Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Plastid genes across the Great Divide

A thesis presented in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

in Evolutionary Genetics/Bioinformatics

Institute of Fundamental Sciences

Massey University, Manawatu

New Zealand

Simon J.L. Cox

2014

Abstract

Nearly all life that is visible to the naked eye is arguably a direct consequence of one or both endosymbiotic events that took place early in evolution and eventually resulted in the mitochondrion and the chloroplast. The timing of the mitochondrial endosymbiotic event weighs argument around the nature of LUCA (Last Universal Common Ancestor) being complex or simple and challenge the commonly taught view of bacteria being the first kingdom to emerge from the primordial state.

The ancient metabolic pathways of amino acid and vitamin biosynthesis are examined and Ancestral Sequences constructed in order to discover the endosymbiotic signature within the nucleus of eukaryotes. Cyanobacterial and plant enzymes from these pathways are tracked as they cross from a prokaryotic coding environment to a eukaryotic one. If the eukaryote that took up the chloroplast ancestor was heterotrophic then it probably got some of its co-factors (vitamins) and essential amino acids from its diet. However, in order to become autotrophic it would have to be able to synthesise these amino acids and co-factors directly. The most likely source of these elements would have been the cyanobacterium; therefore cyanobacterial homologs should be found in the nuclear genome of plants.

Ancestral Sequence Reconstruction (ASR) had a negligible effect on uncovering deeper endosymbiotic homologs. However ASR did confirm ancestral convergence between chloroplast and cyanobacterial homologs and between eukaryote nuclear genes and their cyanobacterial counterparts for vitamin and amino acid biosynthetic pathways. The results, all significant, show that the convergence is much stronger between organisms from the same coding environment (prokaryote [chloroplast] versus prokaryote [cyanobacteria]) than from different coding environments (eukaryote [nuclear] versus prokaryote [cyanobacteria]).

Contents

Table of Figures

Tables

Photographs

Chapter 1. Literature Reviews

Introduction

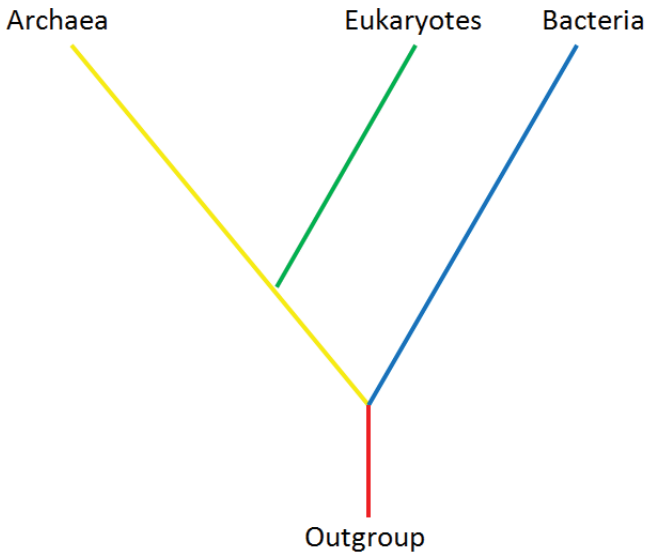
Scientists stand on the shoulders of those that have gone before them. Back in the early nineteenth century the nucleus in cells was discovered in plants by light microscopy but not formally acknowledged until the first half of the twentieth century when Edouard Chatton came up with the term “prokaryote” to describe organisms without a nucleus (Chatton, 1925). From then all life was characterized by this distinction – the presence or absence of a nucleus- from phenotypic observation.

With the advent of electron microscopy in the 1940’s and 50’s, further discoveries reinforced this distinction with the observation of double membrane enclosed organelles as well as the nucleus being found in eukaryotes. This led to the emergence of the concept of bacteria – organisms without a nucleus or any of these organelles - and implied a deep phylogenetic division between these two different types of cells based upon their different organisational plans (Stanier and Niel, 1962).

In 1977, the concept of a third division in the tree of life emerged (Archaea) that effectively split the prokaryotes into two domains (Woese and Fox, 1977) and was formalised in 1990 . Interestingly, this was achieved through the use of alignment based gene trees; the same method we use today to track molecular evolution. And so biologists end up with the readily accepted model of life, known as the Gogarten tree (Gogarten *et al.*, 1989), that is illustrated below in Figure 1.1.

A great deal of effort has gone into elucidating the origin of eukaryotes – the kingdom from within which complex life forms evolved, humans in particular. Eukaryotes are unique in that all have nuclei, mitochondria (or the remnants of) and some have plastids. It is well established that both mitochondria and plastids are endosymbionts and that both of these endosymbiotic events occurred relatively early in the evolutionary formation of eukaryotes – mitochondria first, followed at some later stage by chloroplasts. The typical eukaryotic cell appears to have a chimeric origin with its informational genes having the highest degree of homology with archaea and its operational genes having the highest degree of homology with bacteria (Ribeiro and Golding, 1998, Rivera *et al.*, 1998)- this implies some sort of joining (fusion, endosymbiosis) between these two domains to form the eukaryotic lineage. There are however alternative theories to just how eukaryotes evolved which are discussed later in this chapter.

Figure 1-1-illustrating the 3 domains of life as we now know it. While this model (Gogarten et al., 1989) is readily accepted there is dispute about when the archaea and eukaryotes diverged, and about the nature of the out-group. This is discussed at length, along with associated implications later in this thesis.



Is this readily accepted model of the three domains of life really what happened – at this stage all scientists can say is that we really don't know but Figure 1.1 is the most agreed-upon model.

Current thought concludes that the Last Eukaryotic Common Ancestor (LECA) was a complex organism capable of meiosis, division of genes into many pieces (intron-exon structure), complex transcript processing, and nonsense mediated decay (NMD) as well as possessing mitochondria (Collins and Penny, 2005, Egel and Penny, 2007, Kurland *et al.*, 2006, Roy and Gilbert, 2005, 2009a, Roy and Irimia, 2009b, Pisani *et al.*, 2007). All of these eukaryotic specific attributes would have to have been developed in the uni-cellular lineage leading to LECA (listed in Table 1.1) and presumably took a significant amount of time to develop.

Table 1-1-List of attributes thought to be properties of the Last Eukaryotic Common Ancestor [LECA] from (Poole and Neumann, 2011).

Feature	References
Mitochondrion	(Embley and Martin, 2006, van der Giezen and Tovar, 2005)
Phagocytosis	(Cavalier-Smith, 2002b, Jékely, 2003, Jékely, 2007, Yutin *et al.*, 2009)
Nucleus and nuclear pore complex	(Bapteste *et al.*, 2005, Devos *et al.*, 2004, Devos *et al.*, 2006, Mans *et al.*, 2004, Neumann *et al.*, 2010)
Endomembrane system	(Dacks *et al.*, 2003, Dacks and Field, 2007, Field and Dacks, 2009, Jékely, 2003, Jékely, 2007, Neumann *et al.*,

	2010)
Mitosis and meiosis	(Cavalier-Smith, 2002a, Ramesh *et al.*, 2005, Egel and Penny, 2007)
Introns and spliceosomal apparatus	(Collins and Penny, 2005, Jeffares *et al.*, 2006, Roy and Gilbert, 2005, Roy and Irimia, 2009b)
Linear chromosomes and telomerase	(Nakamura and Cech, 1998)
RNA processing	(Collins *et al.*, 2009, Gardner *et al.*, 2010)
Peroxisome	(Gabaldon, 2010, Gabaldon *et al.*, 2006)
Cytokinesis	(Eme *et al.*, 2009)

As a eukaryotic feature, endosymbiosis is a relatively complicated event. While there are multiple instances of phagocytosis where small cells are engulfed by larger cells for their nutritional value, for a larger cell to engulf a smaller cell and not digest it but keep it alive and functioning for many lifecycles while the endosymbiont and host learn to coexist is a complicated feat.

This thesis deals particularly with the signature of cyanobacterial genes found in the nucleus of archaeplastidal eukaryotic organisms (archaeplastida are the red and green algae, the glaucophytes, and the land plant lineage),(Ball *et al.*, 2011). The ultimate goal is that of characterizing the lineage leading to LECA in terms of autotrophy or heterotrophy. This is to be deduced by the nature of the enzymes that catalyse reactions in the ancient and essential metabolic pathways of vitamin (co-factor) and amino acid biosynthesis. If the pre-plastid state was such that this cell obtained its nutritional requirements from other organisms (heterotrophy) then it is expected that many of the pathways would be incomplete or degraded to some extent because vitamins and some essential amino acids would have been obtained through the diet.

Conversely, if the pre-plastid cell could synthesise these molecules *de novo* (autotrophy) then its pathways would always have been complete. Thus, if these biosynthetic pathways exhibit a high degree of homology with cyanobacteria then this would suggest that the pre-plastid organism was a heterotroph. Many of the hosts pathways would be degraded through lack of use and with the influx of cyanobacterial genes, upon the endosymbiotic event that led to the chloroplast, would restore these pathways to being fully functional. To summarize, if extant Archaeplastidal vitamin and amino acid biosynthetic pathways have a high degree of homology with cyanobacteria then the pre-plastid organism was most likely a heterotroph.

There are 3 scenarios that may explain the origin of vitamin synthesis in archaeplastida (plants and algae)-

- the cyanobacterial genes have been co-opted by archaeplastida in order to synthesize vitamins
- some other bacterial genes have been co-opted by archaeplastida in order to synthesize vitamins
- the ancestral archaeplastida could always synthesize its own vitamins

However, these scenarios are not mutually exclusive and it is likely that there exists a combination of these scenarios throughout archaeplastida genomes (Ribeiro and Golding, 1998, Rivera *et al.*, 1998). The homologous relationships between some eukaryotic and bacterial genes have usually been explained as originating from the eukaryotic symbionts mitochondria and the plastids. In a study only 18% of the genes found in *Arabidopsis thaliana* had homology with the cyanobacteria (Martin *et al.*, 2002) which leaves over 80% to have had other origins. This concept of chloroplast genes in the nucleus has been examined here in the essential amino acid and vitamin biosynthetic pathways to help add to the growing body of knowledge on the origins of the eukaryotes.

To assist in homology detection a standard Ancestral Sequence Reconstruction (ASR) program is used. ASR is a mature, well-established technique as demonstrated in the next section. This program generates an Ancestral Sequence (AS) from an alignment of pertinent extant sequences and a model of evolution (phylogenetic tree construction). The expectation is that deeper homologs will be discovered by using a putative ancestral sequence rather than an extant one (see Collins and Penny, 2005). The ASR methodology is discussed next.

Literature Review - Ancestral Sequence Reconstruction (ASR) Methodology

Introduction

ASR is a technique whereby an ancestral state of a given group of extant proteins is inferred following the alignment of sequences, construction of a phylogenetic tree and building of ancestral sequences. I have used this technique to search, examine, and form putative ancestral enzymes. The technique has been used to construct ancestral proteins both in silico and in vitro. ASR is a mature, well-established technique that has use in many applications as demonstrated in the following sections.

Thornton and colleagues have used ASR to actually synthesise ancestral steroid receptor genes, that reconcile the activities of hormones that direct sexual differentiation, reproduction, behaviour, immunity, and stress response, from which all extant steroid receptors evolved. Results indicated that steroid reception evolved before the development of bilaterally symmetric animals and was subsequently lost in arthropods and nematodes. Thornton used ML (Maximum Likelihood) reconstruction of the protein sequences and MP (Maximum Parsimony) to select the tree; the sequence matrix, that models evolutionary mutation rates, was custom and based on an empirical model of protein evolution with a gamma distribution of evolutionary rates across sites (Thornton *et al.*, 2003). This synthesis of a predicted ancestral protein that carries out a predicted function is a very important test and establishes the veracity of the ASR technique. There are several Thornton papers using this established technique (Finnigan *et al.*, 2012, Ortlund *et al.*, 2007, Bridgham *et al.*, 2006).

Sometimes this method is used to discover the function of now dysfunctional genes. Adey and colleagues (1994) resurrected a mouse "F" LINE-1 retroposon from a sub-family thought to act as a template that encodes reverse transcriptase; there are two types of this retroposon A & F ; "A" being transcriptionally active and "F" considered functionally extinct. Through ASR the "F" was resurrected to a stage where there was promoter activity. The authors comment the broad application of this technique in evolutionary studies. Adey and colleagues used the PAUP phylogenetic package (Phylogenetic Analysis Using Parsimony) (Swofford and Begle, 1993).

Sun and colleagues (2002) used this technique to infer the ancestral states of three *Pax* genes; these genes have an essential role in the embryonic development of organs and tissues (ranging from the eyes and central nervous system to the pancreas and B-lymphocytes) and they have been isolated in a range of Animalia – *Drosophila*, cnidarians, sea urchins and other animals. These Pax genes were functionally grouped and ASR carried out for the two main supergroups in order to determine when

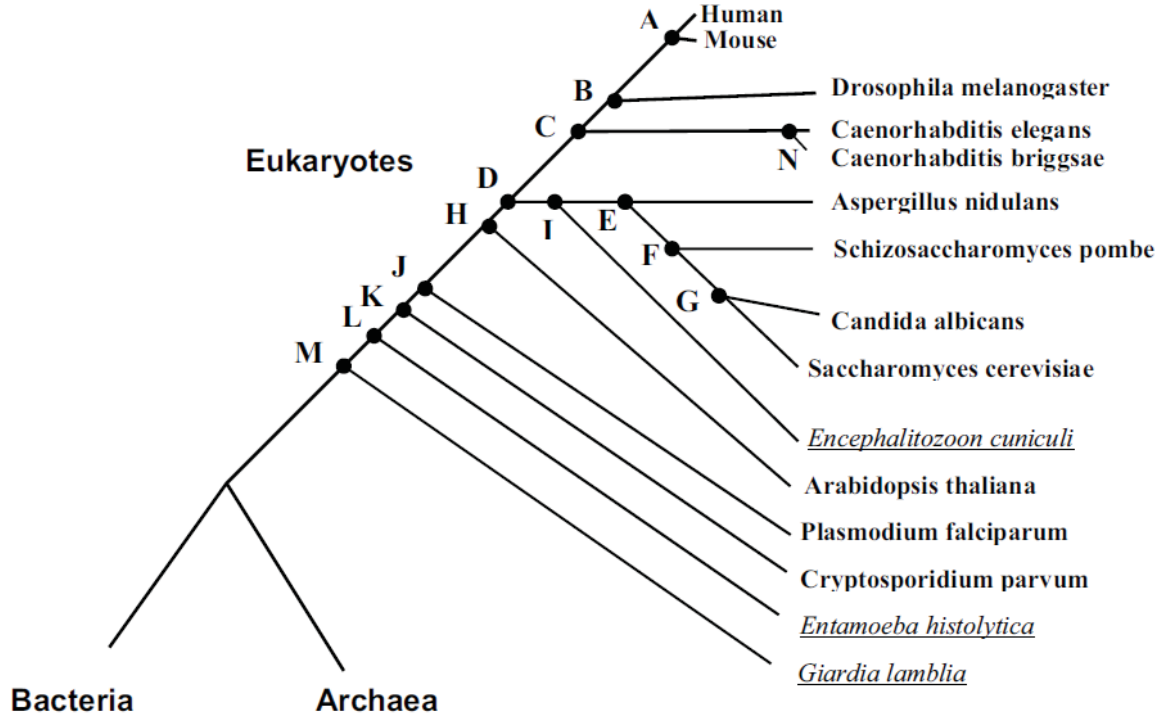
functional differences in the genes appeared – the results also added insight to the larger question of how transcription factors and binding sites evolved. Sun and colleagues used the Bayesian method of ASR that was developed by Zhang and Nei (1997).

More recently, in Voordeckers and colleagues paper (2012), ASR was used to reconstruct the ancestral protein from a set of fungal glucosidase paralogs (enzymes involved in breaking down complex carbohydrates such as starch and glycogen) in order to gain clear experimental evidence on gene duplication and the following functionalization. They used PAML (Yang, 2012) – a standard computer package - for the reconstruction finding the JTT model of protein evolution optimal for the synthesis of the ancestral enzymes. They used MrBayes (Ronquist and Huelsenbeck, 2003) to optimize their phylogenetic tree using a Neighbour joining tree and a GTR model.

ASR may also be used to uncover potential gene homologs as is the case in this thesis. For example, Collins and colleagues (2003) used ASR to uncover RNase P protein homologues in organisms that retrieved no results using extant RNase P proteins in BLAST and HMMER databases. For example, four of these proteins (Pop4, Pop1, Pop5 and Rpp21) had been found in humans and other eukaryotes- mouse, *Drosophila melanogaster*, *Caenorhabditis elegans* and the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*; the authors were interested in finding homologs in three other eukaryotes – two of which were from different supergroups than the discovered homologs (*Entamoeba histolytica* – an Amoebazoa, and *Giardia lamblia* – an Excavata – see Fig 1.2).

Alignments of these homologs were constructed and phylogenetic trees generated in order to ascertain the most likely ancestral amino acids. ClustalX (Thompson *et al.*, 1997) was used for the Multiple Sequence Alignment (MSA) and tree construction, PAML for the sequence construction using the WAG model (Whelan and Goldman, 2001).The construction of the AS allowed more distant homologs to be discovered by entering the AS as a search term, as illustrated in Figure 1.2 above (Collins *et al.*, 2003). For example, no homologs were discovered in Amoebazoa (*Entamoeba histolytica*) and Excavata (*Giardia lamblia*) when the animal and plant homologs were used but they were discovered when the AS was used.

Figure 1-2- Illustrates the depth of AS used to find homologs from Collins *et al.*, 2003. The letters correspond to reconstructions used to aid in the discovery of more distant homologs for the underlined taxa; branch lengths do not correspond to phylogenetic distance.



Daly and colleagues (2013) used ASR to construct in silico ancestral sequences from extant Major Vault Proteins (MVP) in order to search for these vaults in a wide variety of eukaryotes. They used Muscle (Edgar, 2004) for the multiple sequence alignment, MrBayes for tree construction and both PAML (Yang, 2007) and FastML (Ashkenazy *et al.*, 2012) for sequence construction. iTASSER was also used for 3-dimensionsl reconstruction of MVP (Zhang, 2008).

Wolf and Koonin (2013) analysed Ancestral Reconstructions in many domains of life with the aim of tracking gene gain and loss among the many lineages. They argue that evolution is characterised by short periods of gene gain followed by long periods of gradual gene loss; analysis of 120 archaea tracking 1400-1800 gene families showing substantial gene reduction from the putative ancestral state. This constitutes the most compelling evidence of gene reduction from the three supergroups though other cases are found in bacteria and eukaryotes. Table 1.3 below lists some of the other studies where ancestral reconstruction of a last common ancestor has been used in order to demonstrate that ASR is a well-used method at substantial evolutionary depth.

Table 1-2- List of taxa and associated evolutionary reconstructions from Wolf *et al.*, 2013.

Taxa	Depth of evolutionary reconstruction	Subject of Reconstruction	Method	Reference
Mitochondria	Proto-mitochondrial endosymbiosis, LECA	Genes	MP ML	(Embley and Martin, 2006, Lane and Martin, 2010)
Lactobacillales	Last common ancestor of Bacilli	Gene Families	ML MP	(Makarova *et al.*, 2006, Makarova and Koonin, 2007)
Anoxybacillus flavithermus	Last common ancestor of Firmicutes	Gene Families	ML MP	(Saw *et al.*, 2008)
Rickettsia	Last common ancestor of Rickettsia	Genes	MP	(Blanc *et al.*, 2007, Merhej and Raoult, 2011)
Cyanobacteria including chloroplasts	Last common ancestor of cyanobacteria	Genes	MP	(Larsson *et al.*, 2011)
Archaea	Last common ancestor of archaea	Gene families	ML MP	(Wolf *et al.*, 2012)
Eukaryotes	LECA	Protein domain families	MP	(Zmasek and Godzik, 2011)
Eukaryotes	LECA	Introns	ML MP	(Csuros *et al.*, 2011, Rogozin *et al.*, 2012)
Microsporidia	Last common ancestor of Microsporidia	Genes	No trees	(Corradi and Slamovits, 2011)

Fournier and Alm (2015) used ASR in their analysis of the evolution of the genetic code and the amino acid alphabet. One model of the genetic code expansion suggests that tryptophan (Trp) was added to the code by the divergence of aminoacyl-tRNA synthetase (aaRS) families. ASR was used to construct ancestral sequences of paralogs tryptophan and its paralog tyrosine (Tyr) and results suggest that Trp was indeed a subsequent addition to the genetic code. Interestingly this event most likely took place pre-LUCA (Last Universal Common Ancestor) maybe >3Bya and so demonstrates the ability of ASR to work on very ancient proteins. Trees were constructed using ML (WAG model) as were the MSA (JTT model).

There is one factor in particular that limits the depth that ASR effectively may operate. It is that the mathematics of Markov models limits the distance back in time that we can reasonably and confidently infer phylogeny from sequences. At the very deepest times, the Markov models we currently use lose information exponentially (Mossel and Steel, 2004), even though there is also a linear increase in information with the number of sequences (Penny *et al.*, 2014). This is discussed at length later on in this thesis.

As demonstrated above, ASR is a well-established, mature phylogenetic technique that has been successfully used in very many applications (Thompson *et al.*, 1997, Zhang and Nei, 1997, Swofford and Begle, 1993, Diallo *et al.*, 2010, Tamura *et al.*, 2013).

Construction of Ancestral Sequences

Construction of ancestral sequences is a three part process-

1. the primary homology assessment is the multiple sequence alignment (MSA) process,
2. a secondary homology assessment is the tree building process (Morrison, 2009),
3. the inference of the ancestral sequences themselves.

Thus, when inferring an ancestral sequence a multiple sequence alignment is first constructed from an assortment of pertinent extant sequences in order to identify similarities among the sequences.

For the second part of the process, there are two main methods used to calculate the optimal phylogenetic tree- Maximum Parsimony (MP) and Maximum Likelihood (ML. including Bayesian methods). Put simply, MP assumes that the phylogenetic tree with the smallest amount of change (mutation) from extant sequences to ancestral sequence is going to be the correct tree. ML allows an explicit model of character evolution and so chooses the tree with the best fit to this model. Models can estimate probabilities of differing kinds of mutations within the sequences and score an optimal tree from these probabilities. There are adherents for both methods of tree construction (Mount, 2008, Sober, 2004, Steel, 2002, Steel and Penny, 2005). Finally, the ancestral sequences are inferred, usually with ML, these methods of ASR produce much more robust sequences due to the specific models of character evolution than the MP methods but more data is required (Wolf and Koonin, 2013). These models, used in ML constructions, work by weighting the probabilities for changing (evolving) nucleotides or amino acids. For example- differing probabilities would be assigned to the chance of adenine evolving to guanine (both being purines – transitions) than to adenine evolving into a thymine (a pyrimidine –transversions). Similar weights may be modelled to amino acid changes- amino acids with similar physico-chemical properties tend to interchange with each other at higher rates than dissimilar amino acids; for example aspartic and glutamic acids and isoleucine and valine both have similar properties (Yang, 2006). Differing scores are also given to insertions and deletions.

There are many differing models. Use is dependent on which

- genetic unit is to be analysed- nucleotide, codon or amino acid
- genetic code is pertinent- universal, mitochondrion, plastid
- model best describes the weight of transition and transversion; some models will keep the same proportions of the nucleotides or amino acids in the input data as the output data (base heterogeneity), for example the TN93 model
- level of evolutionary pressure is acting on which part of the protein (among site heterogeneity); due to differing roles in the structure and function of different parts of that protein, some areas maybe substitutional hotspots while other areas may be highly

conserved, for example the JTT-Γ model. These models ending in a capital gamma (Γ) allow for differing mutation rates at differing parts of the protein by randomly drawing a mutation rate function from a statistical distribution (Yang, 2006).

In the PAML 4 package, which is used in all reconstructions in this thesis, an Empirical Bayes program is used alongside the ML methods to account for differences in branch lengths of the phylogenetic trees generated. This approach augments/replaces parameters in the model, branch lengths and substitution rates for example, with their ML estimates. This combination of Empirical Bayes and ML methods has been proven optimal (Hanson-Smith *et al.*, 2010, Yang *et al.*, 1995). Care must be taken in the fact that the sequences produced are only pseudo-data not real observed data; this being stated, the sequences produced do have theoretical application (Yang, 2007).

Summary

ASR has been used successfully in a number of differing applications; in this study it has been used initially to assist in discovering deeper cyanobacterial homologs than extant sequences. ASR is a three part process- firstly alignments of extant sequences are produced, secondly an optimal phylogenetic tree is produced using maximum likelihood methods which allow for specific models of character evolution to be used, and finally the ancestral sequences are inferred. There is some discussion about the models and methods used to best model evolution for producing the optimal phylogenetic tree.

Literature Review- Last Eukaryotic Common Ancestor (LECA)

There have been many approaches that have tried to elucidate the origins of the last eukaryotic common ancestor (LECA). Initially, ribosomal RNA was used to construct a phylogeny of the tree of life (Woese *et al.*, 1990) followed by many different single (Brown and Doolittle, 1997) or concatenated sets of genes; differing methods of phylogenetic analysis have also been used (Gribaldo *et al.*, 2010, Poole and Penny, 2007). Advances in these approaches have been aided by the massive influx of fully sequenced organisms, or their proteomes, from all the domains of life and this has meant that there are now large numbers of genomic data to compare. However, to date there has been no definitive answer on the properties of LECA.

One current hallmark of the eukaryotic condition is the presence of endosymbionts. The endosymbiotic uptake of the cell that would become the mitochondrion was an important event that may have eventually powered this radiation of eukaryotes in the era when oxygen was available. In a similar manner, the chloroplast has helped the radiation of the archaeplastida, including green plants and red algae. In both these instances there have been outright gene losses from the endosymbiont (mitochondria and chloroplast) as well as gene transfer to the host nucleus. As both mitochondria and chloroplasts are accepted to be endosymbionts from the bacterial domain (α-proteobacteria and cyanobacteria respectively), this transfer of bacterial genes to the nucleus adds to the chimeric nature of the eukaryote genome.

The Problem of LECA

One problem that scientists face in elucidating LECA is that the eukaryotic cell's genome has a nature that allows two genetically distinct ancestors. Most eukaryotic genes for transcription, translation and replication (that is informational genes) have their closest homologs within the archaea while the genes encoding proteins for many metabolic and biosynthetic functions (that is, operational genes) have their closest homologs within bacteria (Rivera *et al.*, 1998, Ribeiro and Golding, 1998). This poses the question of which genes originated with the host and which genes originated from the endosymbiont; there may also be a uniquely eukaryotic lineage that originated pre-LECA and pre-Mitochondrial Endosymbiosis (ME).

Eukaryotes possess more bacterial-related genes than they possess archaeal-related genes. Of the 5,833 human proteins that have homologs in these prokaryotes at specified thresholds, 4,788 (80%) have greater sequence identity with bacterial homologs, whereas 877 (15%) are more similar to archaeal homologs (196 are ties) (Dagan and Martin, 2006).

Interestingly in a 1999 paper Forterre makes a point about this chimeric nature. He argues that organisms from all kingdoms have this chimeric nature and it is not limited to just eukaryotes but to

both archaea and bacteria. Comparative genomic studies show that organisms from each domain have a mosaic nature and that this the norm rather than the exception in terms of gene content (Forterre and Philippe, 1999). They also argue that eukaryotes could well be the ancestral condition and it is only a prejudice of life moving from simple to complex that sustains the current interpretation of the Gogarten tree (see Fig 1.1).

Another big part of the problem is that there are no identifiable fossil records of cells that were evolutionary intermediates between eukaryotes and prokaryotes (Mayr, 1998). With plant and animal evolution, there are fossils that record the development of features (intermediate forms) that are specific to these lineages so that their evolution could be tracked and dates estimated. There was a discovery of a large fossilised cell (300µm in diameter) found in 3.2 billion year old shale but this has been controversial and the fossil is such that no internal features may be detected (Javaux *et al.*, 2010). With eukaryotes and archaea, all the sequences used to try and unravel their early history, come from extant organisms; that is, there are no intermediate forms.

Further to these problems, are those associated with finding ancient evolutionary relationships using molecular phylogenies. Phylogenetic signals that could help resolve early relationships are likely to be very weak because of successive substitutions. Also ancient phylogenies are on the whole sensitive to tree reconstruction artefacts that potentially affect the resulting topology to such an extent that some suggest that the three domain tree of life is the result of these artefacts (Tourasse and Gouy, 1999), discussed later on in this chapter (Gribaldo *et al.*, 2010, Mossel and Steel, 2004).

Reconstruction of Ancient Relationships

Trying to discover what happened hundreds or thousands of millions of years ago through the use of aligned DNA or amino acid segments produces some interesting challenges. First of all is the question of the strength and quality of any phylogenetic signal left in extant DNA after millions of years of divergence between the three (or two) domains. The first generally accepted eukaryotic fossil is over 1.8 billion years ago (Knoll *et al.*, 2006), an enormous amount of time for accumulated substitutions to have removed any pertinent signal from extant DNA. However there are some genes, known as sometimes as core genes that are very slowly evolving due to the highly conserved nature of their function; fast evolving sites (though useful for species analyses) are quickly saturated and thereby useless for Kingdom delimitation (Williams *et al.*, 2013).

Secondly, there are assumptions used in some phylogenetic analyses that assume compositional homogeneity (the same nuclear base composition and site rate of evolution).The problem with assuming that the there is no difference in GC content and evolutionary rates across species used in an analysis is that this is clearly not the case. Small subunit ribosomal RNA genes vary widely in both

their GC content (Cox *et al.*, 2008) and evolutionary rates having both fast and slow evolving sites within these RNA genes (Olsen, 1987) , known as compositional heterogeneity. This may cause sequences of similar base arrangement to group together in phylogenies even when they are not closely related (Foster, 2004).

Thirdly there is another artefact known as Long Branch attraction (LBA). This is when sequences with a long branch to their nearest common relative in a phylogenetic tree, cluster with other long branches regardless of their relatedness. It has been suggested that the 3 Domain (3D) tree of life (see Figure 1.9 below) is but a LBA artefact between the long bacterial and eukaryotic branches (Tourasse and Gouy, 1999). All of the above factors have to be considered in the construction of the tree of life.

At this juncture the possibility of reconstructing what happened at the beginning of cellular life should be highlighted. Of concern is that the base of the eukaryotic tree has not been satisfactorily resolved due to the mosaic nature of gene content in the 5 major lineages (Forterre and Philippe, 1999). If the phylogenetic tools that we use cannot solve this problem then logically we have even less chance resolving events that took place even earlier than this (Pisani *et al.*, 2007).

Theories on the origin of LECA

Many theories have been proposed for the origin of the eukaryotic lineage. There are many features that are specific to eukaryotes - the nucleus, endomembrane system, mitochondria, spliceosomal introns, linear chromosomes with telomeres synthesised by telomerases, meiotic sex, sterol synthesis, unique cytokinesis structures, and the capacity for phagocytosis. Support for each of these features being present in LECA is outlined in Table 1 (Poole and Neumann, 2011). All of these features have to be accountable to the theory that accurately explains the occurrence of LECA. To date it is unclear in what order these features appeared.

Key to the growth of this eukaryotic kingdom was the mitochondrial endosymbiotic event (Curtis and Archibald, 2010) and the nature of the cell that took up the mitochondrion. In order to explain these chimeric patterns, both fusion and endosymbiosis have been postulated as mechanisms that explain the presence of mitochondria and the nucleus in the eukaryotic cell.

Poole and Penny, 2007, class their analysis of eukaryote origin into four models, shown in Figure 1.4 below. Fusion of an archaeon and a bacterium (Model 1) has been tentatively rejected (Poole and Penny, 2007) as this is not a “known process”, that is , fusion is an undemonstrated biological process between organisms of these two Kingdoms therefore is less likely to have happened in the past . Also there have been extremely few endosymbiotic associations recorded between an

Archaeon and a bacterium. The possibility that the nucleus is an endosymbiont (Model 2) has also been rejected (Martin, 1999, Poole and Penny, 2001) mainly due to the fact that it bears no similarity to any free living archaeon or bacterium (excepting possibly a genus of Planctomycetes discussed below). Model 3 can also be rejected because no archaea are known to be capable of phagocytosis, and no archaea have been documented to harbour any bacterial endosymbionts. That leaves Model 4 as the basis for the more likely scenario (Poole and Penny, 2007) .

While endosymbiosis and a protoeukaryote is the most likely scenario that accounts for the presence of mitochondria, the origin of the nucleus has not been resolved. While fusion is a theory that could possibly explain eukaryote origin, there are other scenarios that can as well. One theory, discussed later on in this chapter is that eukaryotes are the ancestral condition.

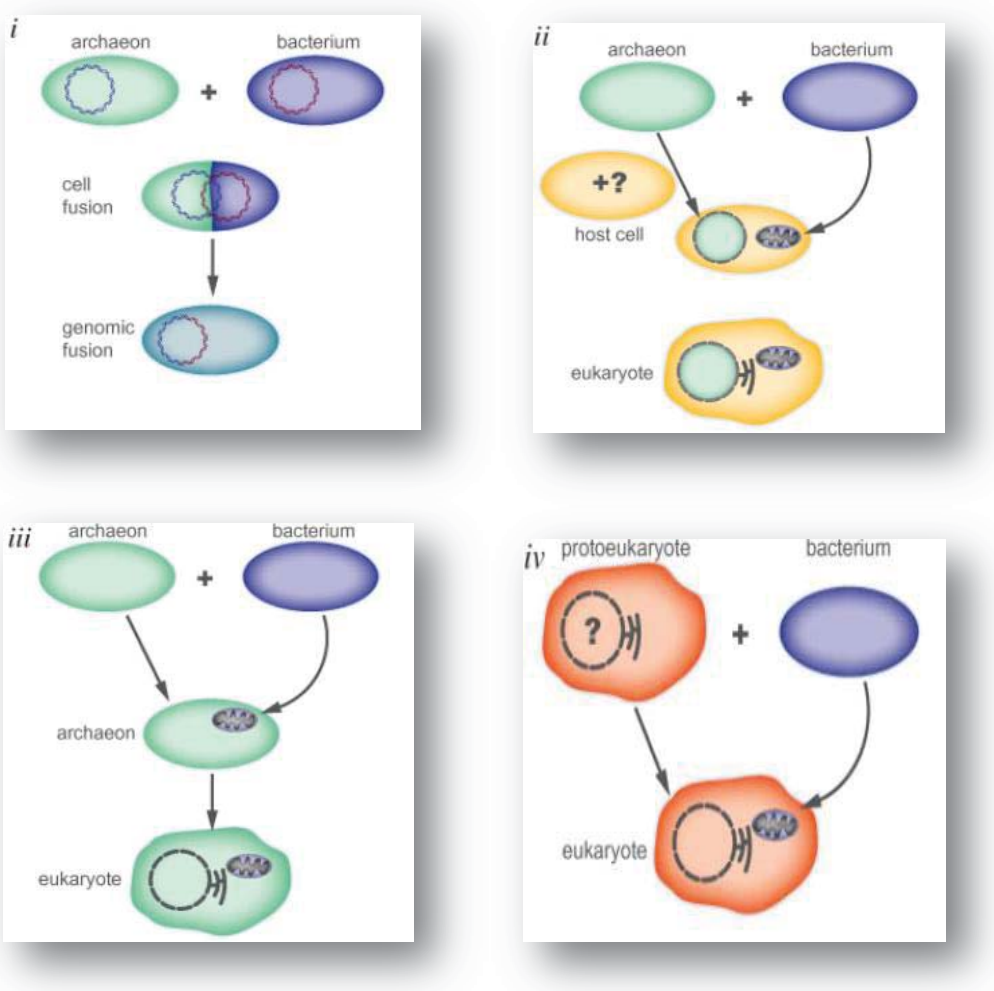
Eocyte theory

The Eocyte theory postulates that eukaryotes evolved from a specific group of thermophiles, the Eocyte archaea (Cox *et al.*, 2008). Using a 40 taxa dataset comprised of the long and short ribosomal RNA sub units the topologies produced were supportive of the Eocyte tree (see Figure 1.5b below), not the 3 Domains tree; interestingly this study used “informational genes” that have the closest homologs within archaea, no use was made of “operational genes” that make up the bulk of genes in eukaryotes, 80% in humans (Dagan and Martin, 2006). The methods used “sophisticated methodologies that accommodate among-site compositional heterogeneity in DNA and protein sequences and lineage specific compositional changes over time” (pp 20049, Cox *et al.*, 2008).

The Eocyte theory is typically dismissed because of the different nature of lipid membrane synthesis between archaea and eukaryotes (archaea possess glycerol-ether membrane lipids, bacteria and eukaryotes possess glycerol-ester lipids). There is a possibility that in the archaeal stem lineage the ester lipid evolved to the ether lipid either before or after the origin of mitochondria (Archibald, 2008, Gribaldo *et al.*, 2010).This raises the question as to the nature of the archaeal host - are extant archaea representative of a type of archaea that may have engulfed a bacterium to form the lineage that becomes LECA?

Figure 1-3- Models for the Origin of Eukaryotes.

Model 1 (from Poole and Penny, 2007) - Fusion: Fusion between an archaeon and a bacterium leading to a singular cellular compartment and a single integrated genome. Model 2- Endosymbiosis by unknown 3rd cell: Endosymbiosis- where a host cell (non-eukaryotic) engulfs both the nucleus and mitochondrial ancestor. Model 3- Endosymbiosis by archaeon: Endosymbiosis where the archaeon engulfs the mitochondrial ancestor. This model implies that no eukaryotic features evolved before this engulfment. Also worth considering is the nature of this archaeon- is it to be found within modern archaea or was it a pre-modern archaea- this is discussed some more in the section on the Eocyte theory. Model 4- Endosymbiosis by protoeukaryote: Endosymbiosis- where the host cell was a proto-eukaryote; implies that the eukaryotes were a separate lineage to the archaea.



2 Kingdoms versus 3 Kingdoms

In a recent perspective, Gribaldo (Gribaldo *et al.*, 2010) summarizes the different arguments and caveats for eukaryogenesis as stemming from a 2 Kingdom (2D) or 3 Kingdom (3D) scenario.

Substantiation for the two scenarios revolves around issues to do with the methods used to construct the phylogenies -see Tables 1.6 and 1.7. All methods have their strengths and weaknesses. Harris and colleagues (Harris *et al.*, 2003) identified a set of 80 universal proteins (genes conserved across the three kingdoms of life) and generated a tree from each protein. Concerns (Gribaldo *et al.*,

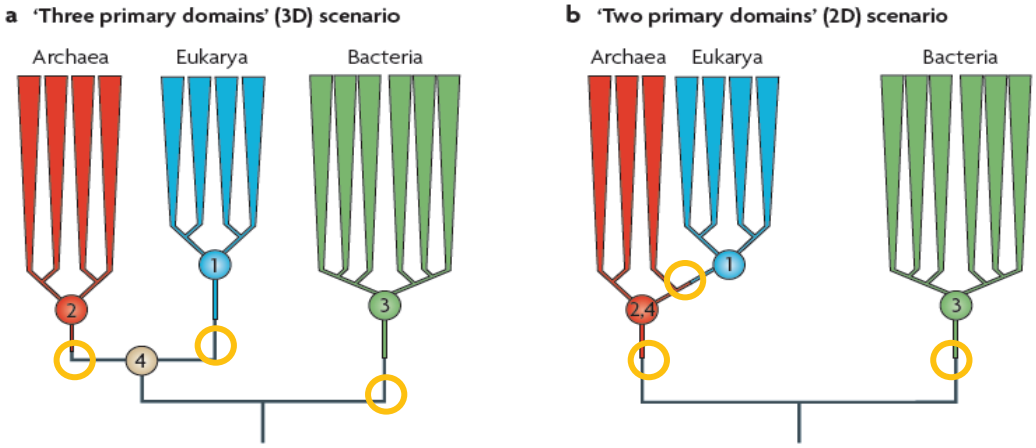
2010) were raised about the low statistical support these individual genes gave for support of the 3 kingdom tree of life.

Ciccarelli and colleagues (2006) used a large taxonomic sample, an objective methodology for identifying Horizontal Gene Transfer (HGT) and concatenated gene sets (they identified 31 universal orthologous proteins) for the phylogenies but used atypical strategies for assembling their datasets. Concerns were raised about the alignment (and manual curation) of genes in each kingdom that was completed before a universal alignment was constructed. A weakness of concatenated datasets is that they require ortholog data sets that overlap between taxa. This limits the number of genes that may be analysed especially when studying the tree of life because of the scarcity of universal genes (Gribaldo *et al.*, 2010).

The 2D scenarios (Figure 1.5b) explain the origin of eukaryotes using known archaeal and bacterial species without postulating the need for a proto-eukaryote (Dagan and Martin, 2007); this is seen by

Figure 1-4- Relationships proposed between eukarya, archaea and bacteria under 2D and 3D scenarios.

The main point of difference between the two is that in the 3D model all of the kingdoms have a specific most recent common ancestor (a- labels 1, 2 & 3) with archaea and eukarya being sister lineages (a– label 4), whereas in the 2D model eukarya derive directly from archaea (b- label 2, 4). Noteworthy are the stems below each kingdom along which kingdom-specific attributes would have arisen, highlighted by the orange circles. From Gribaldo(2010).



its proponents as the 2D argument's main strength. However, there would have to have been proto-eukaryotic lineages that went extinct because of the complexity of eukaryotic specific characteristics that are inferred to have been present in LECA unless all these features evolved in the one lineage without any speciation or diversification – an extremely unlikely scenario.

Table 1-3- Summary of seven recent large scale phylogenetic analyses, based on Gribaldo *et al*, 2010. (ML = Maximum Likelihood, MP = Maximum Parsimony).

Publication	No of Markers	Taxonomic Sampling	AA positions used	Method	Model Supported
(Harris *et al*, 2003)	50	25 bacteria	Variable	Single-gene	3D
		1 crenarchaeaote		ML	
		7 euryarchaeaotes		MP	
		3 eukaryotes			
(Ciccarelli *et al*., 2006)	31	150 bacteria	8,090	Concatenation	3D
		4 crenarchaeaotes		ML	
		14 euryarchaeaotes			
		23 eukaryotes			
(Yutin *et al*, 2008)	136	Variable	Variable	Single-gene	3D
				ML	
(Rivera & Lake, 2004)	Whole genome	2 bacteria	n/a	Gene content	2D
		1 crenarchaeaote			
		2 euryarchaeaotes			
		2 eukaryotes			
(Pisani *et al*, 2007)	Data not available	97 bacteria	Variable	Supertree	2D
		4 crenarchaeaotes			
		17 euryarchaeaotes			
		17 eukaryotes			
(Cox, *et al*., 2008)	45	10 bacteria	5,521	Concatenation	2D
		3 crenarchaeaotes		Bayesian	
		11 euryarchaeaotes		ML	
		16 eukaryotes		MP	
(Foster *et al*, 2009)	41	8 bacteria	5,222	Concatenation	2D
		8 crenarchaeaotes		Bayesian	
		2 thaumarchaeaotes		ML	
		6 euryarchaeaotes		MP	
		11 eukaryotes			

A further argument supporting the 2D scenario is that the eukaryote genome carries genes that are evolutionarily more similar to bacteria and archaea than to its own unique genes. Although the bacterial genes are most likely readily explainable from the eukaryotes organelles, the archaeal ones are harder to justify under a 3D scenario – explanations include that the genes were

- derived from an archaeal/eukaryote exclusive common ancestor (Figure 1.5a, label 4)
- acquired through HGT from archaeal sources

- directly inherited from LUCA (Last Universal Common Ancestor) then lost or replaced in bacteria

A further phylogenetic technique is to use Supertrees. Supertrees are produced using topological information from gene trees (rather than from sequences) to produce a binary matrix. Their advantage is that many datasets, each containing many differing taxa, may be analysed and in so doing bypass the limitations of concatenated datasets. Pisani et al. started with 2,807 genes; first and second attempts produced trees that put cyanobacteria and α-Proteobacteria as sister species to eukarya. Once trees displaying this propensity were removed, as well as those identifying any eukarya sistership with bacteria (a potential source of HGT from endosymbionts) the third tree placed eukarya within Thermoplasmatales, an archaeon. However once trees showing bacterial affinities were removed, only 53 out of the 2807 remained. This study raised a few concerns. First the amount of stripping done in order to obtain trees carrying no bacterial affinity; secondly that Thermoplasmatales is a late diverging archaeon (Gribaldo and Brochier-Armanet, 2006) making sistership with eukarya highly unlikely and thirdly the same concerns that accompany single gene analysis (the trees of which were also used in this study) that these results should be viewed with caution (Gribaldo *et al.*, 2010).

There are certainly a number of opinions on the correct number of kingdoms; key issues revolve around techniques used to construct the trees, summarised in Table 1.7 below. One observation I would make is on the nature of “universal genes”, particularly that these tend to be informational genes, that is, they have an archaeal homology. In light of this it is not surprising that there are 2 kingdom trees produced. These results do not lend confidence to the readily accepted tree of life; other scenarios are presented next in this chapter.

Table 1-4- Summarises the problems with theories about Eukaryotic origins and the number of Kingdoms; adapted from Poole & Penny (2007) and Gribaldo *et al.*, (2010).

Theory	***Problem***	
Eukaryote origin		
Fusion of Archaeon and Bacterium	Poole & Penny (2007)	Fusion is an undemonstrated biological process between organisms of these two Kingdoms therefore is more unlikely to have happened in the past
Nucleus as Endosymbiont	-	It bears no similarity to any free living archaeon or bacterium (excepting a genus of Planctomycetes)
Archaeon engulfs bacterium	-	No archaea are known to be capable of phagocytosis
Eocyte	Cox *et al.*, (2008)	usually dismissed because of differing nature of lipid membranes
No Of Kingdoms	**Author**	**Problem**
2 Kingdoms	Rivera & Lake (1998)	New probabilistic approach- low support for each tree generated from genome fusions- insufficient taxonomic sampling
	Pisani *et al.*,(2007)	Supertrees, conclusion based on most divergent data sets from poorly conserved regions
	Cox *et al.*, (2008)	Concerns over the accurate estimation of a large number of parameters
3 Kingdoms	Harris *et al.*, (2003)	Low statistical support for individual genes
	Ciccarelli *et al.*, (2006)	Concatenation technique may have been biased
	Yutin *et al.*, (2009)	No universal analysis- authors analysed 3 tree topologies – then used combined tree for analysis
	-	1. monophyly of the Crenarchaeota and the Euryarchaeota as support for the 3D scenario
	-	2. the sistership between the Euryarchaeota and the Eukarya as support for the 2D scenario
	-	3. the sistership between the Crenarchaeota and the Eukarya as support for the 2D Eocyte scenario

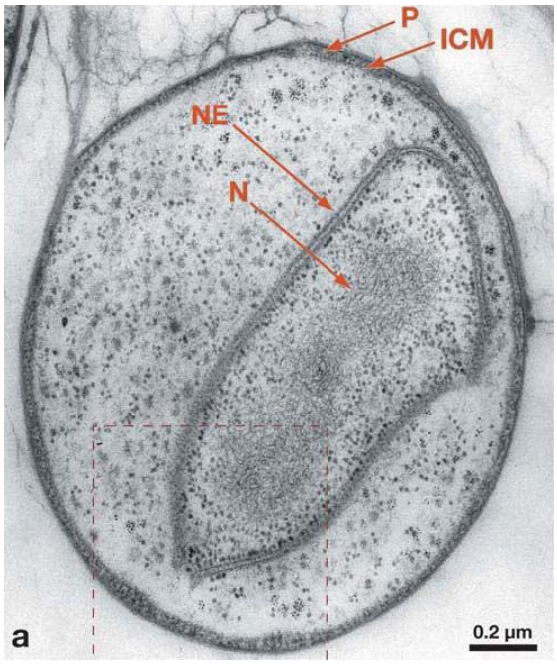
Another possibility

Additional to these theories of fusion and endosymbiosis, is one that suggests that eukarya may be the ancestral condition (Pisani *et al.*, 2007). This is much less mainstream than those theories expounded upon above. Recent research on Planctomycetes is furthering insight into these theories. Planctomycetes are a phylum of bacteria that in some instances are phylogenetically basal in the bacteria tree of life (Goldfarb *et al.*, 2011, Cary *et al.*, 2010). All Planctomycetes carry a unique cell plan that is of special significance when considering the origin of the nucleus; within this phylum, all taxa so far examined contain single-layer membranes that divide the cell cytoplasm (Fuerst, 2005).

One genus in particular, *Gemmata* is of interest for two reasons. Firstly, the nucleoid is surrounded by two membranes suggesting a nucleus-like structure. Figure 1.8 below illustrates this phenomenon in the organism *Gemmata obscuriglobus*. The nucleoid is surrounded by three layers of membrane, one close to the cell wall labelled ICM and the double membrane, labelled NE, surrounding the more immediate nucleiod, labelled N.

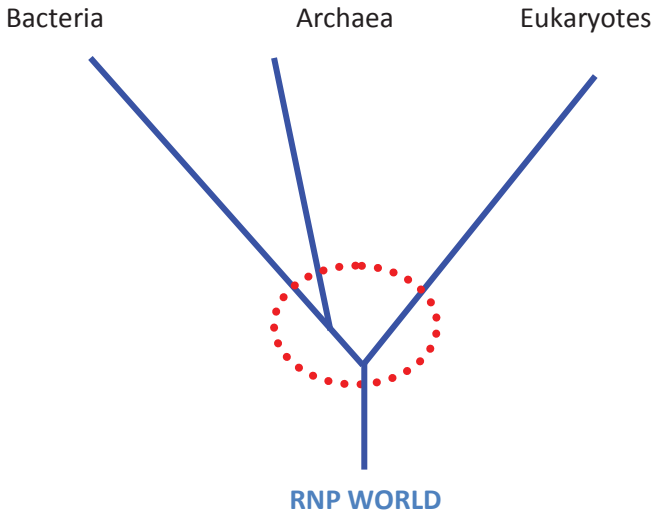
Secondly, it is capable of performing endocytosis (a process by which extracellular material such as macromolecules can be incorporated into cells via a membrane trafficking system) - a quality comparable to the eukaryotic-like trait that led to the endosymbiotic uptake of mitochondrial and chloroplast ancestors (Lonhienne *et al.*, 2010). Both of these observations of nuclear portioning and endocytosis may cast doubt on the bacterial condition being basal in the tree of life.

Photo 1-1- Illustrating *Gemmata obscuriglobus* and the three layers surrounding its nucleiod; from Fuerst, 2005.



As well as this interesting point about Planctomycetes, there is further support for eukaryotes holding the basal position (see figure 1.9 below) from re-analysis of Woese's data that originally proposed the sister relationship of eukaryotes and archaea. These analyses show very strong support for an akaryote empire (archaea and bacteria) that is distinct from the eukarya (Brinkmann and Philippe, 1999, Philippe *et al.*, 2011, Caetano-Anollés, 2002, Harish *et al.*, 2013). Brinkman bases his support of a akaryote/eukaryote tree on a more well curated gene tree using the 54 kDa signal recognition particle and receptor SRα, paralogs that diverged before LUCA. They argue that the archaea/eukarya "sisterhood" is due to long branch attraction and use of fast mutating sites in the original analysis. Focusing on the slowly mutating sites, more conserved and therefore more ancient, gave the akaryote\eukaryote tree. Phillipe *et al.* (2011) argue the need for a high standard of alignment analysis before generating trees –the main sources of error include incorrect identification of orthologs, erroneous alignments and the incorrect reconstruction of multiple substitutions occurring at a given position.

Figure 1-5- Illustration of eukarya properties being basal. The dashed box illustrates uncertainty about the split of bacteria and archaea from the eukaryotic lineage.

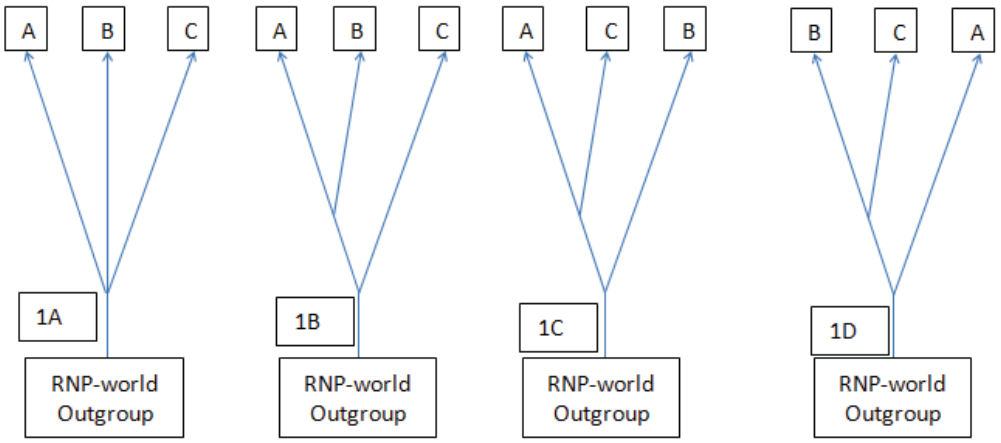


Caetano-Anollés (2012) succeeded in rooting both large and small unit rRNA based on the conserved regions of the genes and the evolving secondary structures; both rooted phylogenies show the sisterhood of the archaea and bacteria. Harish *et al.* (2013) use data generated from the tertiary structure of superfamilies. Superfamilies are defined as "Families, whose proteins have low sequence identifies but whose structures and, in many cases, functional features suggest that a common evolutionary origin is probable (Murzin *et al.*, 1995)". The argument is that as the genetic code is degenerate (different codons produce the same amino acid), valuable information is stored in the three dimensional nature of proteins that is generally invisible in raw sequence alignments. Harish *et al.* (2013) argue that using the tertiary structure of protein kingdoms to construct phylogenies has led to some interesting conclusions. Their analysis of tertiary structures supports the sisterhood of archaea and bacteria. Additionaly, analysis using the superfamilies leads to the conclusion that firstly the organism that gave rise to the three kingdoms was already a complex organism, secondly that since then LACA (Last Akaryote Common Ancestor) and decendents have followed a course of reductive evolution, and thirdly that LECA and decendents have followed a course of duplication. There are some points here that I'd like to make-

Firstly, given the three super kingdoms and a possibly complex RNP outgroup (ribonuclear protein), chosen because a RNP ancestor would be from before DNA had evolved and probably had many regulatory functions as well as coding for protein functions (Collins *et al.*, 2009). This would give the following trees, illustrated in Figure 1.9 below-

Figure 1-6- The unrooted tree (1A) for the three groups' Archaea, Bacteria and Caryotes (eukaryotes).

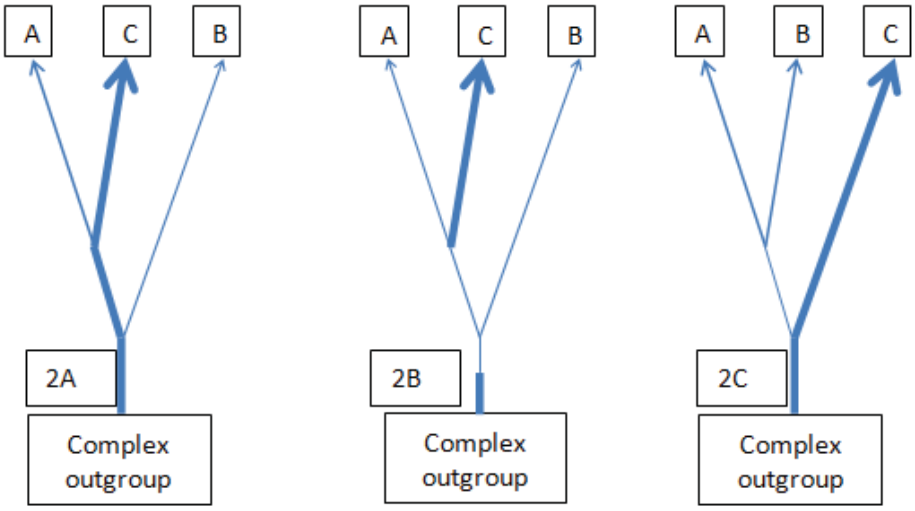
This is followed by the three rooted trees 1B-1D for the groups (Kingdoms). 1B is ((Archaea, Bacteria), Caryotes) recently favoured by Kurland; 1C is the commonly a 'Gogarten' tree ((Archaea, Caryotes), Bacteria); and 1D is ((Bacteria, Caryotes), Archaea) which is a possibility pointed out by Forterre (Forterre and Philippe, 1999).



Now, given that the RNP ancestor was complex then this raises some interesting observations. The currently favoured tree 1C requires either two losses of complexity in the akaryote lineages or a loss of complexity and then a regaining of complexity in the eukaryote lineage, illustrated in Figure 1.11 below (Penny *et al.*, 2014). Given no further information neither scenario would be favoured, and there appears to be no genetic mechanism that favours the redevelopment of an intron/exon structure of genes together with a very complex spliceosmal apparatus (Collins and Penny, 2005).

The are some simulations involving populations of simple single celled organisms (de Nooijer *et al.*, 2009). Given that the standard model of evolution gives eukaryotes a late emergence, simulations were created to test the premise that there was a period of time before eukaryotes emerged when there was no predation. This is based on the idea that in the world of single celled organisms, cells that get their energy from engulfing smaller cells (phagocytosis) are nearly all from eukaryotes. In all instances, starting under a variety of conditions, the strategy of smaller cells and larger phagocytotic cells happened early in the simulations. This may add to support of eukaryotes being basal?

Figure 1-7- The problems of a complex out-group.
If the RNP world used small RNA molecules for regulation, and had an exon/intron gene structure (with complex spliceosome) then there would be two losses of complexity in tree 2A, namely in both the archaea and the bacteria. Conversely, in tree 2B, there may be a loss of complexity before LUCA, and a regain of complexity just in eukaryotes. In 2C the relationships are derived from tree 1B in Figure 1.6 above, and there is only a single loss of complexity, though the relationship in 2C is not favoured by many researchers at all.



The tree of life and phylogenetic markers from the bacterial perspective

In order to construct a tree of all life, a universal phylogenetic marker, i.e. a protein or gene, had to be found that occurs in all three kingdoms of life. This marker had to fill the criteria of universal occurrence, functional constancy, sufficient sequence conservation and complexity. Woese (1990) came up with the rRNA long and short subunits that satisfied these criteria and proclaimed the existence of a three Kingdom tree of life.

Definitions
Homology is the relationship of two characters that have descended, usually with divergence, from a common ancestral character.
Orthology is that relationship where sequence divergence follows speciation.
Paralogy is defined as that condition where sequence divergence follows gene duplication (Ball *et al.*, 2011).

In a paper on bacterial phylogeny Ludwig and Schleifer (2005), made some telling points about marker selection and support for a 3-Kingdom tree of life. Firstly they state that there is general acceptance that rRNA-based conclusions can only roughly reflect evolutionary history. Secondly, markers such as elongation factors, ATPase subunits and RNA polymerases, although giving similar overall tree topologies, gave marker-specific discrepancies in detailed topologies and that these discrepancies were to be expected. Thirdly, markers that fulfilled the criteria of universal occurrence, functional constancy, sufficient sequence conservation and complexity were few. One of the problems they highlighted involved the comparison of sequence data of phylogenetic marker molecules; the issue

revolved around the multiple evolutionarily diverged versions of homologous markers. Their analyses found many examples of duplicates with different degrees of sequence divergence in phylogenetic marker databases. These resulted in the following problems that impeded sound analysis - a true recognition of paralogous markers from orthologous ones, the recognition of copies sharing the same function and selection pressure, and the ability to differentiate clonal duplication and lateral gene transfer.

Han and colleagues (2013) also identified this issue of recognition of paralogous markers from orthologous ones in their study of Eukaryotic Signature Proteins (ESP) and demonstrated the detrimental effects that this misidentification can cause .

The authors (Ludwig and Schleifer, 2005) investigated those markers that fulfilled the above criteria (of universal occurrence, functional constancy, sufficient sequence conservation and complexity) and came up with markers that support the 3-Kingdom model and markers that do not. So, there was no resolution on the question of 2 or 3-Kingdom model from the bacterial perspective.

Viruses - a fourth super kingdom?

There is a long-standing view that viruses don't qualify as life - a view stemming from their small genomes, no metabolic capability and relative simplicity (Pennisi, 2013). However, in recent times there have been discoveries of large viruses that contain more DNA than some organisms from all three super-kingdoms which has been the cause of suggestion that the viruses are a super-kingdom to themselves that may predate the emergence of the three main branches (Raoult *et al.*, 2004, Pál *et al.*, 2005). Raoult and colleagues suggest that the mimivirus descended from a free living cell that gradually lost its other genes on its way to a parasitic lifestyle.

Further discoveries of viruses that are larger than the 1.2 million DNA base mimivirus have added weight to the argument of viruses being an old super-kingdom. A class of viruses, named pandoraviruses, have been found with 1.9 and 2.5 million DNA bases long; they have outstandingly different genes from other viruses as well as conspicuously different phenotypes (Philippe *et al.*, 2013). The suggestion is that these pandoraviruses are an intermediate step between a free-living ancestor and the mimivirus. Of course it could be the other way around, where some viruses accumulated genes until they were able to reproduce independently in the cytoplasm – thereby creating prokaryotes. We simply do not know yet and therefore will not include viruses in any analysis.

The "When" of LECA

If eukaryotes are not the ancestral condition then the crux of the above discussion may be summarized as trying to ascertain the time between mitochondrial uptake and the development of

all the ancestral features of LECA. This is shown in Figure 1.11 below. This thesis looks at the nature of the organisms involved in the mitochondrial endosymbiosis (M.E.). An important point for understanding this is the time it took between ME and LECA. This has implications for both the nature of the host and endosymbiont, and the 2 vs. 3 Kingdom arguments. If the time was short (evolutionarily speaking) between ME and LECA then the 2 Kingdom argument, particularly the Eocyte theory, seems more likely. This would also imply that there was a fusion between an archaeon and a bacterium. If the time was long then the 3 Kingdom arguments seem more likely. A proto-eukaryote endosymbiotic event with a bacterium appears to be the most likely scenario in that instance.

Vitamins, Essential Amino Acids and Metabolic pathways

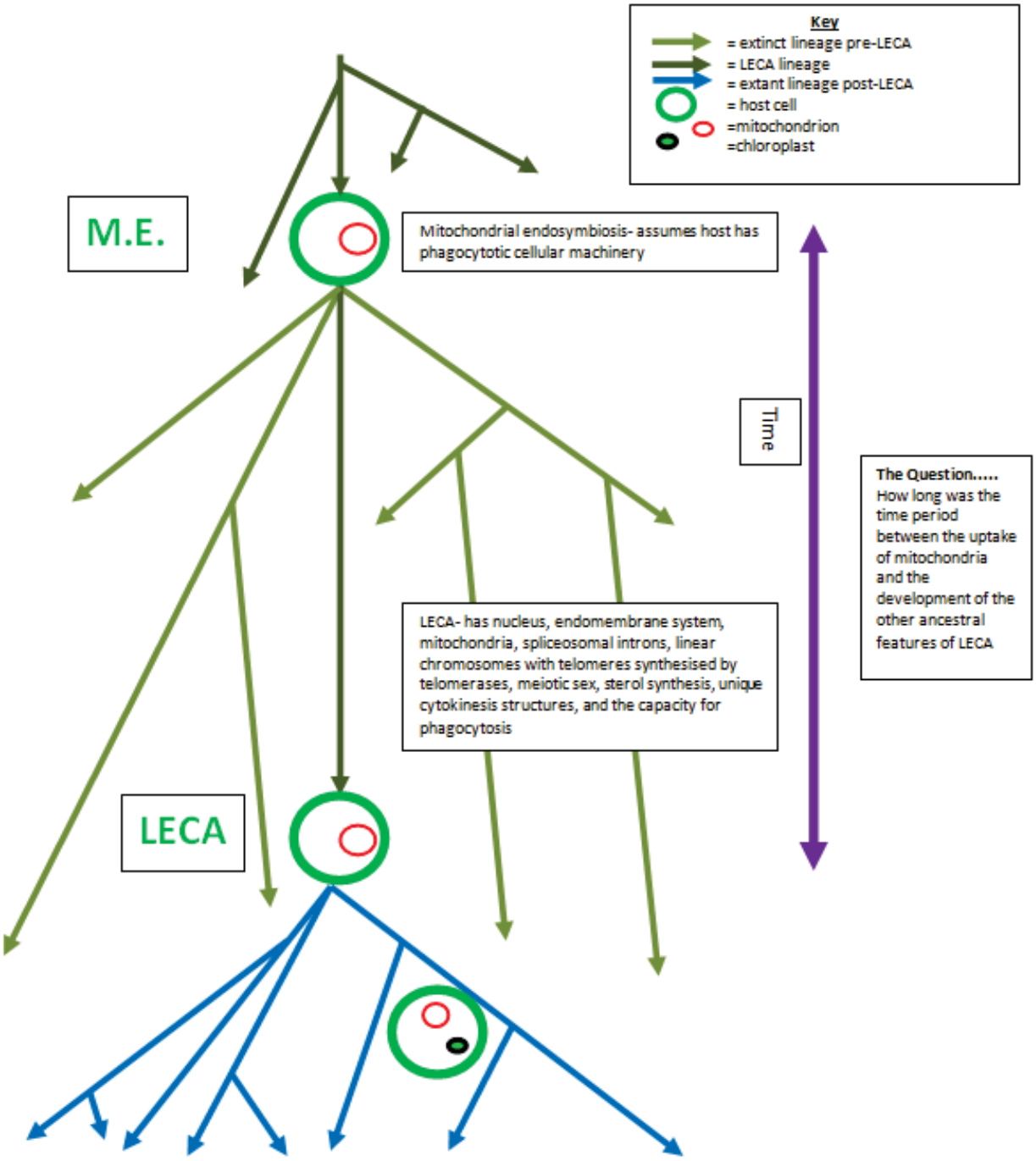
Vitamins/Cofactors

Vitamins are described as organic compounds that are essential nutrients needed in small quantities that must be obtained through the human diet because they are unable to be synthesized (Oxford Dictionaries, 2010). It is likely that they were extant in the pre-biotic RNA world implementing varied metabolic reactions in the absence of proteins (Smith *et al.*, 2007). Vitamins are involved with many central metabolic processes with many acting as co-factors with enzymes, catalysing vital reactions. Humans do not produce these vitamins; they usually come from their diet. The understanding is that as humans gained all their vitamins from their diet the need to synthesise these products became redundant and that therefore being unnecessary to synthesise them we lost the ability to make them. By implication, vitamins were present in the earliest cells (Caetano-Anollés *et al.*, 2009) and it is for this reason that the metabolic pathways used to synthesize these vitamins are being used to track the biosynthetic capabilities of LECA. This is achieved through the analysis of the DNA of the enzymes that catalyse the reactions in these metabolic pathways.

Proteins for the biosynthesis of vitamins from members of archaeplastida (green plants and red and green algae) are examined. Archaeplastida are distinguishable from other eukaryotic super-groups by the presence of plastids arisen from primary endosymbiosis. As such, most of the proteins involved in vitamin biosynthesis are expected to have homology with cyanobacteria due to the chloroplast being most closely related to this prokaryotic bacterial phylum if indeed the ancestral eukaryote was a phagocyte.

Figure 1-8- Summarizes key features of LECA problem.

If the time was short (evolutionarily speaking) between ME and LECA then the 2 Kingdom argument, particularly the Eocyte theory, seems more likely. This would also imply that there was a fusion between an Archaeon and a Bacterium. If the time was long then the 3 Kingdom arguments seem more likely. A proto-eukaryote endosymbiotic event with a bacterium appears to be the most likely scenario in that instance.



There are 3 possible scenarios that may explain the origin of vitamin synthesis in plants-

- the cyanobacterial genes have been co-opted by the plant in order to synthesize vitamins/cofactors

- some other bacterial genes have been co-opted by the plant in order to synthesize vitamins/cofactors
- And that the plant genes from the ancestral eukaryote have displaced the cyanobacterial genes, that is, the ancestral eukaryote could synthesize its own vitamins/cofactors.

However, these scenarios are not mutually exclusive and it is likely that there exists a combination of these scenarios throughout plant genomes. The homologous relationships between some eukaryotic and bacterial genes have usually been explained as originating from the eukaryotic symbionts mitochondria and the plastids. In an earlier study only 18% of the genes found in *Arabidopsis thaliana* had homology with the cyanobacteria (Martin *et al.*, 2002) which leaves over 80% to have had other origins. This concept of chloroplast genes in the nucleus has been examined for the essential amino acid and vitamin biosynthetic pathways.

The point here is that humans and their ancestors have lost the ability to synthesise some of their own vitamins and plants have retained (or gained) this ability.

Essential amino acids

An essential amino acid is an amino acid that cannot be synthesized de novo by an organism and therefore must be obtained from external sources. At the level of the ribosome, eukaryotes normally require up to 20 different amino acids for protein synthesis and consequently a deficiency of any of these amino acids may lead to limited growth or no growth at all. However, most 'animals' can synthesize some of these amino acids from other substrates, therefore, only some amino acids used in protein synthesis are considered essential. Whether a particular amino acid is essential depends upon the species and the stage of development (Rose *et al.*, 1951, Kopple and Swendseid, 1975).

The amino acids essential for human life are phenylalanine, tryptophan, isoleucine, methionine, leucine, valine, threonine, lysine, and histidine. Some others are required by children- tyrosine (or aromatic amino acids), cysteine (or sulphur-containing amino acids), and arginine (de la Torre *et al.*, 2014). There are some instances when populations do not synthesise adequate amounts of the amino acids arginine, cysteine, glycine, glutamine, histidine, proline, serine and tyrosine. These instances involve malnutrition, stress and pregnancy and may also be affected by the growth state of the individual – children needing higher amounts of some of these amino acids at this stage of their development. These amino acids are considered conditionally essential (Fürst and Stehle, 2004, Reeds, 2000).

So the important question here is, was the ancestral eukaryote that took up the chloroplast heterotrophic (and depended on its diet for some essential amino acids and cofactors) then to become autotrophic (and not to depend on its diet) it had to get the necessary enzymes from the

plastid cell (or either modify existing enzymes, or invent them *de novo*). Can we determine which option occurred from the homologs of current cells?

Metabolic Pathways

Several hundred different metabolic pathways and networks are distributed throughout cells that synthesize or break down organic compounds. These biochemical networks are responsible for the metabolic functions that fuel the cell and thus the inner workings of life. These metabolic pathways are highly tailored systems that have been acted upon by billions of years of evolution and as such are central to understanding the evolution of cellular life, multicellularity, and the origins of archaea, bacteria and eukarya. There are many theories on how cellular metabolism evolved. A substantial body of work suggests that metabolic pathways evolve fundamentally by recruitment; enzymes are attracted from close or distant regions of the metabolic network to perform novel chemistries or use different substrates (Caetano-Anollés *et al.*, 2009). There may be signs of this happening in the protein family homologies discussed in the later part of chapter 3. The position of cyanobacterial homologies in the metabolic pathways in this study are discussed later in chapter 4.

Metabolic networks, for the large part, appear to remain relatively stable with archaea's synthesizing ability displaying the most stability, eukarya mid-stability and bacteria less so (Ebenhoh *et al.*, 2006). Bacillales (a gram positive order of Firmicutes) in particular, showing the least stable metabolism; whether this is a result of HGT (horizontal gene transfer) was not discussed by the authors. They do suggest that at various times during evolution, key events occurred that drastically changed metabolism, however, during speciation events metabolic functions appeared only minimally influenced.

Therefore, the choice of vitamin and essential amino acid metabolic pathways, as avenues to further search for the nature and origins of LECA, appear to be both constructive and necessary.

Summary

LECA was undoubtedly a complex organism that had many distinct attributes that would presumably have taken an evolutionarily long period of time to develop. The genes in the eukaryotic nucleus have homologous relationships with both archaea (informational type genes) and bacteria (operational type genes) which have led to theories of fusion and endosymbiosis being postulated as to the likely origin of the eukaryotic lineage. But is this really what happened –alternative theories that eukaryotes are the ancestral condition are presented.

This chimeric nature appears to influence the arguments about a two or three kingdom tree of life. Universal genes suitable for constructing such a phylogeny tend to be informational rather than

operational and so weigh opinion towards a two kingdom tree of life. Adding to the debate is the fact that ancient phylogenies are very sensitive to tree reconstruction artefacts, so much so that both a two and a three kingdom tree have been concluded from the same initial data.

Vitamins and essential amino acids have a very ancient history; both are thought to have been present in the earliest cells. The rationale behind using enzymes from the biosynthesis of these products is that if the pathways display significant amounts of cyanobacterial homologies then this would suggest that the ancestral eukaryote had a predatory lifestyle. This is because a predator would obtain these vitamins and essential amino acids from its diet, thereby no longer having to synthesise its own, these pathways would no longer have use and would degrade (or mutate to a new function). With the introduction of the chloroplast and a sessile, non-predatory lifestyle, on the branch leading to algae and plants, these pathways would once again be essential and the genes needed for the resurrection of these pathways would have come from the endosymbiont cyanobacterium. Therefore if these plant metabolic pathways are full of cyanobacterial homologs then this would lend weight to the argument that the ancestor that took up a cyanobacterium to form an alga was a predator.

Chapter 2. Nucleotide Approach

Introduction

The ultimate goal of this thesis was to determine something of the nature of the cell that formed an endosymbiotic relationship with the bacterium that would become the chloroplast (or even earlier, the mitochondrion). As stated in Chapter One the ancestral eukaryotic cell (LECA) had mitochondria that originated from the endosymbiosis of an α-Proteobacterium. Sometime after this LECA had a similar endosymbiotic event with a Cyanobacterium that led to the formation of the chloroplast and from this all the archaeplastida lineages are derived.

The assumption is that the nature of this cell would be along a spectrum from being either heterotrophic (needs to obtain essential amino acids and vitamins from its diet) to autotrophic (and had the ability to synthesise these acids and vitamins itself). If this cell was heterotrophic then it is assumed that the metabolic pathways that synthesise these essential compounds would be full of cyanobacterial (or α-proteobacterial) homologs. The rationale behind this statement is that if the cell was heterotrophic then these pathways would have been lost through lack of use and upon the endosymbiotic event these pathways would have been revitalised with functioning enzymes with a cyanobacterial (or α-proteobacterial) origin. Thus this putative ancestral cell would have used the genes from these endosymbionts to repair its redundant pathways.

In contrast if this ancestral cell could synthesise all its own essential amino acids and co-factors then the pathways that metabolise these compounds would be essentially intact and have homologs from all over the three kingdoms due to the ancient and essential nature of these pathways. That is very few cyanobacterial (or α-proteobacterial) homologs would be found in these pathways. This thesis examines the chloroplastic endosymbiotic event that happened more recently than the mitochondrial endosymbiotic event and so the signature of this (cyanobacterial genes in the genes of archaeplastida) should be more evident and traceable. Once this has been achieved for plastids, the technique can be expanded to the earlier mitochondrial endosymbiotic event.

Nucleotide approach

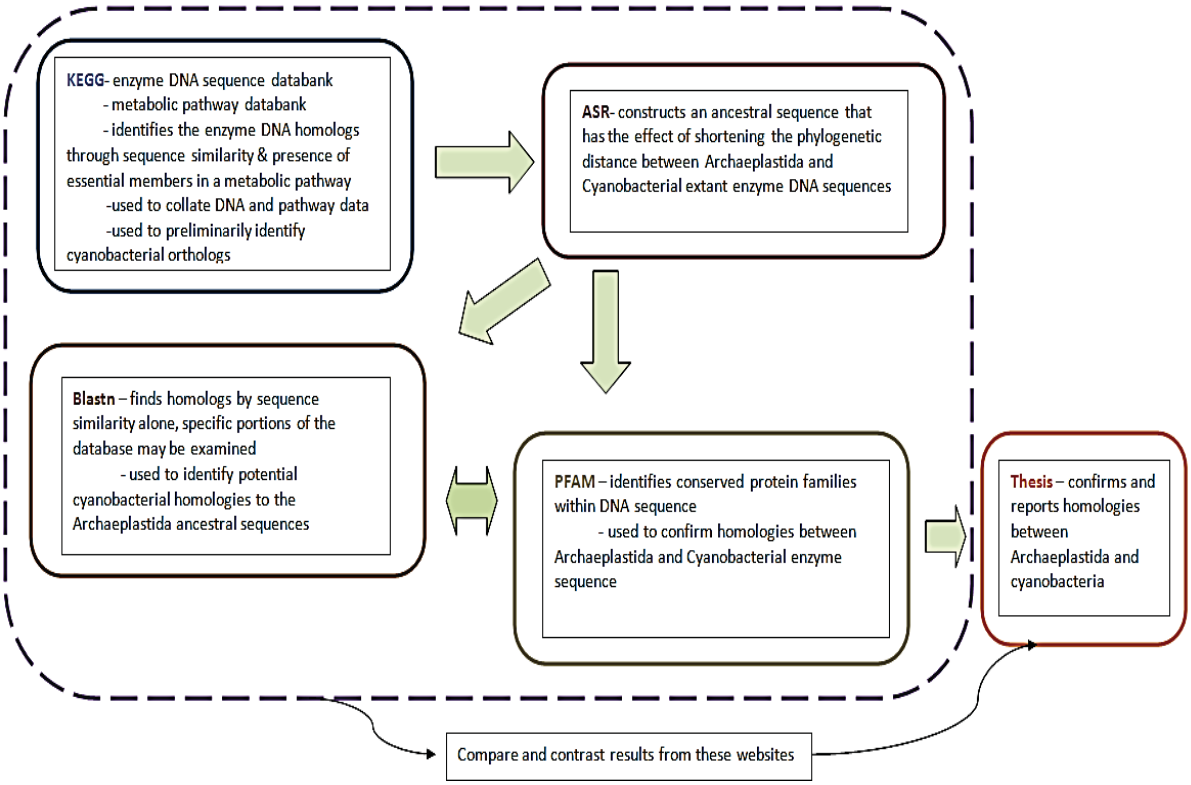
Kyoto Encyclopaedia of Genes and Genomes (KEGG) was used to investigate the taxonomic distribution and the phylogeny of homologs of the enzymes of essential amino acid and vitamin metabolic pathways, from representatives of the archaeplastida from which complete genome sequences were available. Taken together (KEGG plus genomes), these data allow deciphering of the evolutionary history of the pathway and therefore a better understanding of the evolutionary history of the organisms carrying them. As a starting point, phylogenomic approaches rely on the

knowledge of the composition of the metabolic pathway under investigation in at least one representative, this is the model plant *Arabidopsis thaliana* in most instances.

An Ancestral Sequence (AS) (Collins and Lockhart, 2007, Collins *et al.*, 2003) was then constructed using the homologs from the archaeplastida. This sequence was then run through BLAST to find similarity scores amongst the cyanobacteria and thus potential homologies. These ancestral sequences were also run through PFAM (Punta *et al.*, 2011) in order to find conserved protein families. Cyanobacterial homologs identified through BLAST were also run through PFAM.

Those AS homologs with sufficient similarity to cyanobacteria, in particular to an inferred cyanobacterial ancestral sequence, were taken as originating from the chloroplast. This chapter is about refining the pipeline from known processes (that is, the engulfment of the chloroplast ancestor from the cyanobacteria), in order to gain confidence in the results. This brief methodology is summarized in Figure 2.1 below.

Figure 2-1-Overview of methodology



Archaeplastida

There are 12 members of the archaeplastida (the only archaeplastida available on the KEGG database at the time I started), used to identify homologs and construct ancestral sequences, listed below in Table 2.1.

Table 2-1-List of archaeplastida used in this study.

Group	Code	Organism	Organism Number	Assembly Number
Eudicots	ath	Arabidopsis thaliana	T00041	GCF_0000017
	pop	Populus trichocarpa	T01077	GCF_0000027
	rcu	Ricinus communis	T01087	GCF_0001516
	vvi	Vitis vinifera	T01084	GCF_0000037
Monocots	osa	Oryza sativa japonica	T01015	GCF_0000054
	sbi	Sorghum bicolor	T01086	GCF_0000031
	zma	Zea mays	T01088	GCA_0000050
Mosses	ppp	Physcomitrella patens	T01041	GCF_0000024
Green algae	cre	Chlamydomonas reinhardtii	T01039	GCF_0000025
	vcn	Volvox carteri f. Nagariensis	T01330	GCF_0001434
	olu	Ostreococcus lucimarinus	T01029	GCF_0000920
Red algae	cme	Cyanidioschyzon merolae	T00175	GCF_0000912

Pathways

18 different metabolic pathways were chosen for this study. These pathways were chosen for their ancient and essential nature; they (or their equivalents) are assumed to have been present in the Last Universal Common Ancestor (LUCA). They are listed below in Table 2.2.

Table 2-2-The 18 amino acid and vitamin metabolic pathways used in this study

Amino Acid	Vitamin/Co-factor
Histidine	Ascorbate (VitC)
Isoleucine	Coenzyme A
Leucine	Folic acid (B9)
Lysine	Niacin (B3) NAD\NADP
Methionine	Pantothenic Acid (B5)
Phenylalanine	Pyridoxine (B6)
Threonine	Riboflavin (B2) - FAD\FMN
Tryptophan	Thiamine (B1)
Valine	Vitamin H - Biotin (B7)

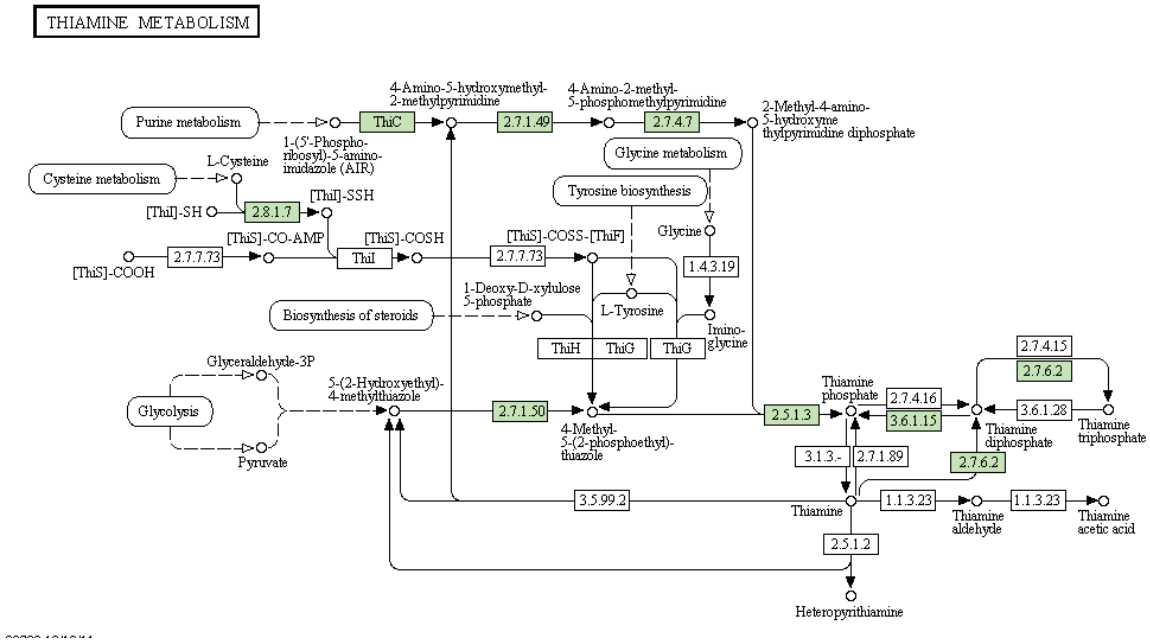
KEGG (Kyoto Encyclopaedia of Genes and Genomes)

KEGG (Kanehisa *et al.*, 2008) is a website that links genomic information with higher order functional information. It lists metabolic pathways that link to sequences from chosen organisms and has a further option to generate each sequences homologies; it also has a function to run a CLUSTALW (Larkin *et al.*, 2007a) analysis on these homologies that may then be displayed as a Neighbour-Joining (NJ) tree. Figure 2.2, below, illustrates a map from KEGG- an organism is selected from a list

then a generic pathway with the reactions. Numbers in the box correspond to the Enzyme Commission (EC) number, or the name of the enzyme, that have been identified for that organism and are shown in a shaded highlight. A further function directs the user to references for each organism's pathway.

Arabidopsis thaliana is the best annotated plant genome and so homologies were found with the KEGG homology function, using *Arabidopsis* enzymes as the basis for the homology search. Enzymatic DNA was collected, for those archaeplastida that expressed that protein, for use in constructing AS. Closest homologous species, outside the archaeplastida, were noted and patterns collated. In all cases, the default settings were used for CLUSTALW and Neighbour-Joining (NJ) trees.

Figure 2-2-Thiamine metabolic pathway; green boxes indicate enzymes that are found in *A.thaliana*, numbers indicate the EC number of that reaction. (http://www.genome.jp/kegg-bin/show_pathway?org_name=ath&mapno=00730&mapscale=&show_description=hide).



When searching for homologs, instead of just searching for sequence similarity, as BLAST does, KEGG considers four aspects in its search for orthologs-

1. sequence similarity
2. whether an organism contains a complete set of genes that constitutes a functional pathway
3. whether these genes are physically coupled on the chromosome
4. what are the orthologous genes among different organisms (Kanehisa and Goto, 2000)

Using these considerations the method employed by KEGG in the search for orthologs can be considered more robust than BLAST.

BLAST

The Basic Local Alignment Search Tool (BLAST) is the most commonly used method for locating homologs in a sequence database (Altschul *et al.*, 1990). BLAST works for both nucleotide and protein sequences. The BLAST algorithm initiates the search by seeding a small subset of letters (be they nucleotides or amino acids) from the query sequence. The query word (sequence) and the related words are located and scored by a scoring matrix and this produces a neighbourhood. BLAST uses a neighbourhood threshold T to determine which words are closely related to the query word. An increase in T means that only very similar words will be included; a decrease in T means that more remote words will be included.

The original query word is then aligned to a neighbourhood word and the BLAST algorithm then extends the alignment in both directions while scoring for matches, mismatches and gaps. The length of this alignment is determined by the number of positions aligned versus the cumulative alignment score. The alignment extension continues until the number of mismatches starts to decrease the cumulative score (Fitzpatrick and Tovar, 2011).

For this work BLAST was used to search for cyanobacterial homologues to the AS that were reconstructed. In each instance only the top two scores (as determined by E-value, see box on right for definition) were used to further explore in the PFAM database.

> Each alignment has a bit score (*S*), which is a measure of similarity between the hit and the query. The E-value of a hit is the number of alignments with bit score $\geq S$ that you expect to find by chance (i.e., with no evolutionary explanation)

PFAM

Pfam (Finn *et al.*, 2010, Punta *et al.*, 2011, Sonnhammer *et al.*, 1997) is a database of protein families, where families are sets of protein regions that share a significant degree of sequence similarity, thereby suggesting homology. Pfam contains two types of families: high quality, manually curated Pfam-A families and automatically generated Pfam-B families. Pfam-A families are built following what is, in essence, a four-step process:

1. building of a high-quality multiple sequence alignment (the so-called seed alignment);
2. constructing a profile hidden Markov model (HMM) from the seed alignment;
3. searching the profile HMM against the UniProt Knowledgebase (UniProtKB) sequence database;
4. choosing family-specific sequence and domain gathering thresholds (GAs); all sequence regions that score above the GAs are included in the full alignment for the family. The gathering thresholds, or GAs, are manually curated ,family-specific, bit score thresholds that are chosen by Pfam curators at the time a family is built (Punta *et al.*, 2011).

PFAM was used to compare the protein families from within the AS with the protein families within the cyanobacterial sequences. If the protein families from the AS and the cyanobacteria matched then this was taken to support the homology of that enzyme between cyanobacteria and archaeplastida.

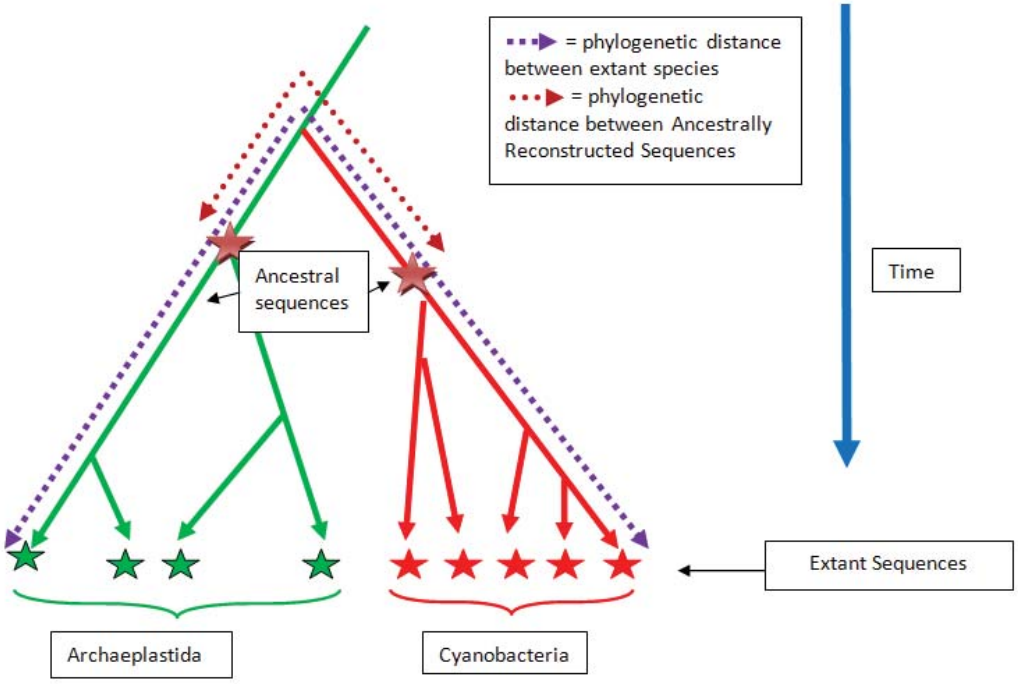
ASR (Ancestral Sequence Reconstruction)

In genome comparisons of distantly related species, it is often difficult to identify homologous protein coding regions that are highly diverged. The ASR program accepts this problem by inferring the ancestral sequence for a group of homologous sequences from a tree. It aligns these sequences and estimates the most likely amino acid or nucleotide for each position at each node in that tree (Collins *et al.*, 2003). The approach is robust to slight changes in the tree (Hanson-Smith *et al.*, 2010).

Ancestral sequences were calculated using the Baseml program in PAML (Yang, 2007) that had been customized into a Perl script. This program uses the marginal reconstruction approach, using most-parsimonious likelihood that compares the probabilities of different amino acids for an interior node at a site and then selects the nucleotide that has the highest posterior probability.

In this instance, using the members of the archaeplastida, listed in Table 2.1, ancestral sequences of the enzymes that catalyse reactions in the essential metabolic pathways listed in Table 2.2 were reconstructed. Similarly, the ancestral sequence of cyanobacteria could also be inferred. This reduced the phylogenetic distance to the homologs within cyanobacteria, as demonstrated in Figure 2.3 below.

Figure 2:3-Demonstrates the shorter phylogenetic distances when using ASR –distances are exaggerated for illustrative purposes.



Method

Identification

Nucleotide sequences for the 12 archaeplastida identified in Table 2.1 were collated from the KEGG database for each of the enzymes catalysing reactions in the 18 metabolic pathways shown in Table2.2**. KEGG nucleotide and amino acid sequences are typically CDS (Coding DNA Sequence) however some are RNA (Ikeuchi, 2015)**. An example of the pathway and enzymatic steps is shown in Figure 2.2. Each step in the pathway, typically identified with an E.C. (Enzyme Commission) number, for which data was available, resulted in a collection of sequences that were stored in fasta format.

For some organisms, there was a choice of sequences. The longest variant was chosen in these cases providing that the title of the fasta sequence did not include the term “chloroplastic-like”. Lengths and number of variants were noted.

1st Assessment

For each sequence, KEGG gives an ortholog function. This was executed and the top 20 orthologs were selected for a multiple sequence alignment (MSA), using CLUSTALW2 (Larkin *et al.*, 2007a) and a rooted neighbour-joining tree was constructed. All of these operations were available through the

KEGG database. The taxonomy of each of the 20 organisms was noted for future reference. Of particular interest were orthologs that derived from Cyanobacteria.

ASR generation

Ancestral sequences were then generated from the fasta files for each enzyme. This involved many steps –

1. The header line in the fasta files could be no longer than ten spaces. A custom Perl program was run to shrink header values to 10 spaces.
2. the ASR perl script requires

- a MSA, generated using CLUSTALW2 (Larkin *et al.*, 2007b) , found at http://www.ebi.ac.uk/Tools/msa/clustalw2/
- a treefile, generated using Splitstrees3.2 (Huson and Bryant, 2006) from the MSA - (a second custom Perl script was used to change the format of the MSA from .paml to .phylip)
- a control file that sets up some of the parameters (see appendix 1)

3. the ASR perl script was modified from the Baseml program in PAML (Yang, 2007) by W.T.J. White (2011) – see appendix 1

2nd Assessment

1. The AS were initially run through Blastn to assess whether the sequences that had a high level of cyanobacterial homology, as indicated by KEGG, were indicating a high level of cyanobacterial homology from Blastn. Blastn results were limited to the bacterial kingdom.
2. The completed AS were also run through the PFAM database (Finn *et al.*, 2010, Punta *et al.*, 2011, Sonnhammer *et al.*, 1997) to assess whether the ASR program was constructing functional protein families within the AS.

Results

KEGG

In the 18 different metabolic pathways, there were 161 enzymes that were analysed; see appendix, table 1. Of these 161 enzymes, 42 (26.1%) of these had some cyanobacterial homology with archaeplastida; 17 (10.6%) had a total cyanobacterial homology with archaeplastida, that is, all the homologs that emerged for that individual enzyme outside of archaeplastida were from cyanobacteria. Alphaproteobacterial homologs were found for 19 of the 161 enzymes; there were no Rickettsiales homologs, the bacterial family from which mitochondria is thought to have originated. There being no Rickettsiales homologs is an interesting finding. Does this imply that the mitochondrial signature is no longer apparent in plant genomes or was it there in the first instance? With the tools that we currently have there is no definitive answer to this question.

PFAM

161 putative ancestral enzymes that were constructed through ASR were analysed with PFAM; 73 of these returned a positive result with a homology being found within the database (see appendix, table 2 for results). Many of the enzymes returned more than one protein family within the database, the 73 enzymes giving 119 protein families. Of these 119 only three enzymes found homology within the PFAM B (computer generated designation rather than the curated PFAM A designation); 49 of the results did not belong to a clan but did to a family (clans are groups of related families) and for 31 of the results the "family" designation was judged to be non-significant.

KEGG identified that 42 of the 161 enzymes analysed had some level of cyanobacterial homology; of the 161 AS constructed 45.4% of these sequences had some level of PFAM recognition and maybe seen as successful reconstructions. There was no identifiable correlation between organisms showing cyanobacterial homology (KEGG) and AS displaying protein family recognition (PFAM).

Discussion

During the course of this initial experiment, a number of problems presented themselves; these are listed below. All but one of these problems has to do with the nature of the bioinformatic programs used in this study.

Enzyme Variants

When searching KEGG metabolic pathways for enzymes, in several instances there are multiple variants of the same gene, that code for the enzyme, within an organism (common in plants). It is generally accepted that there were two rounds of whole genome duplication in the lineage leading to flowering plants (De Bodt *et al.*, 2005). Possible causes of this variation are-

1. The enzyme is involved in multiple processes -e.g. EC 1.2.1.3, an enzyme found in multiple (16) metabolic pathways in Table 2.3 below – and as such has 6 coding regions in *Arabidopsis thaliana*, 11 in *Populus trichocarpa*, etc.

Table 2-3-Tables listing both the number of EC 1.2.1.3 variants per organism (left-hand side) and the variety of pathways EC 1.2.1.3 is involved in (right-hand side).

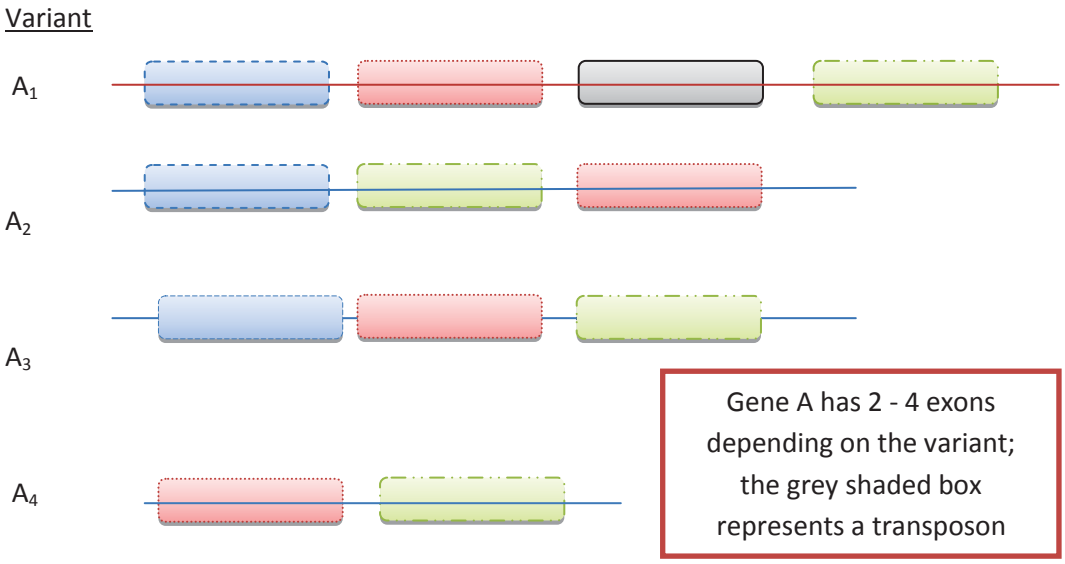
Organism	Variants
Arabidopsis thaliana	6
Populus trichocarpa	11
Ricinus communis	4
Vitis vinifera	9
Oryza sativa japonica	7
Sorghum bicolor	4
Zea mays	9
Physcomitrella patens subsp. patens	7
Chlamydomonas reinhardtii	1
Volvox carteri f. nagariensis	0
Ostreococcus lucimarinus	1
Ostreococcus tauri	1
Cyanidioschyzon merolae	1

Metabolic pathways for EC 1.2.1.3
Glycolysis / Gluconeogenesis
Pentose and glucuronate interconversions
Ascorbate and aldarate metabolism
Fatty acid metabolism
Valine, leucine and isoleucine degradation
Lysine degradation
Arginine and proline metabolism
Histidine metabolism
Tryptophan metabolism
beta-Alanine metabolism
Glycerolipid metabolism
Pyruvate metabolism
Propanoate metabolism
Limonene and pinene degradation
Metabolic pathways
Biosynthesis of secondary metabolites

2. The genes coding regions are made up of multiple exons that in some instances are rearranged, giving different length genes coding for the same enzyme, illustrated below in Figure 2.4 –

When several variant copies were available while searching for homologies, the longest variant was used for further analysis. The longest variants were chosen because of the possibility of additional protein families being identified.

Figure 2-3 Illustrating exon rearrangements leading to differing gene lengths. The differing dash-arrangement surrounding the boxes represents different exons in different orders.



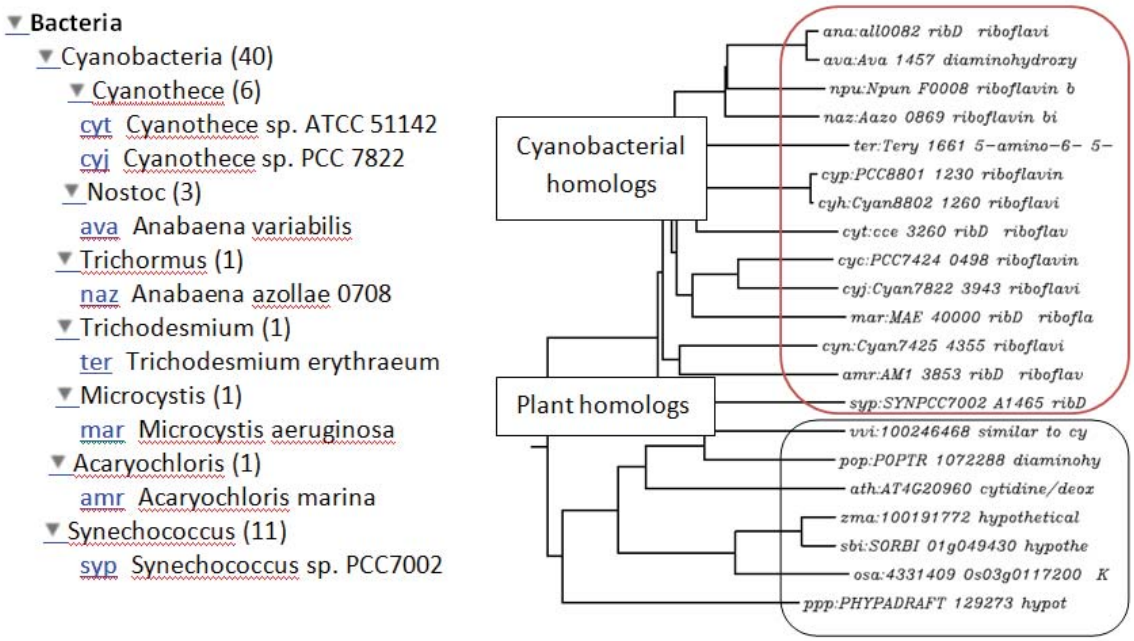
KEGG vs. Blastn

KEGG has an option, after selecting the enzyme of interest on a metabolic map, to search for orthologs in its database. These orthologs can be run through CLUSTALW and a Neighbour Joining tree produced. Initial searches for cyanobacterial orthologs were discovered using this process.

The problem using the KEGG ortholog function is that it searches for orthologs using both sequence similarity and pathway resemblance (Kanehisa and Goto, 2000); Blastn uses only sequence similarity. This difference between the databases leads to disparity in the orthological results.

For example, the enzyme EC 3.5.4.25 found in *Arabidopsis thaliana*, (Riboflavin metabolic pathway, http://www.genome.jp/kegg-bin/show_pathway?map00740) has all its bacterial orthologs (KEGG parameter set at 20 species) in the cyanobacterial family as demonstrated in Figure 2.5 below, which implies that this enzyme probably came from cyanobacteria.

Figure 2-4 - Lists of archaeplastida cyanobacterial homologs from KEGG for E.C. 3.5.4.25 from the Riboflavin pathway.



ASR is then used to find the inferred ancestral sequence for the 13 archaeplastida, only seven of which are shown in Figure 4.5. Once the ASR has been run and the ancestral sequence blasted- (BLASTn, limited to the bacterial domain) results for the ancestral enzyme return different bacterial families- see Table 4.4 below for top 30 bacterial homologs- none of which belong to cyanobacteria. To further confound proceedings, many ancestral enzyme sequences that BLAST results show having cyanobacterial homologs have no corresponding cyanobacterial homology in the KEGG results, illustrated in Table 2.4 below. Thus the difference between KEGG and BLAST results is the nature of the problem. Reasons for this difference stem from the different methods used between BLAST and KEGG for identifying homologies, discussed earlier in the chapter.

Table 2-4-List of AS archaeplastida homologs from BLAST; no cyanobacterial homologs were found in the first 30 bacterial homologs

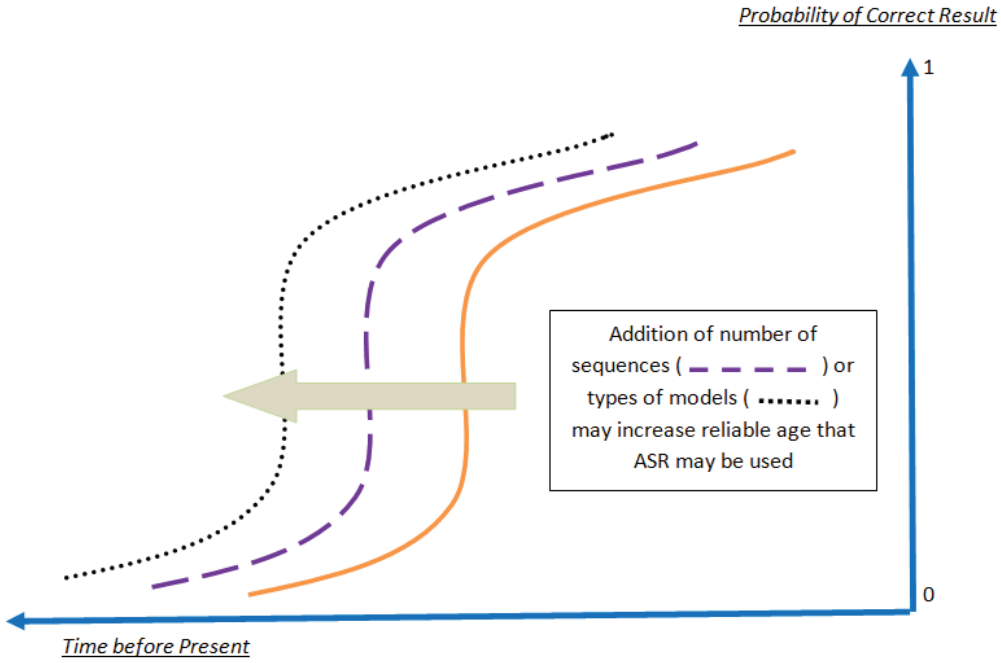
Description	Max score	Total score	Query coverage	E value	Max ident
Thermomonospora curvata DSM 43183, complete genome	286	412	59%	6e-74	87%
Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111, complete ge	273	355	47%	4e-70	85%
Streptomyces scabiei 87.22 complete genome	266	307	76%	6e-68	96%
Myxococcus fulvus HW-1, complete genome	248	373	53%	2e-62	100%
Thermobifida fusca YX, complete genome	244	244	47%	2e-61	71%
Thermobispora bispora DSM 43833, complete genome	230	230	46%	4e-57	71%
Actinosynnema mirum DSM 43827, complete genome	230	401	82%	4e-57	85%
Myxococcus xanthus DK 1622, complete genome	226	308	57%	5e-56	96%
Streptomyces coelicolor A3(2) complete genome; segment 5/29	223	223	65%	6e-55	68%
Catenulispora acidiphila DSM 44928, complete genome	210	335	62%	4e-51	96%
Nocardioides sp. JS614, complete genome	206	206	60%	5e-50	68%
Streptomyces davawensis putative RNA polymerase ECF-subfamily sig	199	199	51%	7e-48	69%
Streptomyces venezuelae ATCC 10712 complete genome	197	363	57%	2e-47	92%
Streptomyces avermitilis MA-4680 DNA, complete genome	187	231	53%	4e-44	96%
Streptomyces ambofaciens ATCC 23877 left chromosomal arm	187	187	76%	4e-44	66%
Azospirillum sp. B510 DNA, complete genome	183	183	50%	5e-43	68%
Streptomyces griseus subsp. griseus NBRC 13350 DNA, complete gen	183	265	56%	5e-43	80%
Streptosporangium roseum DSM 43021, complete genome	181	181	46%	2e-42	69%
Frankia sp. EAN1pec, complete genome	181	227	55%	2e-42	84%
Streptomyces sp. SirexAA-E, complete genome	178	218	56%	2e-41	80%
Geobacter sp. M18, complete genome	178	178	31%	2e-41	72%
Magnetospirillum magneticum AMB-1 DNA, complete genome	178	178	34%	2e-41	71%
Pseudomonas stutzeri A1501, complete genome	176	217	29%	8e-41	93%
Pseudonocardia dioxanivorans CB1190, complete genome	174	376	57%	3e-40	84%
Thermus scotoductus SA-01, complete genome	174	174	32%	3e-40	72%
Amycolatopsis mediterranei S699, complete genome	172	254	56%	1e-39	88%
Amycolatopsis mediterranei U32, complete genome	172	254	56%	1e-39	88%
Pseudomonas stutzeri ATCC 17588 = LMG 11199, complete genome	170	213	29%	3e-39	84%
Frankia sp. CcI3, complete genome	170	211	51%	3e-39	87%

An interesting point is made by Kurland in an unpublished paper – he argues that three-dimensional structures of orthologous proteins would be conserved over long evolutionary distances while the corresponding coding sequences (i.e. sequence DNA) could and would alter extensively in an erratic unrecognizable pattern and as a consequence of these fluctuations only a small minority of the homologous coding sequences could be recognized by BLAST (Kurland and Harish, 2013, unpublished). Inferred 3D structures are addressed in chapter 6.

Markov Models

Markov models (Mossel and Steel, 2005b) are now standard for modelling the evolution of aligned genetic sequence data . Markov models are the algorithms used to construct the substitution/ insertion/deletion rates in phylogenetic programs such as in PFAM, ASR and CLUSTAL. The main criticism of Markov models is that at deeper divergence times there is an exponential drop-off in the reliability/accuracy of the sequences produced. This is the point at which the probability of a substitution (generated by the Markov model to represent the evolution of the nucleotide or codon) throughout the length of the branch of the phylogenetic tree passes a certain critical value (Mossel and Steel, 2005a).

Figure 2-5-Reliability of Markov modelling with addition of other models.
As more sequences or combination of models are added, there is an increase in the time (before present) that the Markov model may be reliably used, as illustrated by the large arrow. Any increase in reliability is linear (addition of sequences) versus an exponential decay (artefact of Markov modelling) so that the gain from the addition of sequences or model combinations is marginal.



To counterbalance this there is a linear increase in reliability with an increase in the number of sequences used; there is a suggestion that the reliability may also be increased by using a combination of models (Mossel and Steel, 2005a, Sober and Steel, 2002) as illustrated in Figure 2.6 above. However, this paradigm will only make a marginal difference as the dynamic is one of linear gain versus exponential decay. This dynamic has not been explicitly stated in the field of evolutionary genetics; rather it is a principle that was proven in the setting of card shuffles that should apply to phylogenetics (Steel and Penny, 2013, pers.comm.).

The tan arrow in Figure 2.6 indicates the cumulative effect of using a larger amount of sequences and more models (3D and PFAM) on the reliability of Markov modelling sequence construction at earlier times than was possible using just Markov modelling (exaggerated for illustration purposes because the effect is marginal).

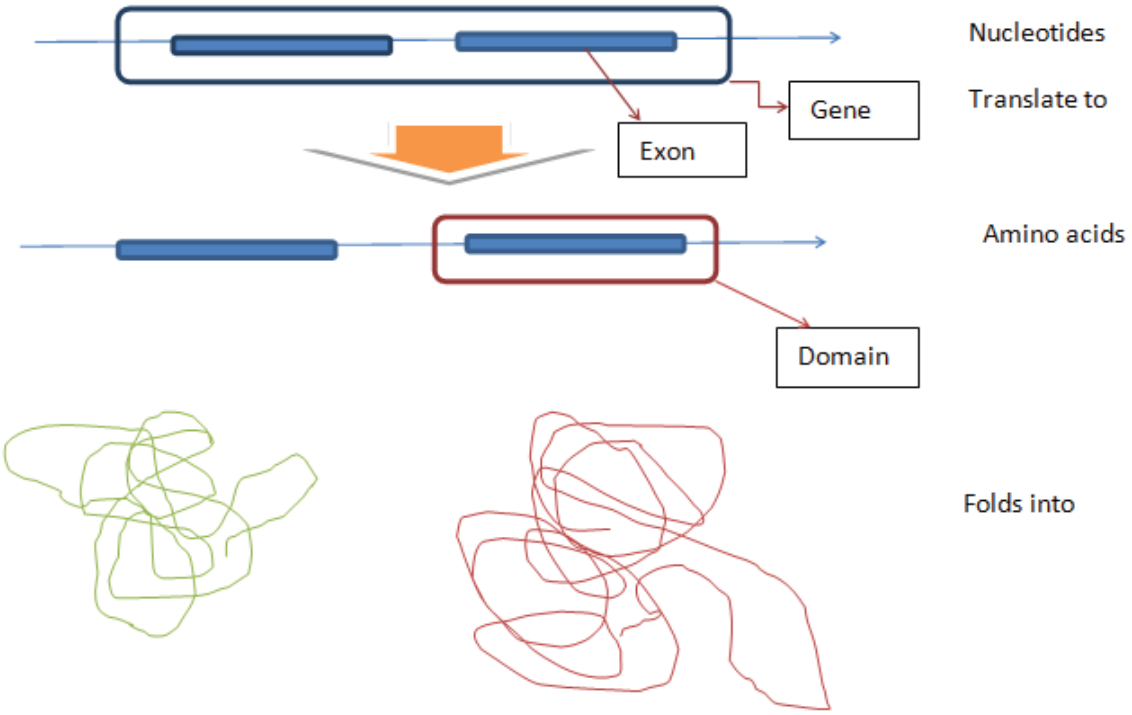
Rationale for the next experiment

Nucleotides versus Amino acids

When the ASR is blasted against cyanobacteria, it is analogous to looking at a haystack and deciding whether you have wheat (good data) or chaff (bad data). The point of searching with amino acids

rather than nucleotides is that the translated protein is 3-dimensional and needs to be folded in the right way for successful operation, illustrated below in Figure 2.7. Using protein BLAST, which builds homologs on amino acid sequence identity rather than nucleotides may effectively separate the wheat from the chaff.

Figure 2-6-From nucleotide to amino acid to domain to motif



For the two proteins to interact, illustrated above in Figure 2.7, they may need motifs that are found in the amino acid sequence that allow them to bind to each other even though the underlying nucleotide sequence may be different.

Thus, for the next part of this thesis amino acids were used rather than nucleotides.

Summary

The method and the nature of the data for the analysis of nucleotides are outlined. KEGG, BLAST and PFAM are introduced and a quick explanation given of how these websites work. An outline of the ASR program is discussed. Of the 161 enzymes analysed using the KEGG database, 42 had some level of cyanobacterial homology and 19 had some level of α-Proteobacterial homology but interestingly no Rickettsiales homology.

Less than half of the AS that were generated returned a homology with a protein family (PFAM) and interestingly the BLAST results showed no homology with cyanobacterial proteins for the top 20 hits returned. The reasons for this are discussed and include –

- Enzyme variants
- The differing methods and rationale behind these method between BLAST and KEGG
- The decreasing reliability of Markov modelling at deeper time periods.

This chapter ends with the rationale for the next chapter being presented.

Chapter 3. Amino Acid Approach

Background

This thesis has been about looking for patterns- in particular a pattern of gene retention from an endosymbionts nucleus to the host's nucleus for the biosynthesis of some essential metabolites namely vitamin and amino acid synthesis. This has been attempted with plants, specifically the transfer of chloroplast genes (accepted to have been a Cyanobacterium) to the plant nucleus, in order to establish a pattern that may then be used with all eukaryotes, specifically the transfer of genes from the mitochondrial endosymbiont (accepted to have been an alpha-proteobacterium) to the eukaryote nucleus. This has been expounded upon at length in chapter 1.

One of the tools that I thought to be capable of strengthening any cyanobacterial signal was the use of the standard Ancestral Sequence Reconstruction (ASR) program using nucleotides. In genome comparisons of distantly related species, it is often difficult to identify homologous protein coding regions that are highly diverged. The ASR program accepts this problem by inferring the ancestral sequence for a group of homologous sequences from an inferred tree.

The idea was that because the chloroplast endosymbiotic event happened later than the mitochondrial endosymbiotic event the signal of any chloroplast genes that were transferred to the plant nucleus would be stronger. The hope/expectation was that the ASR program could identify more chloroplast genes in the plant nucleus and that this technique, or pattern, would be able to be applied to the discovery of mitochondrial genes in all eukaryotic genomes.

Chapter 2 used nucleotides for this analysis; the following results and discussion in this chapter use amino acids for the analysis. At this stage of the thesis a series of checks were initiated in order to verify ASR efficacy by using three different bioinformatics programs. The idea was to build a case using the results from these three different programs to argue the effectiveness of the ancestral reconstructions. This process was planned from a biologist's point of view; in a following chapter, chapter 5, ASR effectiveness is analysed from a statistician's point of view.

Method

Indication of Potential Archaeplastida Enzymes that have a Cyanobacterial Homology

In chapter 2, 161 enzymes were analysed for potential cyanobacterial homology using KEGG. Each enzyme was run through a CLUSTAL program and the top 20 orthologs from all Kingdoms of life examined and listed as a Neighbour joining tree. This process is outlined in more detail in chapter 2. 42 enzymes had some level of cyanobacterial homology. Inclusion for further analysis was dependent upon at least four of the homologs generated by the KEGG homolog function being cyanobacterial. Some of these 42 enzymes showed a total orthology with cyanobacteria with all of the orthologs outside archaeplastida being cyanobacterial, while most of these 42 enzymes' orthologs showed only a partial orthology with cyanobacteria.

42 of these enzymes that indicated cyanobacterial homology were gathered for further analysis in this chapter. Archaeplastida nucleotide sequences were re-collated as amino acids. Additionally, ten cyanobacteria were selected for analysis using ASR (these ten cyanobacteria were the only cyanobacteria available on the KEGG database at the time I started). Amino acid sequences where collated for these organisms from the same positions (as indicated by E.C. numbers) in the same metabolic pathways as the 42 archaeplastida that indicated cyanobacterial homology.

Ancestral Sequence Reconstruction

ASR was run on these enzymes. Amino acids from 12 archaeplastida and 10 cyanobacteria (listed below in table 3.1) were used to construct the relative ancestral sequences. Two AS were constructed for each enzyme, one for the archaeplastida and one for cyanobacteria. Different rate matrices were sampled in order to determine the optimal matrix.

Table 3-1-List of cyanobacterial species

Cyanobacteria Species	KEGG ID	Organism Number	Assembly Number
Gloeobacter violaceus	gvi	T00148	GCA_000011385.1
Synechococcus sp. JA-3-3Ab	cya	T00318	GCA_000013205.1
Synechococcus sp. JA-2-3B'a(2-13)	cyb	T00319	GCA_000013225.1
Cyanothece sp. PCC 7425	cyn	T00836	GCA_000022045.1
Cyanothece sp. PCC 7424	cyc	T00808	GCA_000021825.1
Anabaena variabilis	ava	T00277	GCA_000204075.1
Nostoc punctiforme	npu	T00713	GCA_000020025.1
Synechococcus elongatus PCC6301	syc	T00163	GCA_000010065.1
Prochlorococcus marinus MIT 9303	pmf	T00466	GCA_000015705.1
Prochlorococcus marinus MIT 9215	pmh	T00598	GCA_000018065.1

The same method of ASR was used as described in chapter 2 except that the codeml function was used instead of the baseml, adjusted in the control file.

PFAM – a first check into ASR effectiveness

Each Ancestral Sequence was run through PFAM (Finn *et al.*, 2010) to ensure that the same protein families were being identified as those from the KEGG database (Kanehisa *et al.*, 2008), they were. Next *A.thaliana*, *S.elongatus* and the AS for each enzyme were run through PFAM and an E-score noted to detect if the ancestral sequences were providing lower E-scores as a measure of the effectiveness of the ASR. See Table 3.2 below.

BLAST- a second check into ASR effectiveness

BLAST (Altschul *et al.*, 1990) was used to assess whether the same phyla of organisms were appearing for the ancestrally constructed enzymes as compared with enzymes from *A.thaliana*.. Blastp (a BLAST search using amino acids as input) searches were used to assess the phyla of the two organisms. See table 3.3 below.

Explicitly, *A.thaliana* enzymatic sequences were collated for each position in the Niacin and Isoleucine biosynthetic pathways. These sequences were run through BLASTp, limited to the bacterial kingdom, and the top ten homologs for each pathway position were noted for their phylum. These results correspond to the grey columns in table 3.3. Similarly ancestral enzymatic sequences from the same pathways were collated and run through BLASTp with phyla homology noted. This corresponds to the white columns in table 3.3. So for each row, all the figures in the grey columns add to ten as do all the figures in the white columns.

UGene- a third check into ASR effectiveness

UGene (Okonechnikov *et al.*, 2012) was used to provide simple similarity scores for the Ancestral Sequence and the 23 plant and cyanobacterial sequences for each particular enzyme. The plant Ancestral Sequence was expected to have a higher similarity to the cyanobacteria than the extant plant sequences if ASR was effective. See Tables 3.4, 3.5 and 3.6 below.

Results

Analysis of Protein Families – 2 samples.

Comparisons are made between columns 3 & 4 (archaeplastida A.S. & *A.thaliana*) and 5 & 6 (cyanobacteria A.S. & *S.elongatus*) in Table 3.2 below. In both cases, if the value of the Ancestral Sequence is less than the value of the organism, then ASR may be seen as being sucessful. E-values were determined by PFAM. PFAM describes an e-value as the number of hits that would be expected to have a score equal to or better than this value by chance alone. A good E-value is much less than 1. A value of 1 is what would be expected just by chance.

Table 3-2: Analysis of Protein Family E-scores in the Isoleucine and Niacin Pathways compared to the Ancestral Sequence E-scores

E.C .Number		Archaeplastida A.S.	*A.thaliana*	Cyanobacteria A.S.	*S.elongatus*
Isoleucine					
LeuB (1.1.1.85)	Iso DH	4.40E-141	4.40E-141	7.00E-138	1.10E-135
4.3.1.19	Thr dehydrat C	No hit	3.30E-72	no hit	6.30E-26
	PALP	7.90E-32	8.40E-23	4.00E-81	7.50E-81
1.2.4.1	E1 dh	1.00E-37	6.50E-98	2.50E-103	7.40E-105
2.2.1.6	TPP enzyme N	6.00E-26	3.40E-55	1.60E-58	1.50E-58
	TPP enzyme M	1.40E-29	4.40E-40	2.80E-50	7.90E-50
	TPP enzyme C	No hit	4.00E-46	7.70E-52	4.30E-53
1.1.1.86	Ilv N	6.80E-37	2.60E-32	2.30E-78	1.10E-77
	Ilv C	7.20E-23	5.90E-32	1.80E-58	3.50E-57
4.2.1.9	ILVD EDD	1.80E-210	6.80E-208	7.20E**n**-216	1.90E-196
2.6.1.42	Aminotran 4	3.00E-30	1.20E-33	1.30E-47	8.30E-46
Niacin					
1.4.3.16	FAD binding 2	6.70E-95	2.80E-88	9.60E-76	7.40E-91
	Succ DH Flav C	9.70E-20	6.50E-20	1.80E-10	2.60E-16
NadA	NadA	3.00E-27	1.00E-29	7.00E-116	1.90E-119
	SurE	3.60E-22	8.50E-18	not in Cy	not in Cy
2.4.2.19	QRPTase N	1.10E-22	8.00E-21	3.90E-23	6.30E-20
	QRPTase C	9.70E-50	2.30E-62	3.40E-57	3.40E-60
3.6.1.22	NUDIX	2.50E-19	9.70E-19	not in Cy	not in Cy
	NUDIX-like	No hit	1.20E-08	not in Cy	not in Cy
	zf-NADH-PPase	No hit	9.00E-08	not in Cy	not in Cy
	Stirrup	No hit	0.079	not in Cy	not in Cy
3.1.3.5	SurE	2.40E-48	4.20E-48	1.20E-65	8.60E-63
2.7.7.18	CTPF trans 2	2.80E-24	5.30E-17	2.10E-30	4.00E-14
2.7.7.1	CTPF trans 2	2.80E-24	5.30E-17	only 2 Cy	4.90E-13
2.4.2.11	NAPRTase	1.10E-11	2.10E-15	4.30E-14	3.70E-15
	NAPRTase	4.90E-24	1.10E-28	No hit	No hit
6.3.5.1	CN Hydrolase	3.40E-30	7.20E-23	2.40E-08	4.50E-11
	NADsynthase	7.30E-25	2.80E-24	2.00E-74	2.00E-76
2.7.1.23	NADkinase	1.00E-55	5.50E-62	3.40E-50	1.00E-54

There was no hit for *S.elongatus,* *S.spJA-3-3b* used as a replacement

In 22 out of the 45 protein domains analysed above in Table 3.2, the ancestral sequence has a lower e-value than the extant sequence. Most of these differences between e-values are relatively very small in the other instances and it may be argued that ASR has worked to some degree although not nearly as effectively as hoped. For example the first two values for enzyme 1.1.1.85 are both 4.4E-

141 so there is no advantage in using the AS however the values for Cyanobacteria versus *S.elongatus* are 7E-138 and 1.1E-135 so here there is some advantage.

BLAST analysis of Ancestral Sequence and A. thaliana to bacterial Phylum – 2 samples

From Table 3.3 below, it is apparent that there is some level of agreement between the Ancestral Sequences and *A. thaliana*; however, for Isoleucine enzymes that have a total homology with cyanobacteria using *A. thaliana* as input, this pattern is not repeated when the Ancestral Sequences are used as input, in two out of the three cases (highlighted in green). Therefore, there is not a high level of confidence in the accuracy and efficacy of the Ancestral Sequence Reconstruction program in these instances but again there is some improvement just not as much as expected.

Table 3-3- Comparison of BLAST results to Phyla for the Ancestral Sequence and *A. thaliana*. Each row contains an enzyme and each pair of columns a bacterial phylum. The grey columns hold the BLASTp homology results for the enzyme from A. thaliana; the white columns hold the BLASTp results from the Ancestral Sequence for that enzyme. For example, the BLAST results from A. thaliana EC 4.2.1.9 (highlighted in yellow) show that the top 10 results all come from cyanobacteria, while the BLAST results from the Ancestral Sequence EC 4.2.1.9 show that the top 10 results are spread across 3 phyla – 8 are Planctomycetes, 1 Bacteroidetes and 1 Gammaproteobacteria. This particular pattern may indicate that the cyanobacterial enzymes have changed during evolution.

Niacin	*Gammaproteobacteria*		*Betaproteobacteria*		*Epsilonproteobacteria*		*DeltaProteobacteria*		*AlphaProteobacteria*		*Firmicutes*		*Verrocomicrobia*		*Chlamydiae*		*Spirochaetes*		*Bacteroidetes*		*Cyanobacteria*		*Green Sulphur*		*Planctomycetes*		*Deinococcus-Th*		*Hyperthermophillic*		*Other*	
1.4.3.16			1				4						1					2	4	6				2								
NadA							1			1										2	9	7										
2.4.2.19																			9	10											1	
3.6.1.22		1							10	9																						
3.1.3.5					2	1		3			2	1			3	1			3	3								1			2	
2.7.7.18		6		1							2	2			3				3							1					2	
2.7.7.1		6		1							2	2			3				3	2						1					2	
2.4.2.11								5			4	3							6	2												
6.3.5.1							1	1					2	1					5	6					2	2						
2.7.1.23	1			1			6	4			1																			5	2	
Isoleucine																																
LeuB (1.1.1.85)																					10	10										
4.3.1.19		4	7	4																	3	1										1
1.2.4.1										10											10											
2.2.1.6												4	4	3											5	1			1		4	
1.1.1.86	1	1					1	2	1	1							2	2	5	1												
4.2.1.9		1																		1	10					8						
2.6.1.42							1	9			8																	1			1	

UGene Analysis using Muscle and Similarity scores- 2 Samples –Folate 3.5.4.26 & Leucine B

Tables 3.4 and 3.5 below list the 10 cyanobacteria and the Ancestral Sequences (AS) for each group respectively. The entries are symmetrical, so we will only consider the top half. Reading the top row gives the percentage similarity between the plant ancestral sequence and all other cyanobacterial sequences. Understanding whether the AS of plants are more similar to cyanobacteria is found by comparing the archaeplastida AS to the cyanobacteria and the cyanobacterial AS; if the AS's have a higher similarity to each other than the individual organisms then the ASR has been effective.

Leu B

Table 3-4: Result for Enzyme LeuB- a Simple Similarity Table constructed from a MUSCLE Alignment and expressed as a percentage. A score of 100% would indicate that the two sequences were identical.

	AncLeuBPlants	AncLeuBCyan	syc_syc249	cya_CYA_17	cyb_CYB_16	cyc_PCC742	cyn_Cyan74	gvi_gvip47	npu_Npun_R	ava_Ava_29	pmf_P9303_	pmh_P9215_
AncLeuBPlants	x	62%	63%	61%	61%	62%	60%	58%	64%	63%	57%	51%
AncLeuBCyan		x	72%	97%	95%	70%	73%	72%	72%	72%	68%	55%
syc_syc249			x	67%	69%	77%	77%	69%	81%	79%	69%	56%
cya_CYA_17				x	91%	67%	71%	70%	71%	70%	67%	55%
cyb_CYB_16					x	66%	70%	68%	69%	69%	67%	54%
cyc_PCC742						x	74%	67%	79%	81%	69%	56%
cyn_Cyan74							x	70%	83%	81%	69%	56%
gvi_gvip47								x	72%	70%	63%	50%
npu_Npun_R									x	91%	65%	57%
ava_Ava_29										x	65%	56%
pmf_P9303_											x	56%
pmh_P9215_												x

Folate 3.5.4.26

Table 3-5: Result for Enzyme 3.5.4.26- a Simple Similarity Table constructed from a MUSCLE Alignment and expressed as a percentage.

	AncFol3.5.4.16Pl	AncFol3.5.4.16Cy	syc_syc005	cya_CYA_27	cyb_CYB_01	cyc_PCC742	cyn_Cyan74	gvi_gvip22	npu_Npun_R	ava_Ava_19	pmf_P9303_	pmh_P9215_
AncFol3.5.4.16Pl	x	21%	31%	28%	27%	28%	28%	31%	26%	28%	17%	16%
AncFol3.5.4.16Cy		x	32%	30%	28%	33%	34%	32%	31%	35%	87%	80%
syc_syc005			x	71%	69%	67%	77%	68%	79%	64%	30%	29%
cya_CYA_27				x	86%	63%	66%	66%	62%	63%	29%	26%
cyb_CYB_01					x	64%	65%	68%	62%	63%	26%	25%
cyc_PCC742						x	62%	63%	63%	79%	31%	29%
cyn_Cyan74							x	71%	73%	64%	27%	27%
gvi_gvip22								x	69%	62%	28%	28%
npu_Npun_R									x	61%	27%	28%
ava_Ava_19										x	31%	29%
pmf_P9303_											x	71%
pmh_P9215_												x

Overall, the scores tend to indicate that the ASR program is not making a noticeable difference in establishing a pattern of higher similarity between archaeplastidal AS and cyanobacteria. It should be noted that in most instances, there were only one or two percentage points difference in similarity scores between the archaeplastidal AS and the cyanobacterial sequences.

Table 3.6 below is a summary of all results from simple similarity scores. Table 3.6 is read by looking at the ASR column- a score of 2/7 means that when the archaeplastida Ancestral Sequence was compared to the 7 cyanobacteria and their AS, the cyanobacterial AS had the highest similarity score in 2 of the 7 instances. In the other 5 instances these individual cyanobacteria had higher similarity scores to the archaeplastida AS. (The changing value of the denominator reflects the number of cyanobacteria that were known to have sequences for that particular enzyme).

Table 3-6: Summary of Simple Similarity results for all enzymes.

Vit/AA	EC No	ASR*	Vit/AA	EC No	ASR*
Riboflavin (B2)	1.1.1.93	2/7	Lysine	1.2.1.11	6/10
Riboflavin (B2)	3.5.4.25	8/9	Lysine	5.1.1.7	0/10
Thiamine(B1)	2.8.1.7	1/10	Methionine	1.2.1.11	6/10
Niacin(B3)	NadA	0/10	Methionine	2.1.1.14	2/2
Pyridoxine(B6)	4.2.3.1	2/9	Phenyalanine	1.4.3.21	4/5
Folates (B9)	3.5.4.16	0/10	Phenyalanine	4.3.1.24	1/2
Folates (B9)	2.6.1.85	3/6	Phenyalanine	6.2.1.12	na
Ascorbate(C)	1.13.99.1	0/1	Phenyalanine	2.1.1.104	3/3
Pantothenic acid(B5)	2.2.1.6	10/10	Threonine	4.3.1.19	0/8
Histidine	2.4.2.17	3/10	Threonine	4.2.3.1	2/9
Histidine	4.2.1.19	0/10	Threonine	1.2.1.11	6/10
Histidine	4.1.1.2	1/2	Tryptophan	4.2.1.20	1/10
Histidine	1.2.1.3	0/6	Tryptophan	4.1.3.27 CHO	10/10
Isoleucine	4.2.1.35	5/10		4.1.3.27 GAT	7/8
Isoleucine	LeuC	2/10	Tryptophan	4.1.1.48	7/10
Isoleucine	LeuD	3/10	Valine	4.2.1.35	5/10
Isoleucine	LeuB-1.1.1.85	0/10	Valine	LeuC	2/10
Isoleucine	4.3.1.19	0/8	Valine	LeuD	3/10
Leucine	2.3.3.13	1/10	Valine	LeuB-1.1.1.85	0/10
Leucine	4.2.1.33	0/10	Valine	4.3.1.19	0/8
Leucine	1.1.1.85	0/10			

Key

= This enzyme is duplicated in another pathway

= These were the only 2 enzymes when ASR gave the highest similarity for all 10 Cyanobacteria

Results Summary

It is apparent from these analyses that the ASR program has not fulfilled the expectation of establishing a pattern of strengthening links between endosymbiont and host genes involved in the essential metabolic pathways of amino acid and vitamin biosynthesis. In most instances there is little difference between the scores but not enough to establish a pattern of ASR efficacy; reasons why are discussed below. These questions are re-examined from a different perspective in chapter five.

Problems, Comments and Potential for further study

Problems and Comments

The major problem in trying to construct an Ancestral Sequence is that evolution is dynamic. The enzyme that catalyses a portion of the biosynthesis of a vitamin (or cofactor) in a plant may bear some similarity to the enzyme that holds the same role in a Cyanobacterium but it is reasonable to assume that different evolutionary pressures are acting on these pathways in the different organisms. This is especially pertinent when there are isozymes and different ordering of exons/introns. We know that enzymes transferred to the eukaryotic nucleus gain an intron/exon structure.

Trying to ascertain which isozymes, from up to 16 isozymes that fulfil the same function in a single organism, are on the same evolutionary trajectory has been a particularly challenging aspect of this project. An example of this situation is outlined in Table 3.7 below where the lengths of proteins ostensibly carrying out the same function are compared.

The table below illustrates some of the problems encountered. For this enzyme EC 4.2.1.35, there are 5 different isozymes that have been acknowledged- Leu1L, Leu1S, IPMI2, IPMI1 and IIL1 (Gruer *et al.*, 1997). For each isozyme, there are 1-3 different protein family domains or combinations thereof – IPMI1 has 2 different family combinations- Aconitase C and Aconitase 2N as well as Aconitase C and CReP N. Some of the enzymes have a single copy of this enzyme for each protein family domain combination while some have two. Also note the lack of consistency in protein families across the archaeplastida. It would be a separate project to try and follow the evolution of each form of the enzyme.

Is there any biological relevance to these different isoforms? In the instance below, all these forms contribute to the construction of Isopropylmalate isomerase (IPMI), an enzyme that catalyzes the stereospecific isomerisation of 2-isopropyl- L -malate(α -isopropylmalate) to 3-isopropyl- L -malate (β -isopropylmalate) via the formation of *cis* -dimethylcitraconate in Leucine biosynthesis (Drevland *et al.*, 2007). There are two classes of IPMI based on the forms of protein assembly. Fungal IPMI is a monomeric enzyme; bacterial and archaeal IPMI are heterodimeric enzymes. Both forms consist of large (large subunit, LSU)and small(small subunit, SSU) polypeptide chains that are coded for by two different genes in the leucine operon (Yasutake *et al.*, 2004). In *A.thaliana* IPMI are localised to the stroma of chloroplasts and have homology with the bacterial IPMI (He *et al.*, 2010). It appears that IIL1 codes for the LSU and IPMI1 & 2 code for the SSU; these SSU appear to be specialised in Brassica as these genes are also active in the Methionine pathway (Knill *et al.*, 2009).

There are two forms of IPMI that broadly conform to eukaryotes on one hand and archaea/bacteria on the other; these evolutionary trajectories would have been muddied with the advent of the chloroplast and I speculate that this is part of the reason for the wide array of isoforms. Different gene naming conventions may also play a part.

1). Multiple Isozymes

Table 3-7: EC 4.2.1.35- multiple isozymes, variants and PFAM protein domains that occur in 13 archaeplastida-enzyme lengths are expressed in the main body of the Table. For example *A.thaliana* has four different Aconitase enzymes that range in length from 222-509 amino acids (aa).

Enzyme	4.2.1.35								
Gene Name	LEU1L			LEU1S		IPMI2	IPMI 1		IIL1
PFAM Domains →	Aconitase	Aconitase	Aconitase	Aconitase_C	Aconitase	Aconitase_C	Aconitase_C	Aconitase_C	Aconitase
Organism ↓		Peptidase_S26	SpoVAD		TOBE	Amidase	Aconitase_2_N	CReP_N	DUF3440
Arabidopsis thaliana						256	253	222	509
Glycine max			502	245					
Populus trichocarpa			462	172					
Ricinus communis	621	510					254		
Vitis vinifera			508	253					
Oryza sativa japonica	514			257					
Sorghum bicolor				250	507				
Zea mays				249	505				
Selaginella moellendirffii	489			187					
Physcomitrella patens	518			177					
Chlamydomonas reinhardtii	487			214					
Volvox carteri f. nagariensis	487								
Cyanidioschyzon merolae	512						255		

2). Problem with PFAM and Enzyme Length

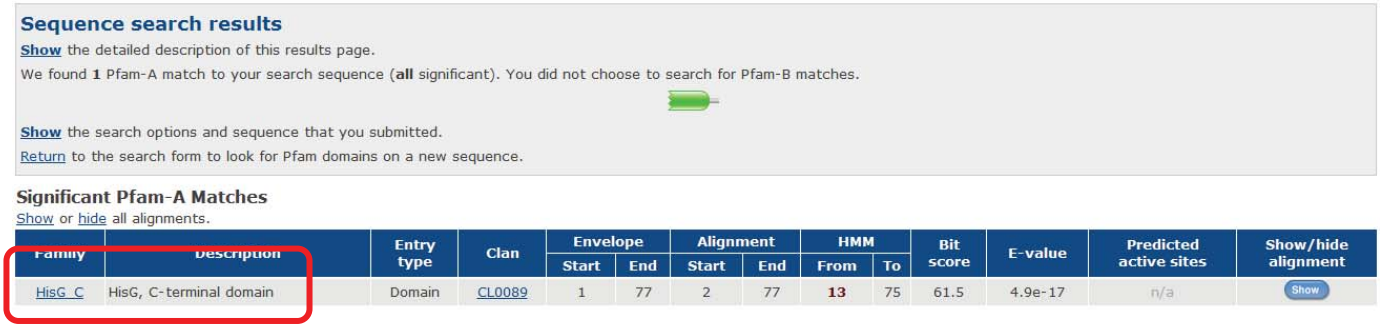
Most of the enzyme lengths for the enzyme EC 2.4.2.17, found in the Histidine pathway, are around 380aa; however for 2 archaeplastida the lengths are significantly shorter. When searching PFAM, both the long and the short sequences give some of the same families. Four of these enzymes, two shorter and two longer than average (380aa) from different archaeplastida are used to illustrate this below – maize, castor bean, *A.thaliana* and *C.merolae*- in Figures 5.1 – 5.4. It looks like HisG C (the enzyme used in the examples below) provides a regulatory function. In the following few pages I give some examples of the sequences from different plants, and the differences in length are apparent as

are the different versions of the HisG and HisG C families, identified in the red boxes. For each protein I give the actual amino acid sequence.

1. *Maize (94aa)*

>zma:MRGNSAEEVAERVLSQTSIWGLQGPTVSPVYRRRDGKVDVEYYAINVVVPQKLLYKSIQQLRSIGGSGVLVT KLTYIFDEETPRWRNLLSELG

Figure 3-1: PFAM result for Maize EC 2.4.2.17



Sequence search results

Show the detailed description of this results page.

We found **1** Pfam-A match to your search sequence (**all** significant). You did not choose to search for Pfam-B matches.

Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

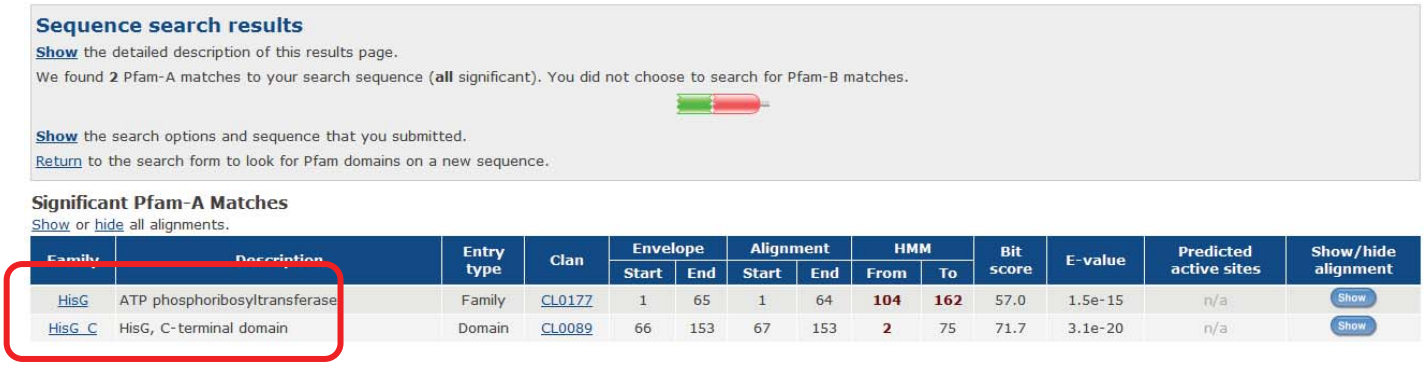
Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To				
HisG_C	HisG, C-terminal domain	Domain	CL0089	1	77	2	77	13	75	61.5	4.9e-17	n/a	Show

2. *Castor Bean (170aa)*

>rcu:MGIADAILDLVSSGTTLRENNLKEIEGGVVLQSQAILVASRKSLMQRKGALDTVHEILERLEAHLRAVGQFTVTANMRGS SAAEVAERVLSQPSLSGLQGPTVSPVFCKRDGKVTPDYYAIVICVPKKALYKSVQQLRAIGGSGVLVSPLTYIFDEETPRWRQLLS KLGL

Figure 3-2: PFAM result for Castor Bean EC 2.4.2.17



Sequence search results

Show the detailed description of this results page.

We found **2** Pfam-A matches to your search sequence (**all** significant). You did not choose to search for Pfam-B matches.

Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

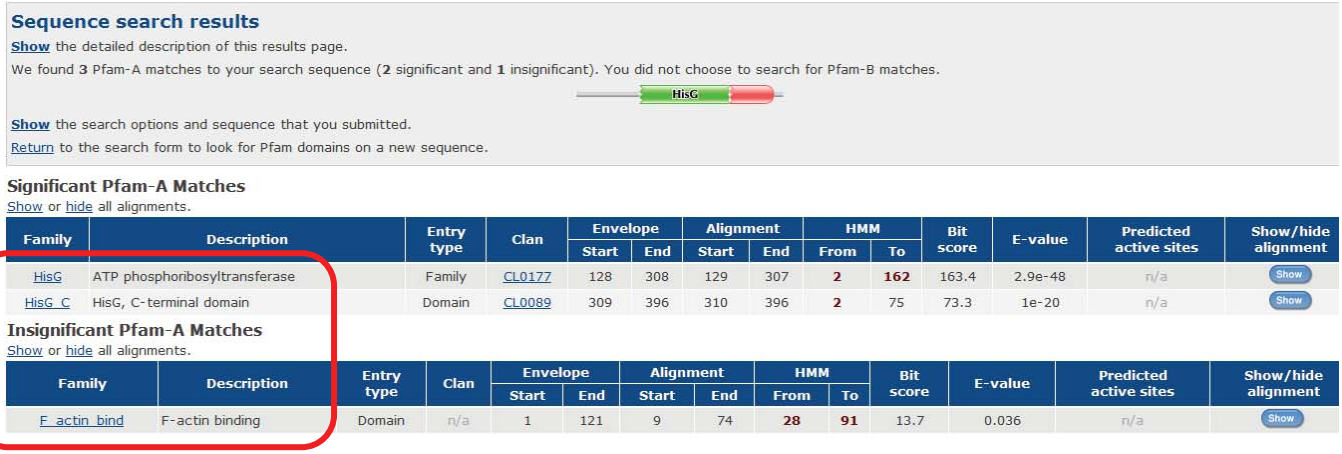
Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To				
HisG	ATP phosphoribosyltransferase	Family	CL0177	1	65	1	64	104	162	57.0	1.5e-15	n/a	Show
HisG_C	HisG, C-terminal domain	Domain	CL0089	66	153	67	153	2	75	71.7	3.1e-20	n/a	Show

3. *Arabidopsis thaliana (413aa)*

>ath:MPISIPLNATLQYSSPSSSSSSSSLVPSSPLFSPIPSTTVSLTGIRQRCLRMVTSCVSNAQKSVLNGATDSVSVVG REQIRLGLPSKGRMAADSLDLLKDCQLFVKQVNPRQYVAQIPQLPNTEVWFQRPKDIVRKLLSGDLDLGIVGLDIV GEFGQGNEDLIIVHEALNFGDCHLSLAIPNYGIFENIKSLKELAQMPQWTEERPLRVATGFTYLGPKFMKDNGIKHV TFSTADGALEAAPAMGIADAILDLVSSGTTLKENNLKEIEGGVVLESQAALVASRRALTERKGALETVHEILERLEAHL KANGQFTVVANMRGTDAEEVAERVKTQPSLSGLQGPTISPVYCKRDGKVTIEYYAIVICVPKKALYESVQQLRAVG GSGVLVSPVTYIFHEETPRWSQLLSNLGL

Figure 3-3: PFAM result for *A. thaliana* EC 2.4.2.17



Sequence search results

Show the detailed description of this results page.

We found **3** Pfam-A matches to your search sequence (**2** significant and **1** insignificant). You did not choose to search for Pfam-B matches.

HisG

Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To				
HisG	ATP phosphoribosyltransferase	Family	CL0177	128	308	129	307	2	162	163.4	2.9e-48	n/a	Show
HisG_C	HisG, C-terminal domain	Domain	CL0089	309	396	310	396	2	75	73.3	1e-20	n/a	Show

Insignificant Pfam-A Matches

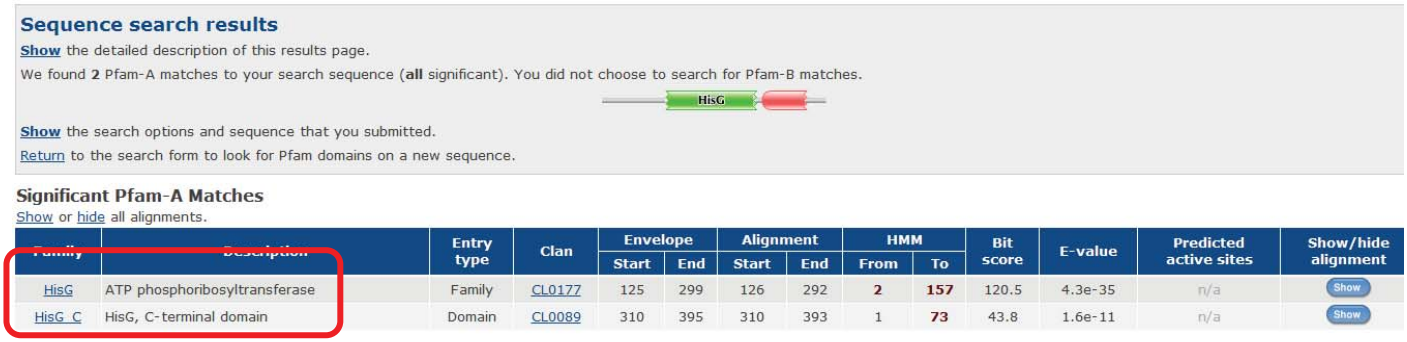
Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To				
F_actin_bind	F-actin binding	Domain	n/a	1	121	9	74	28	91	13.7	0.036	n/a	Show

4. *Cyanidioschyzon merolae (431aa)*

>cme:MAFVVQCSRSSIKERQRISTIVTSVRLRGGSCETRRCLRVRRAQGVRELRSCKVAASANLGNETTSLSADAEA
RDSIRLAIPSKGEIHQAAVELLNSCGLEVDLRNPRQYVGSIKHFPEVELWLQRPADIVRKVKDGTLDIGVAGYDLVAE
YNGTSTDVVIVHDSLGFGICRLGLGVPILWTDIHNIEDFRRFTESRSQPLRIVTKFPNQSEIFLAAHAIRNYRLLYQDG
ALEAATQLGTADCIIDLISSGVTLRENNLKEIPGGTILQSEMQLIGNRSRLAADQTRYPFAERLRDFVRELIERMDAHF
TAQQHYNVIANIRGESAGDVARRLGEFTDLRGLDGPTISTVIPPRGVAHGMYAIGLVVKKTQIYSAMKQLRRVGGS
GVCVLPVTFVFEGTTDRWKRLCEELQIPFENDEWKTMPRPFSSDKL

Figure 3-4: PFAM result for *Cyanidioschyzon merolae* EC 2.4.2.17



Sequence search results

Show the detailed description of this results page.

We found **2** Pfam-A matches to your search sequence (**all** significant). You did not choose to search for Pfam-B matches.

HisG

Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To				
HisG	ATP phosphoribosyltransferase	Family	CL0177	125	299	126	292	2	157	120.5	4.3e-35	n/a	Show
HisG_C	HisG, C-terminal domain	Domain	CL0089	310	395	310	393	1	73	43.8	1.6e-11	n/a	Show

So it appears that the enzymes for all these organisms are fulfilling the same role but with widely different lengths, see table 3.8 below, thereby upsetting the efficacy of the ASR for this enzyme. We thought that these different lengths cause misalignment when constructing the multiple sequence alignment, which is the first part of the ASR - generally, most alignment programs can cope with small indels but when they get too big, misalignment may occur (pers. comm. L. J. Collins, July, 2014).

Table 3-8- Summary of different lengths and identified protein families for organisms coding for EC 2.4.2.17. The average length for the 13 archaeplastida is 380 aa; these four organisms are the exception that may be stymieing the multiple sequence alignment.

Organism	Length (aa)	Families identified		
Maize	94	HisG C		
Castor bean	170	HisG C	HisG	
A.thaliana	413	HisG C	HisG	F Actin bind
C.merolae	431	HisG C	HisG	

This variation in the lengths of the enzymes does appear to be a major problem that must be addressed in the future; calculating the 3D structures using iTASSER (Zhang, 2008) is one possible avenue of investigation.

3). A note on PFAM identification, isozymes and the MUSCLE alignments

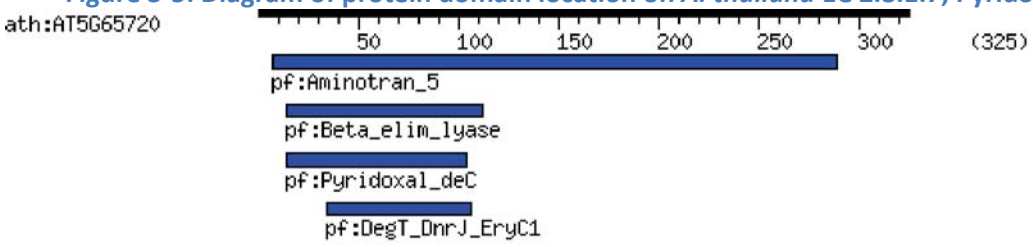
In the thiamine pathway for the enzyme that catalyses the reaction EC 2.8.1.7, *A. thaliana* has 2 isozymes-CPNIFS & NFS1- as illustrated below for both *A.thaliana* and Z.mays, being plants, and two cyanobacteria- *Synechococcus sp. JA-3-3Ab* and *A.variabilis.* As can be seen in the following tables and figures, there are a range of protein families that have been identified both within species and across species and kingdoms.

1. A.thaliana EC 2.8.1.7 - NFS1- Pyridoxal motif

Table 3-9- *A.thaliana* EC 2.8.1.7 Pyridoxal Motif (NFS1) detailing location and E-score of protein domains.

Motif id	From	To	Definition	E value
pf:Aminotran_5	7	289	Aminotransferase class-V	3.80E-62
pf:Beta_elim_lyase	14	112	Beta-eliminating lyase	2.90E-07
pf:Pyridoxal_deC	14	104	Pyridoxal-dependent decarboxylase conserved domain	2.20E-04
pf:DegT_DnrJ_EryC1	34	106	DegT/DnrJ/EryC1/StrS aminotransferase family	5.00E-07

Figure 3-5: Diagram of protein domain location on *A. thaliana* EC 2.8.1.7, Pyridoxal motif

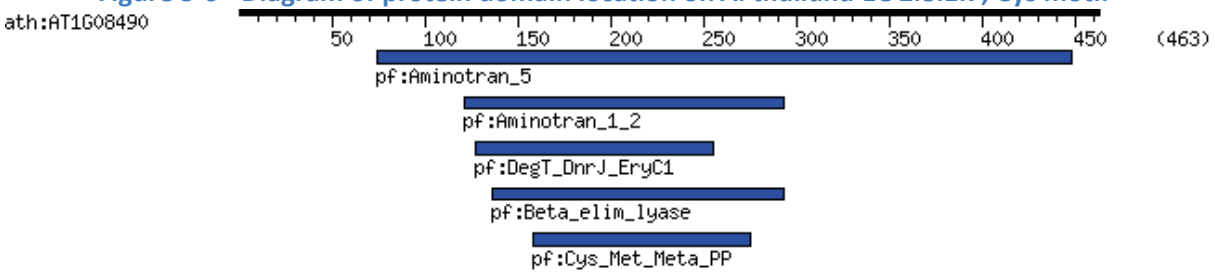


2. A.thaliana EC 2.8.1.7 - CPNIFS-Cys Motif

Table 3-10- *A.thaliana* EC 2.8.1.7 Cys Motif (CPNIFS) detailing location and E-score of protein domains

Motif id	From	To	Definition	E value
pf:Aminotran_5	74	448	Aminotransferase class-V	3E-144
pf:Aminotran_1_2	121	293	Aminotransferase class I and II	0.0075
pf:DegT_DnrJ_EryC1	127	255	DegT/DnrJ/EryC1/StrS aminotransferase family	6.5E-09
pf:Beta_elim_lyase	136	293	Beta-eliminating lyase	0.0011
pf:Cys_Met_Meta_PP	158	275	Cys/Met metabolism PLP-dependent enzyme	0.00000003

Figure 3-6 - Diagram of protein domain location on *A. thaliana* EC 2.8.1.7, Cys motif



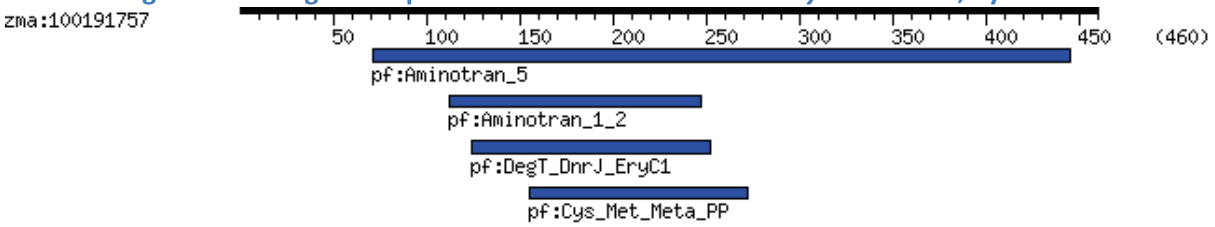
While in *Zea mays* there is a similar pattern -

1. Zea mays EC 2.8.1.7 - Cys Motif

Table 3-11- Zea *mays* EC 2.8.1.7 Cys Motif (CPNIFS) detailing location and E-score of protein domains

Motif id	From	To	Definition	E value
pf:Aminotran_5	71	445	Aminotransferase class-V	9.50E-140
pf:Aminotran_1_2	112	247	Aminotransferase class I and II	0.00065
pf:DegT_DnrJ_EryC1	124	252	DegT/DnrJ/EryC1/StrS aminotransferase family	5.80E-07
pf:Cys_Met_Meta_PP	155	272	Cys/Met metabolism PLP-dependent enzyme	1.60E-06

Figure 3-7 - Diagram of protein domain location on *Z.mays* EC 2.8.1.7, Cys motif

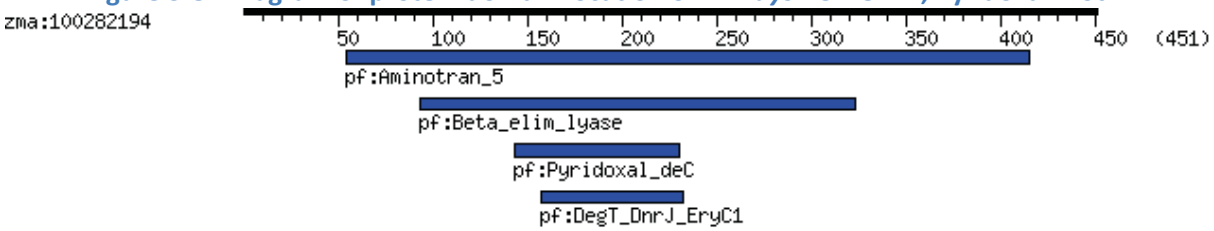


2. Zea mays EC 2.8.1.7 - Pyridoxal motif

Table 3-12- *Zea mays* EC 2.8.1.7 Pyridoxal Motif (NFS1) detailing location and E-score of protein domains

Motif id	From	To	Definition	E value
pf:Aminotran_5	54	415	Aminotransferase class-V	4.10E-88
pf:Beta_elim_lyase	93	323	Beta-eliminating lyase	2.40E-12
pf:Pyridoxal_deC	143	230	Pyridoxal-dependent decarboxylase conserved domain	0.00038
pf:DegT_DnrJ_EryC1	157	232	DegT/DnrJ/EryC1/StrS aminotransferase family	4.10E-06

Figure 3-8 - Diagram of protein domain location on *Z.mays* EC 2.8.1.7, Pyridoxal motif



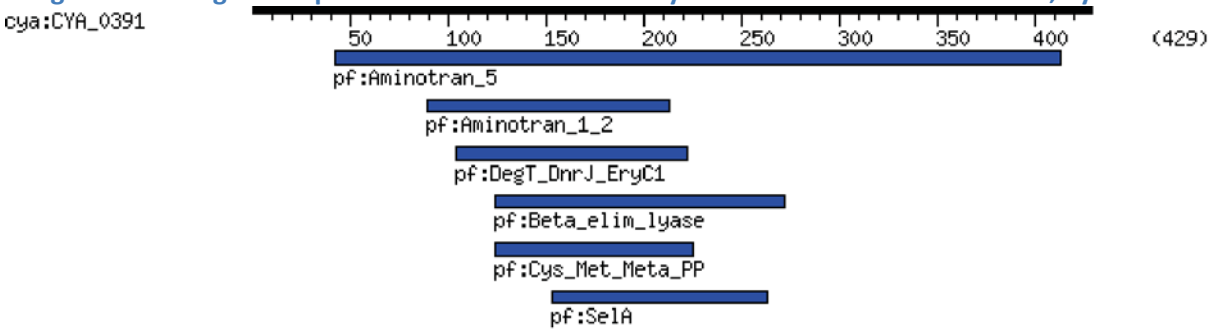
In cyanobacteria, similar motifs to archaeplastida occur but there are another 2 protein families present

1. Synechococcus sp. JA-3-3Ab- Cys Motif

Table 3-13- *Synechoccus JA-3-3Ab* EC 2.8.1.7 Cys Motif (CPNIFS) detailing location and E-score of protein domains

Motif id	From	To	Definition	E value
pf:Aminotran_5	42	413	Aminotransferase class-V	4.20E-156
pf:Aminotran_1_2	89	213	Aminotransferase class I and II	0.00015
pf:DegT_DnrJ_EryC1	104	222	DegT/DnrJ/EryC1/StrS aminotransferase family	4.70E-07
pf:Beta_elim_lyase	124	272	Beta-eliminating lyase	4.50E-06
pf:Cys_Met_Meta_PP	124	225	Cys/Met metabolism PLP-dependent enzyme	0.0092
pf:SelA	153	263	L-seryl-tRNA selenium transferase	0.028

Figure 3-9 - Diagram of protein domain location on Synechoccus JA-3-3Ab EC 2.8.1.7, Cys motif

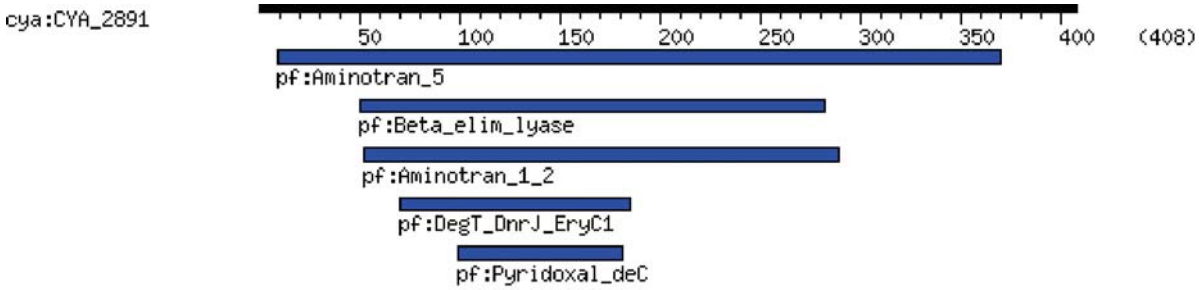


2. Synechoccus sp. JA-3-3Ab- Pyridoxal motif

Table 3-14- *Synechoccus JA-3-3Ab* EC 2.8.1.7 Pyridoxal Motif (NFS1) detailing location and E-score of protein domains

Motif id	From	To	Definition	E value
pf:Aminotran_5	9	370	Aminotransferase class-V	1.30E-83
pf:Beta_elim_lyase	50	282	Beta-eliminating lyase	2.30E-10
pf:Aminotran_1_2	52	289	Aminotransferase class I and II	1.50E-06
pf:DegT_DnrJ_EryC1	70	185	DegT/DnrJ/EryC1/StrS aminotransferase family	0.0085
pf:Pyridoxal_deC	99	181	Pyridoxal-dependent decarboxylase conserved domain	0.12

Figure 3-10 - Diagram of protein domain location on Synechoccus JA-3-3Ab EC 2.8.1.7, Pyridoxal motif

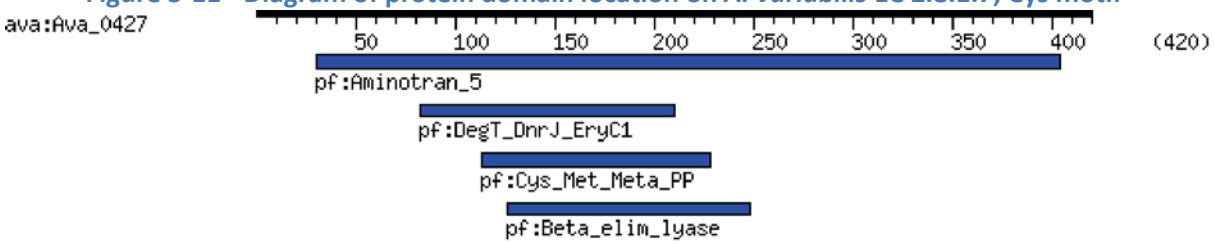


- A.variabilis - Cys Motif

Table 3-15- A. *variabilis* EC 2.8.1.7 Cys Motif (CPNIFS) detailing location and E-score of protein domains

Motif id	From	To	Definition	E value
pf:Aminotran_5	30	404	Aminotransferase class-V	2.30E-150
pf:DegT_DnrJ_EryC1	82	210	DegT/DnrJ/EryC1/StrS aminotransferase family	8.40E-11
pf:Cys_Met_Meta_PP	113	228	Cys/Met metabolism PLP-dependent enzyme	2.20E-05
pf:Beta_elim_lyase	126	248	Beta-eliminating lyase	1.50E-05

Figure 3-11 - Diagram of protein domain location on *A. variabilis* EC 2.8.1.7, Cys motif

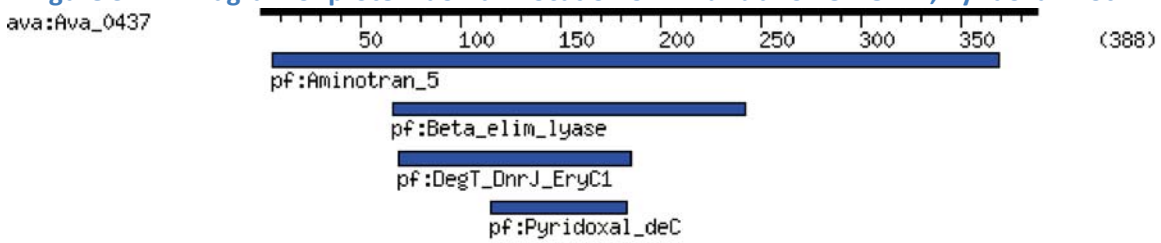


- A.variabilis - Pyridoxal Motif

Table 3-16-*A.variabilis* EC 2.8.1.7 Pyridoxal Motif (NFS1) detailing location and E-score of protein domains

Motif id	From	To	Definition	E value
pf:Aminotran_5	6	369	Aminotransferase class-V	5.40E-92
pf:Beta_elim_lyase	66	242	Beta-eliminating lyase	1.30E-09
pf:DegT_DnrJ_EryC1	69	185	DegT/DnrJ/EryC1/StrS aminotransferase family	8.90E-08
pf:Pyridoxal_deC	115	183	Pyridoxal-dependent decarboxylase conserved domain	0.0086

Figure 3-12 - Diagram of protein domain location on *A.variabilis* EC 2.8.1.7, Pyridoxal motif



As seen above different protein family domains occur in differing places throughout the enzyme for different organisms, and it is these differing placements, lengths and protein families that probably lead to the relatively low scores on the MUSCLE alignments. Interestingly the Aminotran 5 family has a wide range of lengths and also has a range of other protein families that have been identified within it; this identification of other protein families with much lower e-values probably shows more about how enzymes evolve to be capable of a number of functions and also how enzymes may evolve to be part of other metabolic pathways. These problems confound the construction of an ancestral sequence. These results show that evolution is a dynamic process.

Possible reasons on why ASR isn't working as efficiently as hoped.

There are several reasons postulated as to why the ASR did not work as efficiently as we had hoped. Firstly that the time period that this endosymbiosis was theorized to have occurred is just too deep- Markov models having an exponential drop off in reliability after a certain period (Mossel and Steel 2005b). Secondly that the transfer of the cyanobacterial genes from a bacterial coding environment to a eukaryotic nuclear environment effectively meant that AS generated from the eukaryotes had the associated "eukaryotic-noise" and so was incompatible with the AS generated from the cyanobacteria. Thirdly, eukaryote genes, in particular, can be rather variable in their length.

Markov Models

The problem with Markov models was discussed earlier in this chapter. The basic problem is that at deeper divergences the data produced by these models becomes less reliable (Mossel and Steel, 2005a). Adding parameters helps but the situation is one of linear help from the parameters verses exponential decay from the model.

Problems with Multiple Sequence Alignments

In a recent paper, Morrison (Morrison, 2009) makes a few telling comments about the nature of the multiple sequence alignment process. He analysed 1280 papers from the year 2007 that had used this two stage process; he ordered them into subject areas and analysed the processes used. In the subject of "evolution" he discovered that in the first stage of the process 55% of the papers had the MSA manually curated as opposed to automated (by a program, typically CLUSTAL). For the second part of the analysis, nearly all the tree building was automated. The reasons why, he speculates, are to do with CLUSTAL not being able to effectively deal with gaps of variable length. He goes on to argue that substitutions and indels, i.e. sequence variation, are dealt with by computerised algorithms as being a random sample from all sequence variation. This is clearly not correct as DNA/RNAs are not an arbitrary string of characters but are macromolecules with clear biological constraints; he argues that this assumption is the ultimate cause of all problems with phylogenetic sequence alignment.

Summary

The case for ASR effectiveness is built around the use of three tools – BLAST, Pfam and UGene. The results from all three sites are inconclusive with little to differentiate between extant sequences and AS. Probable reasons for this are presented and include-

- multiple isozymes – there are 16 different isozymes in the Archaeplastida organisms chosen for the enzyme EC 4.2.1.35
- Pfam results for enzyme lengths in archaeplastida show differing lengths (54-161 AA) and inconsistent appearance of families across all subjects
- these both have flow on effects when conducting simple similarity scores using UGene

Some thoughts are presented on why ASR is not working as efficiently as hoped. The age since divergence of these organisms may be too deep for the Markov models to produce reliable results and the differences between prokaryotic and eukaryotic transcriptional and translational machineries. The eukaryotic transcriptional machinery may require additional sequence to move the RNA around the cell. ASR has been tested to its limits and the issues that are apparent at this limit have been discussed.

Chapter 4. Further Analysis

Rationale

The ASR in the previous chapters has shown that, unlike the earlier findings of Collins et al, 2003, there was little measureable difference between the extant sequences and the Ancestral Sequences. In this chapter a range of different scenarios are presented.

Firstly, manual curation of the Multiple Sequence Alignments (MSA) is performed. This process is subjective, there being no hard and fast rules about what to delete from an alignment and when to stop with the deletion of difference between the sequences. This is discussed more in the methods section.

A range of the number of sequences to include in the alignment is also investigated. When this project was started there were only 13 Archaeplastida available for analysis through the KEGG database; over the intervening four years this number has increased to 52 Archaeplastida so analysis has been rerun on varying number of sequences- 5, 10, 25 and 50.

As the number of sequences has increased so has the taxonomic distribution of these sequences. This taxonomic distribution is also investigated to the extent that is available – for example, where there were four Monocotyledons available there are now ten available in the KEGG database but conversely there is still only one Bryophyta (moss) and no gymnosperms available for analysis in either the NCBI and KEGG databases .

The quality of the sequencing process was also examined. When the KEGG database was first launched sequencing technology was in its infancy. All the sequences used in this chapter have been checked for their method of assembly to ensure a deep enough coverage ,whether short reads from a Next-Generation sequencing technology have been used, or that Sanger sequencing (whose long-reads ensure accuracy) has been used. See appendix, table 4.

A more thorough investigation of the enzyme variants has been performed. When there are a number of variants available, a MSA is conducted in order to examine what parts of the enzyme are conserved and from this analysis a variant is chosen that contains the most of this conserved region. This is shown in more detail in the methods section.

Methods

Initially, a database of both nucleotide and amino acid sequences was constructed for the first five enzymes in the Lysine pathway and, as a positive control, the first 5 enzymes in the Ether Lipid pathway from the organisms listed in table 4.1 below and the enzymes listed in table 4.2 below.

Table 4-1-Organisms used in this study from the KEGG database.

Group	Family	Code	Organism
Eudicots	Mustard family	ath	Arabidopsis thaliana (thale cress)
		aly	Arabidopsis lyrata (lyrate rockcress)
		crb	Capsella rubella
		eus	Eutrema salsugineum
		brp	Brassica rapa (field mustard)
	Rue family	cit	Citrus sinensis (Valencia orange)
		cic	Citrus clementina (mandarin orange)
	Mallow family	tcc	Theobroma cacao (cacao)
	Myrtle family	egr	Eucalyptus grandis (rose gum)
	Pea family	gmx	Glycine max (soybean)
		pvu	Phaseolus vulgaris (common bean)
		mtr	Medicago truncatula (barrel medic)
		cam	Cicer arietinum (chickpea)
	Rose family	fve	Fragaria vesca (woodland strawberry)
		pper	Prunus persica (peach)
		pmum	Prunus mume (Japanese apricot)
		mdm	Malus domestica (apple)
		pxb	Pyrus x bretschneideri (Chinese white pear)
	Cucumber family	csv	Cucumis sativus (cucumber)
		cmo	Cucumis melo (muskmelon)
	Spurge family	rcu	Ricinus communis (castor bean)
		jcu	Jatropha curcas
	Willow family	pop	Populus trichocarpa (black cottonwood)
	Grape family	vvi	Vitis vinifera (wine grape)
	Nightshade family	sly	Solanum lycopersicum (tomato)
		sot	Solanum tuberosum (potato)
	Amaranth family	bvg	Beta vulgaris (sugar beet)
Monocots	Grass family	osa	Oryza sativa japonica (Japanese rice) (RefSeq)
		dosa	Oryza sativa japonica (Japanese rice) (RAPDB)
		obr	Oryza brachyantha (malo sina)
		bdi	Brachypodium distachyon
		sbi	Sorghum bicolor (sorghum)
		zma	Zea mays (maize)
		sita	Setaria italica (foxtail millet)
	Palm family	pda	Phoenix dactylifera (date palm)

		egu	Elaeis guineensis (African oil palm)
	Banana family	mus	Musa acuminata (wild Malaysian banana)
Basal Magnoliophyta	Amborella family	atr	Amborella trichopoda
Ferns		smo	Selaginella moellendorffii
Mosses		ppp	Physcomitrella patens subsp. patens
Green algae		cre	Chlamydomonas reinhardtii
		vcn	Volvox carteri f. nagariensis
		olu	Ostreococcus lucimarinus
		ota	Ostreococcus tauri
		bpg	Bathycoccus prasinos
		mis	Micromonas sp. RCC299
		mpp	Micromonas pusilla
		csl	Coccomyxa subellipsoidea
		cvr	Chlorella variabilis
Red algae		cme	Cyanidioschyzon merolae
		gsl	Galdieria sulphuraria
		ccp	Chondrus crispus (carragheen)

Table 4-2-the 10 enzymes from the Lysine and Ether Lipid pathways

Pathway	Enzyme (E.C. No)
Lysine	1.2.1.11
	1.1.1.3
	4.3.3.7
	1.17.1.8
	2.6.1.83
Ether Lipid	2.7.8.1
	3.1.1.4
	3.1.4.4
	3.1.4.3
	LPCAT

The enzyme for *Arabidopsis thaliana* was used initially as a search term to find putative homologs for the 51 other organisms. If there were variants in the *A.thaliana* enzyme then all variants were put into a MSA and the variant that was shortest with the most conserved region of sequence was chosen as the variant to search for homologs, as demonstrated below in figure 4.1.

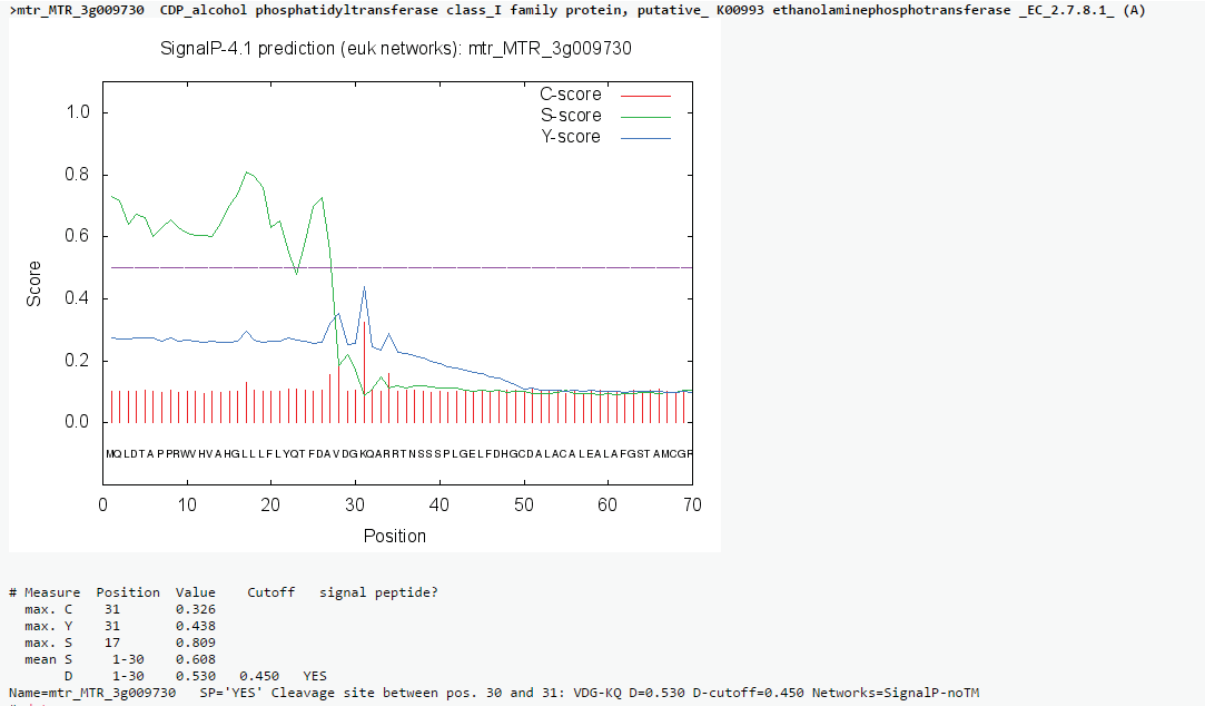
Figure 4-1- MSA of 12 variants for 3.1.4.4 AA. AthAT4G11840, circled in red, was chosen for its region of conserved sequence shown as a solid black line. AthAT4G11830 could also have been chosen.

As KEGG links directly to NCBI, sequences for the other organisms effectively were “blasted”. A homolog was accepted for further analysis if its blasted e-value was greater than 0.001. In some instances no homolog was found for the organisms; typically this was for the algae – see appendix, table 3.

Sequences were then separated into groups as enumerated in the third paragraph of this chapter. Groups were chosen with an even taxonomic distribution in mind however an even sampling was not always possible because of an uneven taxonomic distribution of organisms in the 50 and 25 organism groups.

All 1020 sequences were run through Signal 4.1 (Petersen *et al.*, 2011) to remove any transit peptides. 14 of the 1020 sequences showed transit peptides and these were scattered around the 10 different enzymes. Signal output is displayed below in figure 4.2.

Figure 4-2- Signal output for Medicago truncatula E.C.2.7.8.1 from the ether lipid pathway. The predicted cleavage site is between position 30 and 31.

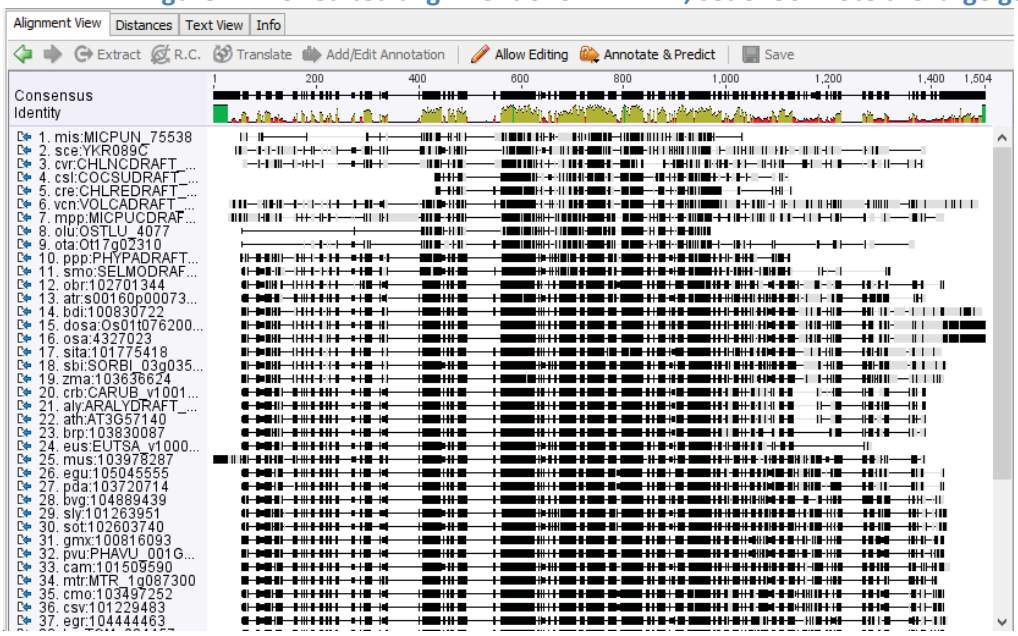


MSA was conducted using Geneious (Kearse *et al.*, 2012) aligning with the Muscle plug-in. At first any obviously long sequences were removed (over 50% longer than the next longest sequence) as its extra length would skew the resulting ancestral sequence, as illustrated below in figure 4.3.

Figure 4-3-Pre-alignment of EC 3.1.1.4 amino acids (AA) (set of 50). The third sequence "bpg" was removed for its excess length.

Then the MSA was run, as illustrated below in figure 4.4.

Figure 4-4- Unedited alignment of 3.1.1.4 AA, set of 50. Note the large gaps.

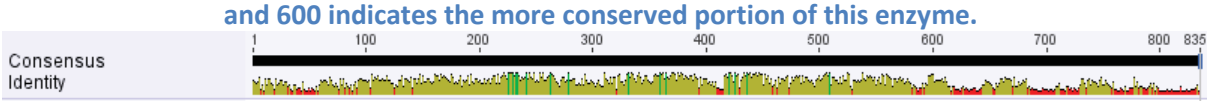


Any gaps were removed where over 90% of the sequences had this gap leaving an edited alignment as demonstrated below in figure 4.5

Figure 4-5- Edited alignment of 3.1.1.4 AA (set of 50).

At the top of the alignment a consensus sequence is generated, as illustrated in figure 4.6 below. These consensus sequences indicate where the more conserved portions of the genes lie.

Figure 4-6- the consensus sequence from 3.1.1.4 amino acid (set of 50) alignment. Green indicates 100% agreement, brown between 100%- 30% and red below 30%. The brown area between 60 and 600 indicates the more conserved portion of this enzyme.



.

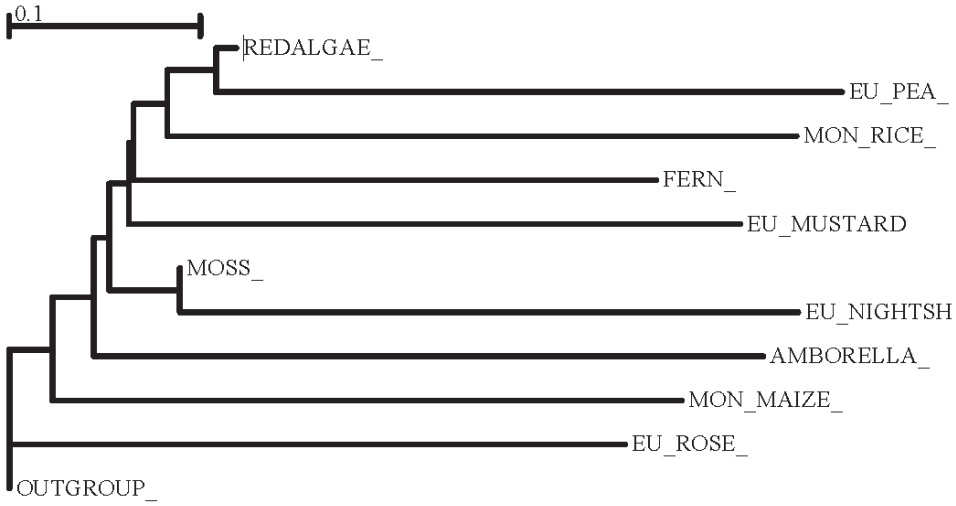
Once the multiple sequence alignment was edited, this was used as input in an ASR program. In this instance, Datamonkey (Pond and Frost, 2005) was used. From the MSA a Neighbour joining (N-J) tree was constructed; an example of this is shown below in figure 4.7.

For the amino acids the JTT protein substitution model was used in all cases with site to site variation modelled using a beta-gamma model and 5 rate classes. Datamonkey gives an Ancestral Sequence (AS) for each node on the NJ tree. In all cases the sequence corresponding to the root node (Node 0) was used as the AS.

Ancestral Sequences were generated for the ten enzymes from two metabolic pathways. Each enzyme class was further split into two major classes, amino acids and nucleotides, and each of these split into groups for the number of sequences in the MSA – 5, 10, 25 and 50.

Figure 4-7- N-J tree for 3.1.1.4 AA (set of 10).

This tree is very different from the standard phylogenetic tree for archaeplastida. A reason for this may be the MSA process- manual curation has removed many of the markers unique to individual organisms skewing the "standard" tree. Lineage sorting may also have an influence.
Key – Some labels are self explanatory- REDALGAE, FERN, MOSS and OUTGROUP- AMBORELLA is a basal magnoliophyta, and the other labels are initially divided into EU, for Eudicot, and MON, for Monocot then followed by the organism's family, EU_ROSE for "mdm" or the common name of the organism , EU_MAIZE for maize.



Results

In order to quantitatively assess whether the AS has shortened the phylogenetic distance between extant sequences from bacteria, the initial *A.thaliana* sequence is blasted against bacteria and the top e-score noted. The AS for the corresponding sequence is also blasted against bacteria and the top e-score noted. The rationale is that the AS should be phylogenetically closer to the outgroup (bacteria) and so have a lower e-value than the sequences from *A.thaliana*. This technique does assume that phylogenetic distance between these enzymes is measurable in an ordered, sequential manner, but is this really the case? The results for nucleotides are expressed below in table 4.3 and for amino acid sequences in table 4.4.

Table 4-3- Comparison of Ancestral Sequences (expressed as nucleotides) and *A. thaliana* blasted against bacteria.

The table details the e-values from the Blast results of *A. thaliana* and the four sets of Ancestral Sequences. NT5 is the AS generated from 5 organisms using nucleotides, NT10 is the AS generated from 10 organisms using nucleotides, etc. The same rationale applies to the following table 4.4. This is a way of quantitatively measuring how close each sequence is to the bacterial kingdom as a proxy for phylogenetically closeness.

Pathway	Enzyme	BLAST	E-value
Lysine	1.2.1.11	A.thaliana	2.00E-12
		NT5	8.00E-17
		NT10	3.00E-23
		NT25	3.00E-61
		NT50	3.00E-86
	1.1.1.3	A.thaliana	5.00E-18
		NT5	3.00E-14
		NT10	9.00E-15
		NT25	2.00E-16
		NT50	4.00E-13
	4.3.3.7	A.thaliana	4.00E-27
		NT5	6.00E-31
		NT10	0
		NT25	2.00E-24
		NT50	7.00E-18
	1.17.1.8	A.thaliana	5.00E-07
		NT5	4.00E-07
		NT10	6.00E-06
		NT25	7.00E-05
		NT50	6.00E-06
	2.6.1.83	A.thaliana	4.00E-16
		NT5	9.00E-37
		NT10	2.00E-37
		NT25	5.00E-33
		NT50	6.00E-26

Pathway	Enzyme	BLAST	E-value
Ether Lipid	2.7.8.1	A.thaliana	0.002
		NT5	1.00E-11
		NT10	6.00E-08
		NT25	4.00E-09
		NT50	1.00E-08
	3.1.1.4	A.thaliana	1.00E-78
		NT5	3.00E-58
		NT10	5.00E-78
		NT25	2.00E-62
		NT50	2.00E-57
	3.1.4.4	A.thaliana	2.00E-35
		NT5	8.00E-26
		NT10	9.00E-24
		NT25	9.00E-16
		NT50	0.069
	3.1.4.3	A.thaliana	5.00E-70
		NT5	9.00E-71
		NT10	2.00E-52
		NT25	7.00E-27
		NT50	5.00E-44
	LPCAT	A.thaliana	0.013
		NT5	0.02
		NT10	0.006
		NT25	0.003
		NT50	0.005

Table 4-4-Comparison of Ancestral Sequences (expressed as amino acids) and *A. thaliana* blasted against bacteria.

Pathway	Enzyme	Sequence	E-value
Lysine	1.2.1.11	A.thaliana	3.00E-144
		AA5	3.00E-140
		AA10	5.00E-138
		AA25	2.00E-132
		AA50	2.00E-137
	1.1.1.3	A.thaliana	0
		AA5	0
		AA10	0
		AA25	0
		AA50	0
	4.3.3.7	A.thaliana	1.00E-116
		AA5	7.00E-116
		AA10	1.00E-116
		AA25	6.00E-118
		AA50	1.00E-114
	1.17.1.8	A.thaliana	2.00E-90
		AA5	1.00E-93
		AA10	2.00E-101
		AA25	4.00E-94
		AA50	2.00E-87
	2.6.1.83	A.thaliana	6.00E-161
		AA5	9.00E-164
		AA10	1.00E-167
		AA25	3.00E-169
		AA50	2.00E-167

Pathway	Enzyme	Sequence	E-value
Ether Lipid	2.7.8.1	A.thaliana	0.001
		AA5	9.00E-10
		AA10	5.00E-10
		AA25	1.00E-08
		AA50	9.00E-08
	3.1.1.4	A.thaliana	1.00E-78
		AA5	7.00E-78
		AA10	6.00E-76
		AA25	1.00E-79
		AA50	2.00E-79
	3.1.4.4	A.thaliana	2.00E-35
		AA5	6.00E-33
		AA10	7.00E-32
		AA25	9.00E-33
		AA50	3.00E-30
	3.1.4.3	A.thaliana	5.00E-70
		AA5	2.00E-76
		AA10	4.00E-66
		AA25	5.00E-66
		AA50	5.00E-72
	LPCAT	A.thaliana	0.009
		AA5	1.40E-02
		AA10	0.022
		AA25	0.028
		AA50	0.043

After running the sequences pre-ASR through SignalP (see figure 4.2) it was suggested that these sequences be run through ChloroP (Emanuelsson *et al.*, 1999) to determine whether these sequences contain chloroplast transit peptides (cTP). Three datasets of 25 AA were selected- two from the Lysine pathway and the other from the Ether Lipid pathway. The results are shown below in figure 4.8.

The cTP trimmed sequences were then run through the ASR protocol with the same parameters as the untrimmed sequences and these results are presented below in table 4.13.

Table 4-5: cTP results from the Ether Lipid enzyme 2.7.8.1

Columns 3, 8 & 13 show the score –the higher the number, the more certain there is an N-terminal cTP. *Columns 4, 9 & 14 (cTP) indicate the absence (-) or presence (Y) of a cTP*-highlighted in the green boxes. Columns 5, 10 & 15 show the cleavage site score. **N.B. - the prediction of the transit peptide length is carried out even if its presence is not predicted**. The rest of the column meanings are self explanatory.

	Lysine 1.2.1.11					Lysine 4.3.3.7					Ether lipid 2.7.8.1				
Name	Length	Score	cTP	CS-Score	cTP Length	Length	Score	cTP	CS-Score	cTP Length	Length	Score	cTP	CS-Score	cTP Length
ath_AT1G14810	375	0.554	Y	2.686	47	365	0.545	Y	5.814	38	337	0.455	-	-1.771	51
atr_s00032p002	379	0.542	Y	2.686	51	366	0.535	Y	7.086	39	343	0.444	-	0.104	76
ava_Ava_3606	347	0.444	-	1.566	10	294	0.446	-	6.201	25	389	0.474	-	-1.548	59
cit_102608194	376	0.570	Y	3.600	32	359	0.473	-	2.481	20	388	0.437	-	-0.175	13
csl_COCSUDRAFT	380	0.530	Y	6.230	38	354	0.567	Y	1.508	34	387	0.478	-	1.1	58
csv_101206329	376	0.577	Y	3.267	32	367	0.499	-	7.481	22	389	0.471	-	-3.927	25
egr_104433627	377	0.541	Y	3.367	28	365	0.451	-	6.522	38	389	0.473	-	1.073	60
eus_EUTSA_v100	375	0.554	Y	2.686	47	364	0.505	Y	5.814	37	389	0.51	Y	-1.587	60
fve_101292991	378	0.568	Y	2.686	50	368	0.529	Y	4.603	41	389	0.48	-	0.562	60
gmx_100814757	379	0.572	Y	2.686	51	363	0.564	Y	4.029	36	394	0.436	-	1.966	80
gsl_Gasu_31010	440	0.529	Y	0.044	15	356	0.515	Y	3.356	46	389	0.468	-	-1.587	60
mdm_103451531	380	0.566	Y	2.686	52	365	0.453	-	6.522	38	353	0.442	-	-4.363	24
mtr_MTR_8g1058	379	0.562	Y	2.686	51	366	0.563	Y	4.369	38	389	0.489	-	1.797	60
mus_103968558	388	0.564	Y	4.331	30	371	0.562	Y	3.389	44	389	0.469	-	-3.533	61
osa_4334188	375	0.522	Y	4.449	31	380	0.493	-	2.11	28	740	0.461	-	2.11	32
ota_Ot03g05540	n/a	n/a	n/a	n/a	n/a	334	0.457	-	9.045	21	389	0.48	-	1.134	60
pop_POPTR_0008	375	0.566	Y	2.686	47	365	0.458	-	3.063	2	383	0.469	-	-1.094	60
ppp_PHYPADRAFT	344	0.431	-	2.686	16	311	0.43	-	1.043	2	389	0.482	-	-1.587	60
pxb_103943087	380	0.566	Y	2.686	52	365	0.453	-	6.522	38	441	0.448	-	-1.587	70
rcu_RCOM_00535	377	0.570	Y	2.686	49	367	0.531	Y	5.333	52	391	0.442	-	-3.611	62
sly_101258686	381	0.574	Y	2.686	53	356	0.491	-	6.916	29	96	0.481	-	1.134	60
smo_SELMODRAFT	346	0.436	-	2.686	16	368	0.52	Y	7.082	41	394	0.444	-	5.42	65
tcc_TCM_011294	374	0.570	Y	2.686	46	365	0.492	-	5.333	50	389	0.478	-	1.134	60
vcn_VOLCADRAFT	377	0.532	Y	10.078	33	344	0.485	-	5.314	9	384	0.44	-	-0.27	78
vvi_100245641	379	0.553	Y	2.686	51	365	0.485	-	5.333	50	389	0.49	-	1.134	60
zma_100191334	377	0.537	Y	4.449	33	377	0.539	Y	5.789	62	389	0.461	-	-3.395	61

Table 4-6: Comparison of cTP-trimmed AS with untrimmed sequences.

AS = Ancestral sequence and AS cTP Tr = Ancestral sequences that have had the chloroplast transit peptides removed. The same notation applies to the *A.thaliana* sequences.

Pathway	Enzyme	Sequence	E-value
Lysine	1.2.1.11	*A.thaliana*	3.00E-144
		AS	2.00E-132
		A.thaliana cTP Tr	1.00E-138
		AS cTP Tr	5.00E-136
	4.3.3.7	*A.thaliana*	1.00E-116
		AS	6.00E-118
		A.thaliana cTP Tr	7.00E-123
		AS cTP Tr	1.00E-120
Ether Lipid	2.7.8.1	*A.thaliana*	1.00E-03
		AS cTP Tr	1.00E-08
		AS	1.00E-08

Discussion

Chloroplast Transit Peptides

From the information presented in tables 4.5 and 4.6, it is apparent that the exclusion of cTP has had a minor effect on the efficacy of the ASR procedure. Table 4.6 demonstrates that for E.C. 1.2.1.11, in both cTP included and excluded examples, *A.thaliana* remains phylogenetically closer to the bacterial kingdom although the distance has lessened by ten orders of magnitude (a 7% decrease). For E.C. 4.3.3.7, the situation has reversed; in the untrimmed instance the AS was phylogenetically closer, once trimmed *A.thaliana* is now phylogenetically closer to bacteria. In both tables E.C. 2.7.8.1 effectively acts as a control as ether lipids are not expected to be exported to the chloroplast, although the instance of the mustard Eutrema salsugineum (eus) having a cTP is an interesting exception. This is possibly an artefact of the cTP recognition process.

This is an interesting result and could be attributable to number of causes. Examination of table 4.5 shows that EC 4.3.3.7 had cTP in 46% of the sequences while EC 1.2.1.11 had cTP in 87.5% of sequences. This may well have skewed the MSA in the case of EC 4.3.3.7 by half the sequences being significantly shortened as opposed to the case of EC 1.2.1.11 where nearly all of the sequences were significantly shortened. Of interest too is that in table 4.4 E.C. 4.3.3.7 had a range of values – larger, smaller and the same – in comparison with the A.thaliana score, depending on how many sequences were in the MSA. This may highlight the issue of taxonomic balance as a potential issue.

An examination of the score columns, in table 4.5 (Columns 3, 8 & 13), show that the cut-off value for cTP identification are values greater than 0.5. In the E.C. 4.3.3.7 section many of the values are very close to this threshold; this may indicate that the cut-off is an arbitrary value and that inclusion or exclusion of a cTP is not as black and white as this cut-off value.

KEGG Efficacy

The accuracy of the KEGG database has been questioned and so has been examined in light of the automatic generation of many of the sequences. Comparisons were made with the UniProt (Consortium, 2015) database - a portion of which, called SwissProt (Bairoch *et al.*, 2004), has been manually curated and the results displayed in table 4.7 below.

Table 4-7: Comparison of KEGG and UniProt databases for E.C. 1.1.1.3 found in the Lysine biosynthetic pathway.

The initial comparison is one of length between KEGG and UniProt (columns 4 & 5) with column 6 indicating where on the sequence the difference is. Column 7 indicates whether the UniProt sequence has been manually curated – i.e. found in SwissProt. The Pairwise percentage identity (column 7) was found by aligning the two sequences in Geneious (Kearse *et al.*, 2012) and column 8 lists the pathways identified by UniProt that these enzymes are found in. (F) indicates that this sequence is a fragment.

1.1.1.3							
		Length		Difference	Curated	Pairwise	Pathway
Taxon	**Organism**	**KEGG**	**UniProt**			**%ge Identity**	
Eudicots	Arabidopsis thaliana (thale cress)	859	916	at 3' end	Y	92.3	Lys/M/T
	Capsella rubella	920	378	conserved 3'	N	15.0	Meth/Thr
	Eutrema salsugineum	930	377	conserved 3'	N	15.8	Meth/Thr
	Brassica rapa (field mustard)	921	378	conserved 3'	N	15.2	Meth/Thr
	Citrus sinensis (Valencia orange)	917	267(f)	conserved 3'	N	11.7	Meth/Thr
	Citrus clementina (mandarin orange)	915	381	conserved 3'	N	16.1	Meth/Thr
	Theobroma cacao (cacao)	952	376	conserved 3'	N	14.9	Meth/Thr
	Eucalyptus grandis (rose gum)	912	373	conserved 3'	N	14.8	Meth/Thr
	Glycine max (soybean)	913	916	23 scattered	N	97.4	no entry
Monocots	Oryza sativa japonica (Japanese rice)	912	384	conserved 3'	N	15.0	Meth/Thr
	Zea mays (maize)	920	920	1 AA	Y	99.9	Lys/M/T
Fern	Selaginella moellendorffii	861	375	conserved 3'	N	16.7	Meth/Thr
Moss	Physcomitrella patens subsp. patens	854	289	conserved 3'	N	70.0	Meth/Thr
Green Algae	Chlamydomonas reinhardtii	917	416	conserved 3'	N	21.6	Meth/Thr
	Volvox carteri f. nagariensis	923	n/a	n/a	n/a	n/a	n/a
	Ostreococcus lucimarinus	807	286	conserved 3'	N	17.5	Meth/Thr
	Ostreococcus tauri	811	353	conserved 3'	N	20.0	Meth/Thr
Red Algae	Galdieria sulphuraria	1028	1028	none	N	100.0	None shown
	Chondrus crispus (carragheen)	802	n/a	n/a	n/a	n/a	n/a

Of interest are the disparate lengths of proteins in the two databases except where the sequences have been manually curated and those of *Glycine max* (soybean) and *Galdieria sulphuraria*. These four sequences also correspond to Lysine or "no entry/none shown" entries in the final column. These results are unsurprising given the generalist nature of E.C. 1.1.1.3 and would indicate that not much work has been done on the Lysine pathway by UniProt and, further, that for this enzyme and range of organisms very little manual curation has been completed. Analysis of another protein E.C. 4.3.3.7 from the same pathway is presented in table 4.6 below.

In this instance the UniProt database does not contain many of the relevant sequences. For those sequences that are there the Pairwise percentage identity is high (column 8). These results establish that the KEGG database is complete for the organisms and pathway chosen. There has also been suggestion that the KEGG sequences are truncated (part of a larger sequence); the results from table 4.5 indicate that it may be the UniProt sequences that are truncated as demonstrated by the short

enzyme lengths – columns 3 and 4. Further study indicated that none of the KEGG-sourced Lysine sequences showed any sign of truncation when entered as a search term in Blast.

Table 4-8: Comparison of KEGG and UniProt databases for E.C. 4.3.3.7 found in the Lysine biosynthetic pathway

4.3.3.7							
		Length		Difference	Curated	Pairwise	Pathway
		KEGG	UniProt			%ge Identity	
Eudicots	Arabidopsis thaliana (thale cress)	365	365	same	Y	100.00	Lysine
	Capsella rubella	365	no hit	n/a	n/a	n/a	n/a
	Eutrema salsugineum	364	no hit	n/a	n/a	n/a	n/a
	Brassica rapa (field mustard)	364	no hit	n/a	n/a	n/a	n/a
	Citrus sinensis (Valencia orange)	364	no hit	n/a	n/a	n/a	n/a
	Citrus clementina (mandarin orange)	346	346	same	N	100.00	Lysine
	Eucalyptus grandis (rose gum)	365	337	5' end short	N	90.50	Lysine
	Glycine max (soybean)	363	332	5' end short	Y	93.00	Lysine
Monocots	Oryza sativa japonica (Japanese rice)	380	no hit	n/a	n/a	n/a	n/a
	Zea mays (maize)	377	380	5' end short	Y	81.80	Lysine
Fern	Selaginella moellendorffii	368	no hit	n/a	n/a	n/a	n/a
Moss	Physcomitrella patens subsp. patens	311	311	same	N	100.00	Lysine
Green Algae	Chlamydomonas reinhardtii	345	no hit	n/a	n/a	n/a	n/a
	Volvox carteri f. nagariensis	344	no hit	n/a	n/a	n/a	n/a
	Ostreococcus lucimarinus	369	no hit	n/a	n/a	n/a	n/a
	Ostreococcus tauri	334	no hit	n/a	n/a	n/a	n/a
Red Algae	Galdieria sulphuraria	356	356	same	N	100.00	no hit
	Chondrus crispus (carragheen)	161	no hit	n/a	n/a	n/a	n/a

Truncated Sequences

The possibility of truncated sequences being used was also investigated. Amino acid *A.thaliana* sequences from all 10 enzyme sets were used a s a search term in Blastp. If the sequences were truncated this would be apparent in the graphic portion of the results. There was no evidence of this. A typical result is shown in figure 4.8 below.

AS Nucleotide and Protein Comparison

A comparison of nucleotide and protein ancestral sequences from the Lysine pathway has been requested and is presented in table 4.9 below. The ancestral sequences expressed as nucleotides were translated in Geneious and a Muscle pairwise alignment then conducted with the protein AS and the percentage of identical sites noted. In many instances the lengths of the sequences were different due to the manual curation of the respective MSA. A further factor confounding direct comparison is the degenerate nature of the genetic code.

Figure 4-8: Graphic Blastp result of E.C. 1.1.1.3

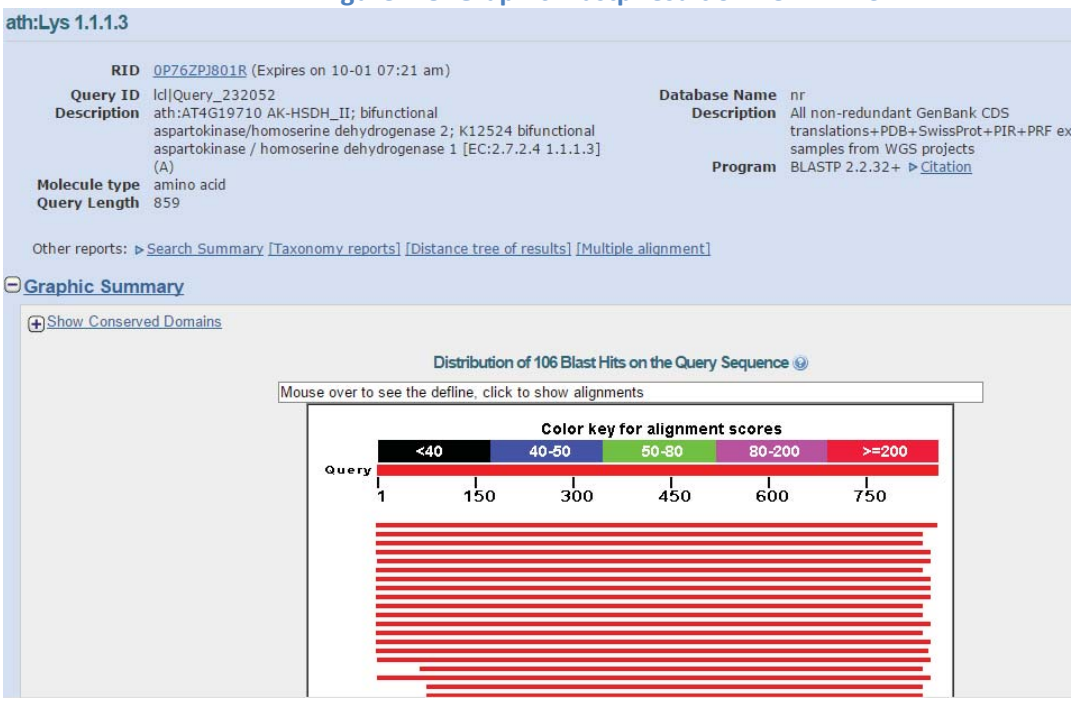


Table 4-9: Comparison of nucleotide and protein ancestral sequences from the Lysine pathway. The top line reads- that in the comparison of nucleotide and protein ancestral sequences for E.C. 1.2.1.11 the AS that were generated from 5 organisms; the sequences contained 55.9% identical sites.

Pathway	Enzyme	Sequence	Identical Sites
Lysine	1.2.1.11	5	55.9%
		10	36.8%
		25	73.6%
		50	15.4%
	1.1.1.3	5	14.5%
		10	59.6%
		25	14.6%
		50	22.4%
	4.3.3.7	5	21.2%
		10	19.5%
		25	54.3%
		50	43.9%
	1.17.1.8	5	12.8%
		10	39.7%
		25	13.7%
		50	22.7%
	2.6.1.83	5	37.8%
		10	17.3%
		25	15.4%
		50	49.1%

Results

From the results section, tables 4.3 and 4.4, pages 81 and 82; it is apparent that for Lysine, there are two instances when the e-value for *A. thaliana* is greater than all the ancestral sequences (that is, the AS are more similar to prokaryotes than *A. thaliana*), and that is in the cases of E.C. 1.2.1.11 NT and 2.6.1.83 AA and NT. For the positive control data set of Ether Lipids the enzyme 2.7.8.1 stands out with AS having lower values of up to 7 orders of magnitude; the most "successful" case. Therefore, strictly speaking the ASR has worked as expected in five out of the twenty instances; with most of the other enzymes the difference between the AS and *A. thaliana* is very little – the *A.thaliana* sequence almost working out as an average of all the e-values.

E.C. 1.1.1.3 AA (**a**mino **a**cids) has e-values of zero; this indicates that the enzyme is nearly identical to those found in the prokaryote kingdom. That a similar result is not apparent within E.C. 1.1.1.3 NT (nucleotides) is caused by the degenerate nature of the genetic code- different codons produce the same amino acid. The nucleotides that code for the prokaryote enzyme that gave the zero scores for the 1.1.1.3AA enzyme are different from those that 1.1.1.3 NT consists of – that is, different nucleotides gave the same codons.

There appear to be no clear patterns as to the relationship between extant and ancestral sequences apart from a general similar e-value in most of the instances; it appears that evolution does not conform to an ordered, sequential pattern. So what may be inferred from these set of results.

As previously discussed the possible reasons for this type of result are plentiful and have been enumerated in chapters two and three. This new analysis in this chapter allows examination of differing numbers of organisms in each enzyme dataset, a thorough manual curation of the multiple sequence alignments, a positive control dataset, an examination of enzyme variants at the initial stage of the process and the excision of any transit peptides at the N-terminal end of the proteins.

Has the number of organisms used in the datasets made any difference to the efficacy of the ASR program? I think that the answer to this question lies in the MSA. At this stage of the process it is apparent that there are conserved portions of the genes that all tend to align and give higher values in the consensus sequence, see figure 4.5 above, page 78. I think it is in these conserved portions that the ASR process has value. The remaining portions of the sequences tend to be less conserved and so more highly variable as can be seen in figures 4.3 and 4.5, pages 78 and 79. ASR tends to average out these areas of highly variable sequence and so does not add value to the concept of an AS.

Therefore, the number of sequences adds value if

- There is an even distribution of the taxa used in a study
- The conserved portion of the gene is in the same area in all the aligned genes.

In this instance there are a maximum of 52 sequences used and unfortunately most of these are from higher order plants (i.e. monocotyledons and dicotyledons) and Chlorophycean green algae. There are few or no gymnosperms, ferns, lycophytes, hornworts, liverworts and mosses, see table 4.1 above, page 74. This lack of coverage from the middle orders of plant life could potentially skew ASR if the conserved areas of these genes were significantly different, or in a different location within the gene, from the algae and higher order plants. This leaves the datasets of five and ten organisms as providing the most taxonomically balanced datasets in this study although there is no indication that these datasets consistently give lower e-values than *A. thaliana.*

The positive control dataset of Ether Lipids confirms that the patterns discussed here are universal and not an anomaly of the Lysine pathway.

The choice of enzyme variants at the initial stage of the ASR process, when putative homologs are chosen from the range of archaeplastida available, is probably not as important as taxonomic balance. This is an issue that should be resolved in the future. It is the conserved portion of the gene that gives the most value to the concept of the AS. Variants tend to have the same area of conserved gene in them, see figure 4.1 page 76.

Summary

While there have been no major changes to the results from this new dataset, this work has addressed some of the shortcomings of design and implementation of chapters two and three. This work has highlighted the importance of the manual curation of MSA and perhaps has shown the inappropriateness of ASR over very long phylogenetic distances. This supports the mathematical results of Mossel and Steel (2004; 2005a).

Chapter 5. Ancestral Sequence Reconstruction

This chapter consists of a relatively short introduction, together with a manuscript that has been submitted for publication.

Cox, S.J.L., White W.T.J, Collins, L.J. and Penny, D. (2003). Ancestral Sequence Reconstruction – how far back can we go? *Genome Biology and Evolution*. (Submitted).

It is well accepted that the chloroplast/plastids were originally free-living cyanobacteria that were taken up by a eukaryote host cell by an endosymbiotic process. As discussed earlier, only a few genes were retained in the chloroplast (see Table 5.1 for the genes analysed), a larger number were transferred to the nucleus, but probably a larger number again are thought to have eventually been lost – because they were redundant? The central issue studied here was whether we could learn anything about the biosynthetic abilities of the eukaryote that initially took up the cyanobacterial chloroplast. There are a spectrum of possibilities, ranging from an animal-like eukaryote (that obtained some amino acids and vitamins (cofactors) from its diet), to being fully autotrophic (and not requiring any factors from its diet). So it is possible that we could learn something about the metabolism of the ancestral eukaryote that took up the cyanobacterium.

The approach we use depends on Ancestral Sequence Reconstruction (ASR), and this may be used to uncover potential gene homologs as is the case in the paper here. Collins and colleagues (Collins et al. 2003) used ASR, for example, to uncover RNase P protein homologues in organisms that retrieved no results using extant RNase P proteins in either BLAST or HMMER databases. Four of these proteins (Pop4, Pop1, Pop5 and Rpp21) had been found in humans and some other eukaryotes - mouse, *Drosophila melanogaster*, *Caenorhabditis elegans* and the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*; the authors were interested in finding homologs in three other eukaryotes – two of which were from different supergroups than the known homologs, namely in (*Encephalitozoon cuniculi* – a fungus; *Entamoeba histolytica* – an Amoebozoan; and *Giardia lamblia* – an Excavate) – see figure 1.2, page 15.

Inferring an ancestral sequence, or most any phylogenetic analysis, is a three part process; the primary homology assessment is the multiple sequence alignment (MSA), the second is the tree building process (Morrison, 2009) and the third is the construction of the ancestral sequence. When inferring an ancestral sequence an alignment (of multiple sequences) is first constructed from relevant extant sequences in order to identify similarities among the sequences. This alignment of sequences is not a direct sequence comparison rather they are averaged changes of compositions within the sequence alignment (Kurland and Harish, 2013, unpublished).

There are many differing models. Use is dependent on which

- genetic unit is to be analysed - nucleotide, codon or amino acid
- genetic code that is pertinent - universal, mitochondrion, or plastid
- model best describes the weight of transition and transversion; some models will keep the same proportions of the nucleotides or amino acids in the inputted data as the outputted data (base heterogeneity), for example the TN93 model
- level of evolutionary pressure is acting on which part of the protein (among site heterogeneity); due to differing roles in the structure and function of different parts of that protein, some areas maybe substitutional hotspots while other areas may be highly conserved, for example the JTT-Γ model. These models ending in a capital gamma (Γ) allow for differing mutation rates at differing parts of the protein by randomly drawing a mutation rate function from a statistical distribution (Yang, 2006)

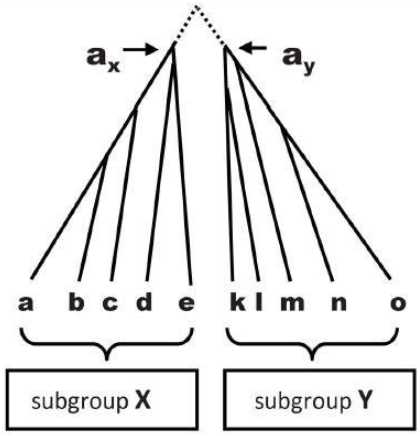
The data used is illustrated in Table 5.1, and shows that it is many of the genes central to photosynthesis are the ones analysed. An important general point for the method is that the two subgroups are defined first by prior information as independent subgroups (see Figure 5.1). However, we always test this later on the data as it is used, and for example, this also checks that they really are two separate subgroups – for example that red algae and green algae are discrete on the chloroplast genes.

I did the primary data gathering, alignment and writing of the paper, Tim White ran the data through his computer program, LJC and DP gave general oversight and planning, and obtained the funding. All authors approved to the final manuscript.

Table 5-1-List of all core chloroplast genes present in essentially all chloroplast genomes from photosynthetic organisms. Adapted from Barbrook et al, (2010).

Photosystem I	Photosystem II	Cytochrome b_6f complex	ATP synthase	Ribosomal protein large subunit	Ribosomal protein small subunit	RNA polymerase	Unknown protein	Rubisco	rRNA
psaA	psbA	petA	atpA	rpl2	rps2	rpoA	ycf4	rbcL	23S
psaB	psbB	petB	atpB	rpl14	rps3	rpoB			16S
psaC	psbC	petD	atpE	rpl16	rps4	rpoC1			
psaJ	psbD	petG	atpF	rpl20	rps7	rpoC2			
	psbE		atpH	rpl36	rps8				
	psbF				rps11				
	psbH				rps12				
	psbI				rps14				
	psbJ				rps18				
	psbK				rps19				
	psbL								
	psbN								
	psbT								
	psbZ								

Figure 5-1- The trees for two subgroups X and Y are illustrated along with the ancestral sequences ax and ay, from (White et al., 2013). Although the subgroups X and Y are based initially on 'prior knowledge', they are tested objectively on the data as used.



Ancestral Sequence Reconstruction – how far back can we go?

Simon J Cox[1,*], W Tim J White[1,3] Lesley J Collins[2] and David Penny[1]

[1]Institute of Fundamental Sciences, Massey University, Palmerston North, P.O. Box 11,222, New Zealand.

[2] Faculty of Health Sciences, Universal College of Learning, Palmerston North, New Zealand.

[3] Current address, Institut für Informatik, Friedrich Schiller Universität, Jena, Germany

*Author for correspondence: Simon Cox, Institute of Fundamental Sciences, Massey University, telephone +64 6356 9099, ext 84728; fax, +64 6 3505682; e-mail S.J.Cox@massey.ac.nz

Abstract:

Ancestral Sequence Reconstruction (ASR) was used to quantify how far back we could detect homologs that originated from cyanobacteria, and that were later taken up in the eukaryotes. Earlier work had shown that the technique worked effectively for at least as far back as the two groups of green algae (Chlorophytes and Charophytes – including the Land Plants). We report here (for genes still coded for in chloroplasts) that the approach also works effectively for red algae versus green algae, and for all algae versus cyanobacteria. However, the technique appears to work less well for genes that probably originate it the cyanobacteria but which were then transferred to the nucleus. This may be because the genes transferred to the nucleus become more eukaryotic in nature, with an exon/intron structure, and longer linker regions on the loops that go between the central portions of the 3-D structure of the proteins. This could be contributing to uncertainties in alignment. The location of cyanobacterial homologies in eukaryotic vitamin and amino acid metabolic pathways is also considered [172 words].

[3300 words, 4000 allowed]

Keywords: Ancestral Sequence Reconstruction, metabolic pathways, cyanobacteria.

Introduction

Our primary interest here was to learn more about the metabolism of the eukaryote that took up the cyanobacterium which became the plastid/chloroplast. Could we infer whether that eukaryote was more 'animal-like' (and obtained many of its essential amino acids and its cofactors [vitamins] from its diet), or was it somewhere at the other end of the spectrum and more autotrophic (and able to synthesize all these requirements by itself).

However, a major question has been raised by Mossel and Steel (e.g. 2005a) who point out that the Markov models used with sequence data lose information at deeper divergences; and that the falloff is exponential at deeper times. The situation is improved linearly by increasing the number of sequences, and so more sequences will always help. However, an exponential decay at deeper times, versus a linear increase with the number of sequences, there is no doubt that the exponential decay with time will eventually become the dominant factor. This could be a reason why some deeper aspects of eukaryote phylogeny still appear unresolved.

In order to investigate these questions outlined above, a two-pronged approach is used that revolves around the analysis of ancient metabolic pathways synthesising essential amino acids and cofactors (vitamins) in both Archaeplastida (plants and algae) and Cyanobacteria. Initially, cyanobacterial homologs were identified from these pathways and the location that they hold within these pathways – early, middle or final – was analysed. Metabolic pathways are thought to develop by recruitment (Caetano-Anollés *et al.*, 2009) so the location of the cyanobacterial homologs within these pathways may help allude to the state (animal-like to autotrophic) of the cell that took up the plastid ancestor. If results show that large portions of these pathways have cyanobacterial homology then this might indicate that the ancestor was animal-like in its diet because parts of these pathways would be redundant due to these vitamins and amino acids being obtained from its diet. Upon evolving into an essentially autotrophic organism (plants and algae) these pathways may have recruited enzymes from the endosymbiont in order to obtain these otherwise redundant pathways and it is this signature of cyanobacterial homology that is investigated. However, being able to infer accurately the relationships of the enzymes does rely on the expectation of a stable relationship between the protein enzyme and its sequence; the second problem alluded to above.

The second part of the approach is to use ASR in the hope of discovering deeper cyanobacterial homologs that may have evolved beyond the recognition of current homology

software. This approach was used successfully within the eukaryote kingdom by Collins *et al.* (2003) which is discussed below.

Ancestral Sequence Reconstruction (ASR)

In general, ASR is an established technique whereby an ancestral state of a given group of extant proteins is inferred through the alignment of sequences and construction of a phylogenetic tree, and has been used in a variety of applications. For example, Sun et al. (2002) have used this technique to infer the ancestral states of three *Pax* genes; these genes have an essential role in the embryonic development of organs and tissues (ranging from the eyes and central nervous system to the pancreas and B-lymphocytes) and have been isolated in a range of Animalia – *Drosophila*, cnidarians, sea urchins and other animals. These Pax genes were grouped functionally and AS constructed for the two main supergroups in order to determine when functional differences in the genes appeared. Thornton et al. (2003), and subsequent papers (Thornton, 2004, Harms and Thornton, 2010, Hart *et al.*, 2014), also used ASR to infer, and then to synthesize actual ancestral steroid receptor genes, that reconcile the activities of hormones that direct sexual differentiation, reproduction, behaviour, immunity, and stress response, from which all extant steroid receptors evolved, they called it Ancestral Sequence Resurrection, also 'ASR'. Results indicated that steroid reception evolved before the development of bilaterally symmetric animals and was subsequently lost in arthropods and nematodes.

ASR may also be used to uncover potential gene homologs as is the case in this paper. Collins and colleagues (Collins et al. 2003) used ASR to uncover RNase P protein homologues in organisms that retrieved no results using extant RNase P proteins in BLAST and HMMER databases. Four of these proteins (Pop4, Pop1, Pop5 and Rpp21) had been found earlier in humans and some other eukaryotes- mouse, *Drosophila melanogaster*, *Caenorhabditis elegans* and the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, and using ASR allowed homologs to be readily detected in remote eukaryotes, such as *Trichomonas* and *Giardia* (Collins et al. 2003). Sometimes this technique is used to discover the function of now dysfunctional genes. Early on Adey et al. (1994) resurrected a mouse "F" LINE-1 retroposon from a sub-family thought to act as a template that encodes reverse transcriptase; there are two types of this retroposon - A & F –"A" being transcriptionally active and "F" considered functionally extinct. Through ASR the "F" was

resurrected to a stage where there was promoter activity. So ASR has a long history, and has been used for many applications in evolutionary studies.

Alignments of these homologs were constructed and phylogenetic trees generated in order to ascertain the most likely ancestral amino acids. The construction of the AS allowed more distant homologs to be discovered by entering the AS as a search term (Collins et al. 2003). Recently, Wolf and Koonin (2013) analysed Ancestral Reconstructions in many kingdoms of life with the aim of tracking gene gain and loss among the many lineages.

For the second half of the ASR process, inferring the AS from the tree, Maximum Likelihood (ML) allows an explicit model of character evolution and so has chosen the tree with the best fit to the model. Models can estimate probabilities of differing kinds of mutations within the sequences and score an optimal tree from these probabilities. ML methods of ASR produce more robust sequences due to the specific models of character evolution than the MP methods, though more data is required (Wolf and Koonin, 2013). These models, used in ML constructions, work by weighting the probabilities for changing (evolving) nucleotides or amino acids. For example - a higher probability would be assigned to adenine evolving to guanine (both being purines – a transition) than to adenine evolving into a thymine (a pyrimidine – a transversion). Similar weights may be modelled to amino acid changes - amino acids with similar physico-chemical properties tend to interchange with each other at higher rates than dissimilar amino acids; for example aspartic and glutamic acids, and also isoleucine and valine acids each have similar properties (Yang, 2006). Differing scores are also given to insertions and deletions.

In the PAML 4 package, which is used in all reconstructions here, an Empirical Bayes program is used alongside the ML methods to account for differences in branch lengths of the phylogenetic trees generated. This approach replaces parameters in the model (branch lengths and substitution rates for example) with their ML estimates. Care must be taken in the fact that the sequences produced are not real observed data; this being stated, the sequences produced do have theoretical application (Yang, 2007), and has been shown to be effective (White et al. 2013), at least for early eukaryotic divergences.

Ancestral Sequences were generated from 13 plant and algal species however these failed to uncover additional cyanobacterial homologs. This analysis led to the question of the depth at

which ASR would operate successfully; the second half of this approach addresses this question.

Results and Discussion

Position of Cyanobacterial Homologs

161 enzymes from 18 metabolic pathways were examined - 25 of the enzymes showed sufficient cyanobacterial homology (see methods) and their position within the pathway was then analysed. 56% of the homologs are found in initial or early positions in the pathways (see Table 1); this percentage rises to 72% when homologs found in the late or final positions of these metabolic pathways are added. These positions in the pathways could suggest that pre LECA was more autotrophic (prey) than heterotrophic (predator) as the pathways studied are not full of cyanobacterial homologs but it does raise questions as to the process of pathway degradation when metabolites are not needed to be constructed *de novo*; presumably, degradation starts at the ends of the pathway working back to the core but there has been no work done on this as far as we know.

The location of these cyanobacterial homologs in the pathways studied could be replacements for a partially degraded pathway caused by redundancy due to a change in diet (indicating a predator lifestyle for LECA) or they may have been recruited over enzymes displaying a non-cyanobacterial homology due to superior fitness. A third possibility is that they may be there by chance as evidence shows the individual enzymes are highly variable while their function is not (Forst and Schulten, 2001).

ASR Depth

Three examples, drawn from the pool of 41 genes analysed, of the results for different genes from the green algae versus red algae are shown in Figure 1. (A full list of genes and the results from all 3 scenarios, outlined in Methods, are shown in the Supplementary Tables). The small circle represents the alignment score between the reconstructed ancestral sequences of each group. If the score (y-axis) is high then this indicates that the ancestral red alga is more similar to the ancestral green alga than most extant red algae are to most extant green algae. **(**This is what is expected if evolution is working.) Thus the ASR program is working well as indicated in the plots for genes atpE and psaB for the green and red algae. The x-axis shows the alignment score which increases with the length of the proteins. This tends to be associated with the short lengths of the proteins for this gene, and is similar to the

effect reported by White et al. (2013). But also, the plot for psbF may indicate that the AS alignment score is not optimal.

ASR appears to be working well for all three scenarios as illustrated by the last three rows in Table 2, particularly for the genes within chloroplasts (the first 4 rows of results are from White et al. (2013). It is apparent that there are high probabilities that the deeper groups being compared on each row all ancestrally converge (although there is an effect of sequence length of the protein, shorter proteins show more variability, as reported in White et al. 2013). The degree to which this is happening may be discerned from the $p(X^2)$ column statistic. The top six rows all have statistics below 1×10^{-19} which are extremely low probabilities of the convergences happening by chance, compared with the $p(X^2$, control) statistic which compares control data from different genes where there is no convergence (White et al. 2013). These first six rows are all genes within the chloroplast, and so they retain their 'prokaryote' gene structure (even though now in the chloroplast). So it appears that the ASR test is working for around 1.5 billion years for genes that are retained in the chloroplast.

In contrast, the final row shows genes that are inferred to come from the original cyanobacterium (as evaluated by BLAST searches that found the closest relative to be cyanobacterial), but that have now been transferred to the nucleus. This row still compares eukaryote and bacterial genes which have been selected beforehand for their homologous relationship. The first important point here is that the probability of observing the convergence for chloroplast genes is about 3.9×10^{-23} for the 28 genes retained in the chloroplast (there are 2 d.f. per gene) but only about one chance in 2.3×10^{-4} for the 17 genes transferred to the nucleus. This is definitely still significant convergence, but much less than the significance of the convergence shown by the genes retained in the chloroplast, and the difference in the results between genes retained in the plastids, and those transferred to the nucleus, does need explanation. The relationship may be more tenuous because the genes are operating in environments that diverged around 1.5 billion years ago. However, eukaryotes and bacteria have evolved differing strategies and machinery for transcription, translation and transport through the cytoplasm. Perhaps the first point though is that genes retained in the plastid do retain their prokaryote structure and secondly it is this change in environment that may cause these cyanobacterial-originating nuclear genes to increase in length compared to their cyanobacterial homologs.

Martin (2003) points out that chloroplast genes that were transferred to the nucleus had to acquire the correct expression and targeting that allowed the proteins to be constructed on ribosomes in the cytosol and then reimported into the chloroplast with the help of transit peptides. A similar line of argument is proposed in Dorrell and Howe (2012) who state that chloroplast derived genes that have been transferred to the nucleus require the acquisition of elements such as nuclear promoters (needed to initiate eukaryotic transcription) and transport elements needed to move the gene products through the cell.

One important difference between prokaryotic and eukaryotic transcriptional and translational machinery is that prokaryotic RNA polymerase is directly coupled with mRNA-processing enzymes and is effectively mature after transcription, while the eukaryotic mRNA requires extensive processing a 5'cap, UTR's, polyadenylation (Nevins, 1983). The physical differences between prokaryotic and eukaryotic transcriptional and translational machinery is one of complexity, eukaryotes being the more complex (Kapp and Lorsch, 2004, Hahn, 2004) - for example, there are three translation initiation factors in bacteria, 12 in eukaryotes, made up of 23 different proteins (Kapp and Lorsch, 2004). In transcription, bacteria and archaea have only one Pol (RNA Polymerase) where eukaryotes have three (Pol I-III) to synthesise different classes of RNA; associated with these polymerases are the σ-factors- one in bacteria, up to almost 60 associated with Pol II in eukaryotes (Hahn, 2004).

There is also the issue of the location at which these events take place - in bacteria, which have no nuclear compartment, transcription and translation are linked together so that as the Shine-Dalgarno sequence emerges from the polymerase, the small ribosomal subunit is there ready to begin translation (Kapp & Lorsch, 2004). In eukaryotes transcription occurs in the nucleus and translation in the cytoplasm and so the mRNA must be transported (Nevins, 1983).

It is well established that eukaryote protein length is almost double that of their bacterial homologues (Brocchieri and Karlin, 2005, Wang *et al.*, 2011, Zhang, 2000). In Wang et al (2011), they examine protein structures in 745 genomes, the ratio of eukaryote Super Family domains to the non-domain regions (internal linkers and terminal tails) of the protein is approximately 1:1, while for archaeal and bacterial proteins the ratio is 3.5:1, an apparent three and a half fold increase in eukaryote linker length. However, the Super Family domain lengths for orthologs from all three kingdoms remain very similar.

Another potential reason for the lower score for the comparison of algal nuclear-encoded genes for vitamin and amino acid synthesis and their cyanobacterial homologs is the degenerative nature of Hidden Markov Models (HMM). Markov models (Mossel and Steel, 2005a) are now standard for modelling the evolution of aligned genetic sequence data . Markov models are the algorithms used to construct the substitution/ insertion/deletion rates in phylogenetic programs such as in PFAM, ASR and CLUSTAL. The main criticism of Markov models is that at deeper divergence times there is an exponential drop-off in the reliability/accuracy of the sequences produced. This is the point at which the probability of a substitution (generated by the Markov model to represent the evolution of the nucleotide or codon) throughout the length of the branch of the phylogenetic tree passes a certain critical value (Mossel and Steel, 2005b).

To counterbalance this there is a linear increase in reliability with an increase in the number of sequences used; there is a suggestion that the reliability may also be increased by using a combination of models (Mossel and Steel, 2005a, Sober and Steel, 2002) as illustrated in Figure 2. However, this paradigm will only make a marginal difference as the dynamic is one of linear versus exponential decay. This dynamic does not appear to have been explicitly stated in the field of evolutionary genetics.

Materials and Methods

Homolog Identification

161 enzymes from 18 different metabolic pathways (Table 3) were identified and analysed using the Kyoto Encyclopaedia of Genes and Genomes (Kanehisa *et al.*, 2008). KEGG is a database that links genomic information with higher order functional information. It lists metabolic pathways that link to sequences from chosen organisms, and has a further option to generate each sequences homologies. It also has a function to run a CLUSTALW (Larkin *et al.*, 2007a) analysis on these homologies that may then be displayed as a Neighbour-Joining (NJ) tree. This was the method used to find cyanobacterial homologies.

Arabidopsis thaliana is the best annotated plant genome and so homologies were found with the KEGG homology function, using *Arabidopsis* enzymes as the basis for the homology search. Closest homologous species, outside the archaeplastida, were noted and patterns collated if the enzyme of interest had four or more cyanobacterial homologies out of a

possible score of twelve. In all cases, the default settings were used for CLUSTALW and Neighbour-Joining (NJ) trees.

ASR Efficacy

We tested three scenarios through the use of ancestral sequence reconstruction (ASR). Firstly, to compare the depth at which ASR may be reliably used, Ancestral Sequences (AS) generated from green algal plastids have been compared to chloroplast encoded genes in red algae, and secondly for those retained in both red and green algae versus those in cyanobacteria. This second approach will estimate sequences that occurred more recently than the original eukaryotic endosymbiotic event but more distant than the estimated 1000 MYA events successfully used in White et al. (2013) using ASR. This will estimate events that took place around 1 - 1.5 BYA (Leliaert *et al.*, 2011).

The third scenario was to reconstruct the AS generated from algal nuclear-encoded genes for vitamin and amino acid synthesis and compare these with AS generated from their cyanobacterial homologs. This will test the closeness of cyanobacterial and algal nuclear homologs that have been identified as being transferred to the eukaryotic archaeplastidal nuclei after the chloroplastic endosymbiotic event transpired. This will also further test the depth at which ASR may be used across the prokaryote/eukaryote boundary.The genes for this part of the experiment where chosen by homolog assessment, that is, all genes from these pathways were run through the KEGG (Kanehisa et al. 2008) ortholog function and those genes that were identified as having cyanobacterial homology were chosen.

A seven step method was used for quantifying convergence between two subgroups - this is outlined in detail in White et al. (2013). Step 1 - take two natural subgroups - for example, chloroplast sequences from red and green algae. Step 2 – independently align the subgroups; step 3 - infer the subtrees for each independent subset, and step 4 independently generate the ancestral sequences a_x and a_y for each subtree. The program PAML was used to generate the ancestral sequences; it is robust to small changes in the tree and is a well-established method (Yang, 2007). Step 5 is to calculate the alignment scores between all sets of sequences (that is using one sequence from each of the two subgroups) using the program MUSCLE (Edgar, 2004).This compares a sequence from say subgroup X with all other sequences from subgroups X and Y as well as the inferred ancestral sequences a_x and a_y. The resulting distributions (Step 6) of between alignment scores are used to calculate the probability (Step

7) of observing scores at least as high as the ancestral score under the null model. The null model is that the similarity between the ancestral sequences is equal to the similarity between the extant sequences which in effect would imply that there is no evolution happening.

Acknowledgements

This work was supported by the New Zealand Marsden Fund

.

Literature Cited

Brocchieri, L. & Karlin, S. 2005. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research,* 33**,** 3390-3400.

Caetano-Anollés, G., Yafremava, L. S., Gee, H., Caetano-Anollés, D., Kim, H. S. & Mittenthal, J. E. 2009. The origin and evolution of modern metabolism. *The International Journal of Biochemistry & Cell Biology,* 41**,** 285-297. Available: DOI DOI: 10.1016/j.biocel.2008.08.022.

Collins, L. J., Poole, A. M. & Penny, D. 2003. Using ancestral sequences to uncover potential gene homologues. *Applied Bioinformatics* 2**,** S85-S95.

Daly, T.K., Sutherland-Smith A.J., & Penny, D. (2013) In silico resurrection of the Major Vault Protein suggests it is ancestral in modern eukaryotes. *Genome Biology and Evolution* 5: 1567-1583

Dorrell, R. G. & Howe, C. J. 2012. What makes a chloroplast? Reconstructing the establishment of photosynthetic symbioses. *Journal of Cell Science,* 125**,** 1865-1875.

Edgar, R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics,* 5**,** 113.

Forst, C. V. & Schulten, K. 2001. Phylogenetic Analysis of Metabolic Pathways. *Journal of Molecular Evolution,* 52**,** 471-489. Available: DOI 10.1007/s002390010178.

Hahn, S. 2004. Structure and mechanism of the RNA polymerase II transcription machinery. *Nature Structual & Molecular Biology,* 11**,** 394-403. Available: DOI 10.1038/nsmb763.

Harms, M. J. & Thornton, J. W. 2010. Analyzing protein structure and function using ancestral gene reconstruction. *Current Opinion in Structural Biology,* 20**,** 360-366. Available: DOI http://dx.doi.org/10.1016/j.sbi.2010.03.005.

Hart, K. M., Harms, M. J., Schmidt, B. H., Elya, C., Thornton, J. W. & Marqusee, S. 2014. Thermodynamic System Drift in Protein Evolution. *PLoS Biol,* 12**,** e1001994. Available: DOI 10.1371/journal.pbio.1001994.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. & Yamanishi, Y. 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Research,* 36**,** D480-D484. Available: DOI 10.1093/nar/gkm882.

Kapp, L. D. & Lorsch, J. R. 2004. The molecular mechanics of eukaryotic translation. *Annual Review of Biochemistry,* 73**,** 657-704. Available: DOI 10.1146/annurev.biochem.73.030403.080419.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics,* 23**,** 2947-2948.

Leliaert, F., Verbruggen, H. & Zechman, F. W. 2011. Into the deep: New discoveries at the base of the green plant phylogeny. *BioEssays,* 33**,** 683-692. Available: DOI 10.1002/bies.201100035.

Martin, W. 2003. Gene transfer from organelles to the nucleus: Frequent and in big chunks. *Proceedings of the National Academy of Sciences,* 100**,** 8612-8614. Available: DOI 10.1073/pnas.1633606100.

Mossel, E. & Steel, M. 2005a. How much can evolved characters tell us about the tree that generated them? *In:* Gascuel, O. (ed.) *Mathematics of Evolution and Phylogeny.* Oxford University Press.

Mossel, E. & Steel, M. 2005b. Random biochemical networks: the probability of self-sustaining autocatalysis. *Journal of Theoretical Biology,* 233**,** 327-336. Available: DOI 10.1016/j.jtbi.2004.10.011.

Nevins, J. R. 1983. The Pathway of Eukaryotic mRNA Formation. *Annual Review of Biochemistry,* 52**,** 441-466. Available: DOI doi:10.1146/annurev.bi.52.070183.002301.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A. & Finn, R. D. 2011. The Pfam protein families database. *Nucleic Acids Research,* 40**,** D290-D301. Available: DOI 10.1093/nar/gkr1065.

Sober, E. & Steel, M. 2002. Testing the Hypothesis of Common Ancestry. *Journal of Theoretical Biology,* 218**,** 395-408. Available: DOI http://dx.doi.org/10.1006/jtbi.2002.3086.

Sun, H., Merugu, S., Gu, X., Kang, Y. Y., Dickinson, D. P., Callaerts, P. & Li, W.-H. 2002. Identification of Essential Amino Acid Changes in Paired Domain Evolution Using a Novel Combination of Evolutionary Analysis and In Vitro and In Vivo Studies. *Molecular Biology and Evolution,* 19**,** 1490-1500.

Thornton, J. W. 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat Rev Genet,* 5**,** 366-375.

Thornton, J. W., Need, E. & Crews, D. 2003. Resurrecting the Ancestral Steroid Receptor: Ancient Origin of Estrogen Signaling. *Science,* 301**,** 1714-1717. Available: DOI 10.1126/science.1086185.

Wang, M., Kurland, C. G. & Caetano-Anollés, G. 2011. Reductive evolution of proteomes and protein structures. *Proceedings of the National Academy of Sciences,* 108**,** 11954-11958.

White, W. T. J., Zhong, B. & Penny, D. 2013. Beyond Reasonable Doubt: Evolution from DNA Sequences. *PLoS ONE,* 8**,** e69924.

Wolf, Y. I. & Koonin, E. V. 2013. Genome reduction as the dominant mode of evolution. *BioEssays***,** n/a-n/a. Available: DOI 10.1002/bies.201300037.

Yang, Z. 2006. *Computational molecular evolution.* Oxford University Press Oxford.

Yang, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution,* 24**,** 1586-1591. Available: DOI 10.1093/molbev/msm088.

Zhang, J. 2000. Protein-length distributions for the three domains of life. *Trends in Genetics,* 16**,** 107-109.

Tables and Figures

Table 1: lists the position of the cyanobacterial homolog within the metabolic pathways of *Arabidopsis thaliana*. "Early" and "Midway" refer to the first and second thirds of a pathway. For example, if a pathway contains 14 enzymes progressively catalyzing the metabolite and the homolog is at the 4th position, then the enzymes position would be classed as early as it lies within the first third of the pathway.

Pathway	**Enzyme (EC no.)**	**Position**
Ascorbate VitC	1.13.99.1	midway
Folate B9	2.6.1.85	first
	3.5.4.16	early
Histidine	2.4.2.17	early
	4.2.1.19	midway
	4.1.1.2	end
	1.2.1.3	end
Isoleucine/Leucine/Valine	4.2.1.35	first
	LeuC	first
	LeuD	first
	LeuB-1.1.1.85	first
	4.3.1.19	first
	1.2.4.1	first
Leucine	2.3.3.13	early
	4.2.1.33	early
Lysine	1.2.1.11	early
	5.1.1.7	midway
Methionine	1.2.1.11	midway
	2.1.1.14	end
Pantothene(B5)	2.2.1.6	first
Threonine	4.2.3.1	end
	1.2.1.11	early
Tryptophan	4.2.1.20	midway
	4.1.3.27	midway
	4.1.1.48	midway

Table 2: Demonstrating the likelihood of ancestral convergence between groups of organisms. The data for the first four columns are from White, et al., (2013).

Data Type	Group X	Group Y	Divergence Times	X^2	d.f.	$p(X^2)$	X^2(control)	$p(X^2)$(control)
Chloroplast	Eudicot	Monocot	125mya	289.058	102	1.94E-19	93.69	7.09E-01
Chloroplast	Angiosperm	Gymnosperm	305mya	363.527	104	1.23E-29	85.647	9.05E-01
Chloroplast	Seed plant	Fern	390mya	457.118	102	1.69E-44	100.451	5.25E-01
Chloroplast	Streptophyta	Chlorophyta	700mya	300.162	94	2.23E-23	90.982	5.69E-01
Chloroplast	Red Algae	Green Algae	~ 1000mya	341.014	82	2.60E-34	70.928	8.03E-01
Chloroplast	Algae	Cyanobacteria	~ 1500mya	231.079	56	3.90E-23	62.718	2.50E-01
Nuclear	Algae	Cyanobacteria	~ 1500mya	70.479	34	2.30E-04	39.342	2.43E-01

Table 3: List of metabolic pathways analysed for cyanobacterial homology detection

Amino Acid	Cofactor (Vitamin)
Histidine	Ascorbate (C)
Isoleucine	Coenzyme A
Leucine	Folic acid (B9)
Lysine	Niacin (B3) NAD/NADP
Methionine	Pantothenic acid (B5)
Phenyalanine	Pyridoxine (B6)
Threonine	Riboflavin (B2)
Tryptophan	Thiamine (B1)
Valine	Biotin (B7)

Figure 1: illustrating the cumulative frequency plots of pairwise alignment scores for three different photosynthetic genes from red and green algae. The x-axis shows the alignment score, which increases with the length of the protein(s). The y-axis shows for three proteins distribution of values for pairs of taxa selected so that one member is from one subgroup, red algae, and the other member is from the second subgroup, green algae. The circle represents the alignment score between the reconstructed ancestral sequences of each group. If the score is high then this indicates that the ancestral red alga is more similar to the ancestral green alga than most extant red algae are to most extant green algae. As expected, the ancestral sequences are more similar the longer the length of the sequences.

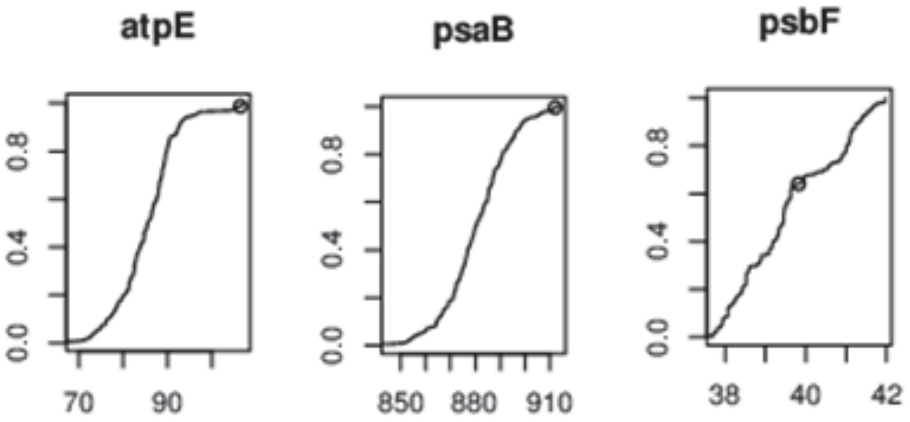
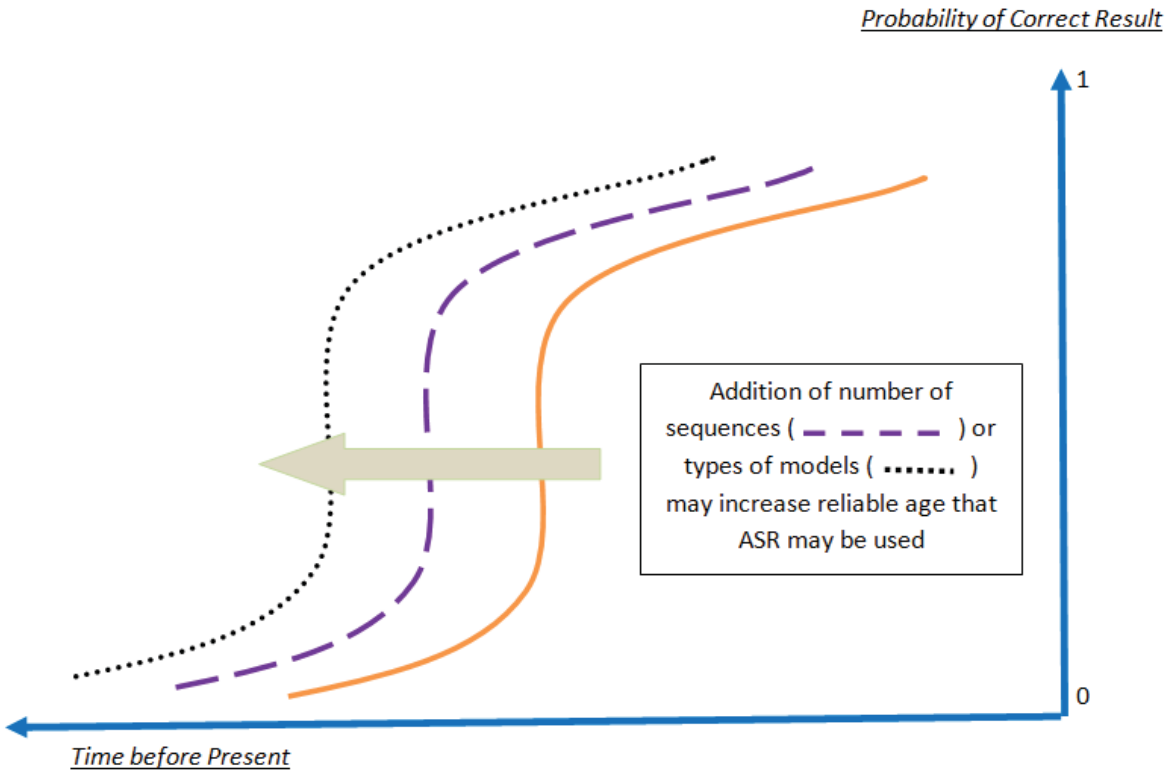


Figure 2: Reliability of Markov modelling with addition of other models. As more sequences or combination of models are added, there is an increase in the time before present that the Markov model may be reliably used, as illustrated by the large arrow. Any increase in reliability is one of linear (addition of models) versus exponential decay (artefact of Markov modelling) so that the effect of the addition of sequences or model combinations is marginal.


Probability of Correct Result
1
0
Addition of number of sequences (– – – –) or types of models (.....) may increase reliable age that ASR may be used
Time before Present

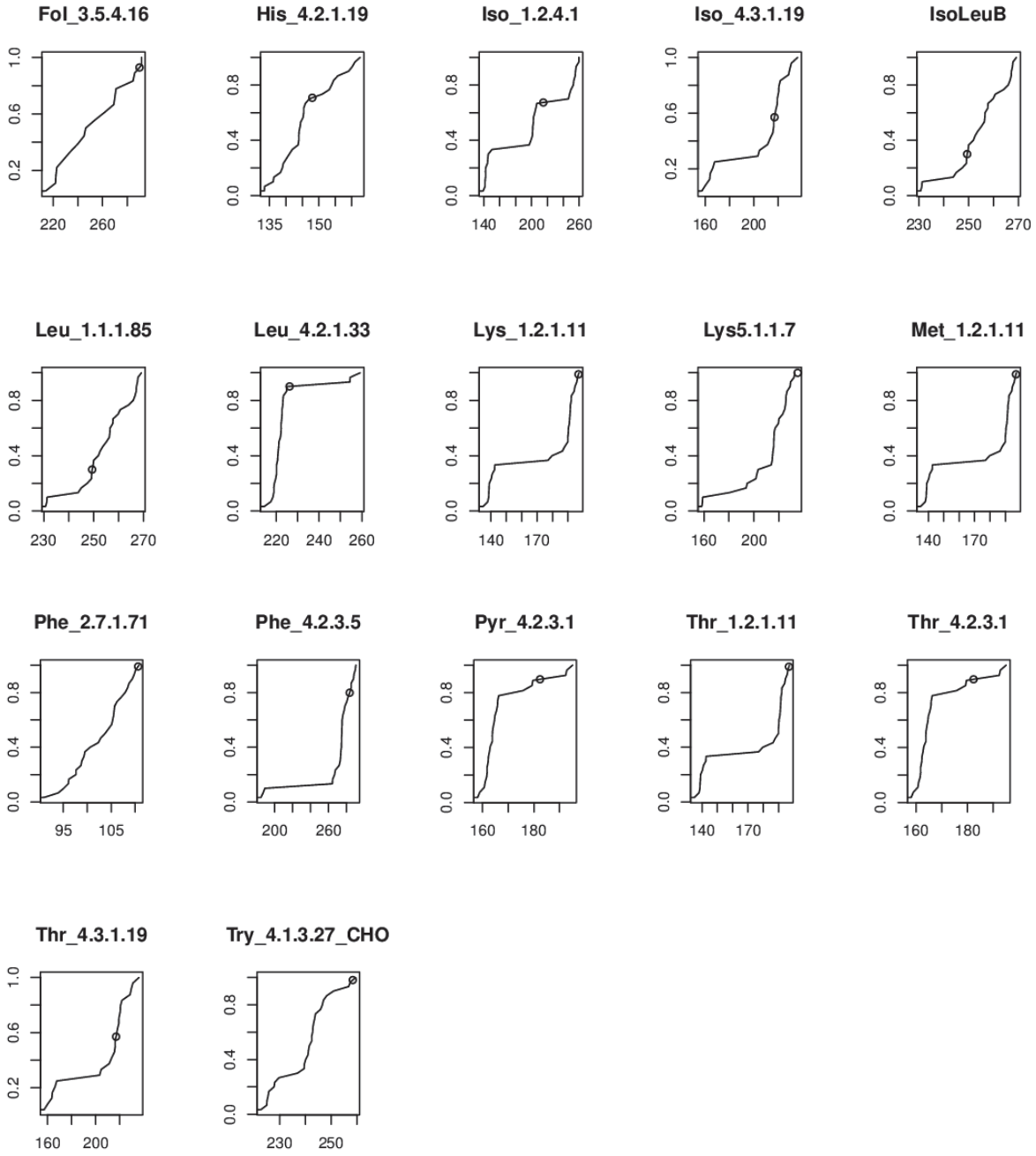
Supplementary Tables

Algal Nuclear Genes compared to their Cyanobacterial Homologs

Table 1: Individual gene results in tabular form

Gene	Ancestral alignment score	Number of higher alignment scores	Total number of alignment scores	Proportion of higher alignment scores	Chi-squared term
Fol_3.5.4.16	289.9	2	18	0.111111	4.394449
His_4.2.1.19	148	9	30	0.3	2.407946
Iso_1.2.4.1	214.9	10	30	0.333333	2.197225
Iso_4.3.1.19	217	11	24	0.458333	1.560317
IsoLeuB	249.4	22	30	0.733333	0.62031
Leu_1.1.1.85	249.4	22	30	0.733333	0.62031
Leu_4.2.1.33	226.2	3	30	0.1	4.60517
Lys_1.2.1.11	196.8	1	30	0.033333	6.802395
Lys5.1.1.7	235.1	0	30	0.033333	6.802395
Met_1.2.1.11	196.8	1	30	0.033333	6.802395
Phe_2.7.1.71	110.6	1	30	0.033333	6.802395
Phe_4.2.3.5	283.6	7	30	0.233333	2.910574
Pyr_4.2.3.1	182.4	3	27	0.111111	4.394449
Thr_1.2.1.11	196.8	1	30	0.033333	6.802395
Thr_4.2.3.1	182.4	3	27	0.111111	4.394449
Thr_4.3.1.19	217	11	24	0.458333	1.560317
Try_4.1.3.27_CHO	258.4	1	30	0.033333	6.802395
Total chi-squared value:					70.47988
Degrees of freedom:					34
Overall p (Fisher's Method):					**0.000237**

Figure 1: Individual gene results in graphic form

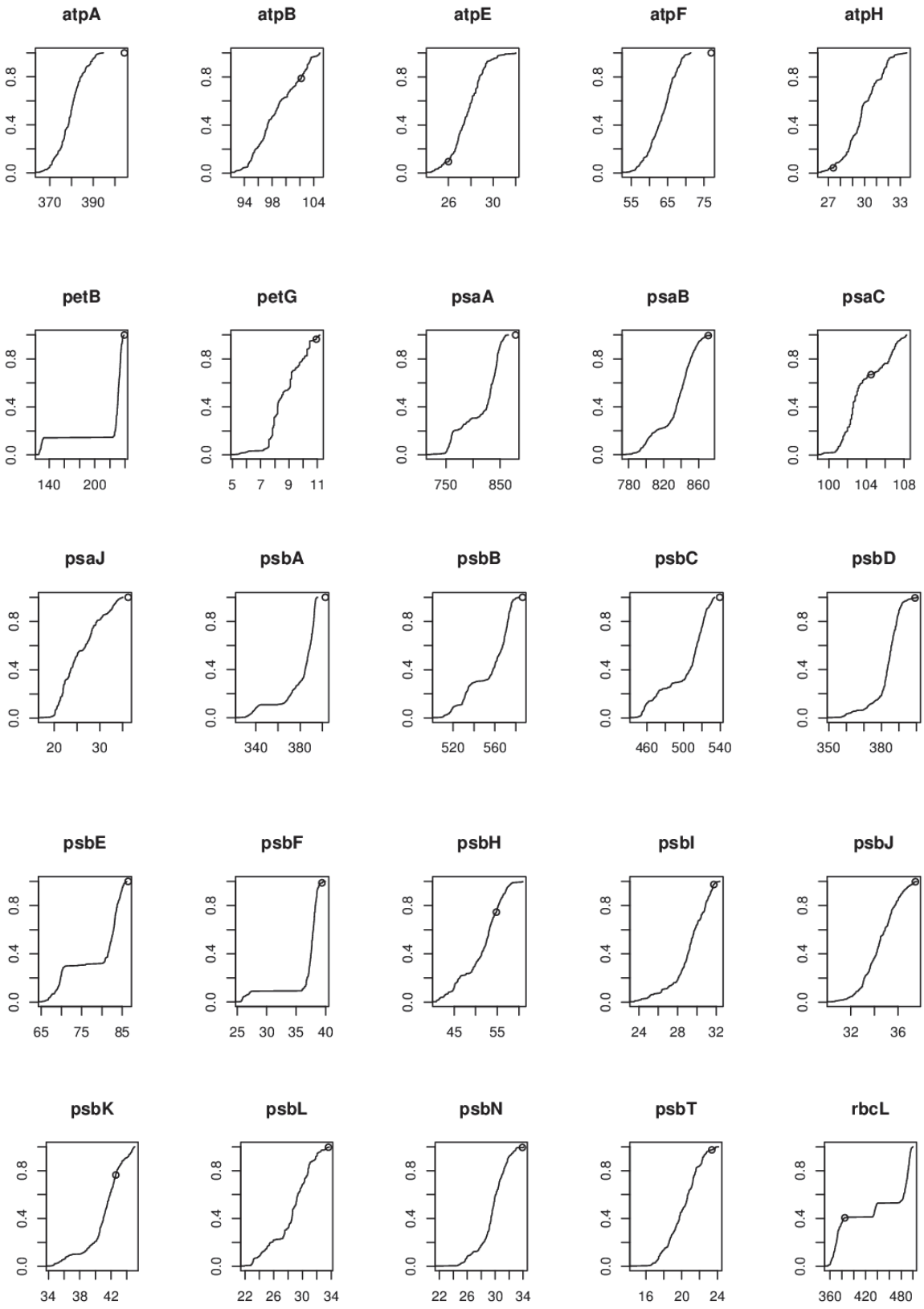


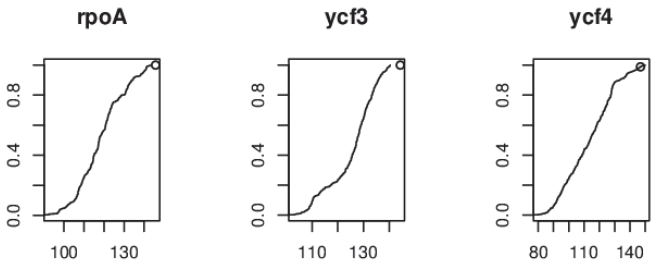
Algal Chloroplast Genes compared to their Cyanobacterial Homologs

Figure 2: Individual gene results graphical

Gene	Ancestral alignment score	Number of higher alignment scores	Total number of alignment scores	Proportion of higher alignment scores	Chi-squared term
atpA	404.3	0	234	0.004274	10.91064
atpB	102.2	34	156	0.217949	3.046991
atpE	26	177	195	0.907692	0.1937
atpF	77	0	195	0.005128	10.546
atpH	27.4	150	156	0.961538	0.078441
petB	238.9	0	273	0.003663	11.21894
petG	10.93	12	312	0.038462	6.516193
psaA	878.6	0	390	0.002564	11.93229
psaB	871	3	546	0.005495	10.40801
psaC	104.5	78	234	0.333333	2.197225
psaJ	36.2	0	234	0.004274	10.91064
psbA	402.7	0	1443	0.000693	14.54896
psbB	586.4	0	390	0.002564	11.93229
psbC	539.5	0	390	0.002564	11.93229
psbD	399.3	4	663	0.006033	10.22096
psbE	86.49	0	390	0.002564	11.93229
psbF	39.38	6	429	0.013986	8.539395
psbH	54.79	100	390	0.25641	2.721953
psbI	31.74	11	390	0.028205	7.136503
psbJ	37.46	1	312	0.003205	11.48601
psbK	42.61	93	390	0.238462	2.867094
psbL	33.63	2	351	0.005698	10.33528
psbN	33.91	3	390	0.007692	9.735069
psbT	23.37	11	390	0.028205	7.136503
rbcL	385	232	390	0.594872	1.038819
rpoA	145.7	0	234	0.004274	10.91064
ycf3	144.3	0	390	0.002564	11.93229
ycf4	146.9	5	390	0.012821	8.713418

Total chi-squared value:	231.0789
Degrees of freedom:	56
Overall p (Fisher's Method):	**3.92E-23**


atpA
0.0 0.4 0.8
370 390
atpB
0.0 0.4 0.8
94 98 104
atpE
0.0 0.4 0.8
26 30
atpF
0.0 0.4 0.8
55 65 75
atpH
0.0 0.4 0.8
27 30 33
petB
0.0 0.4 0.8
140 200
petG
0.0 0.4 0.8
5 7 9 11
psaA
0.0 0.4 0.8
750 850
psaB
0.0 0.4 0.8
780 820 860
psaC
0.0 0.4 0.8
100 104 108
psaJ
0.0 0.4 0.8
20 30
psbA
0.0 0.4 0.8
340 380
psbB
0.0 0.4 0.8
520 560
psbC
0.0 0.4 0.8
460 500 540
psbD
0.0 0.4 0.8
350 380
psbE
0.0 0.4 0.8
65 75 85
psbF
0.0 0.4 0.8
25 30 35 40
psbH
0.0 0.4 0.8
45 55
psbI
0.0 0.4 0.8
24 28 32
psbJ
0.0 0.4 0.8
32 36
psbK
0.0 0.4 0.8
34 38 42
psbL
0.0 0.4 0.8
22 26 30 34
psbN
0.0 0.4 0.8
22 26 30 34
psbT
0.0 0.4 0.8
16 20 24
rbcL
0.0 0.4 0.8
360 420 480


rpoA
ycf3
ycf4
0.0
0.4
0.8
100
130
110
130
80
110
140

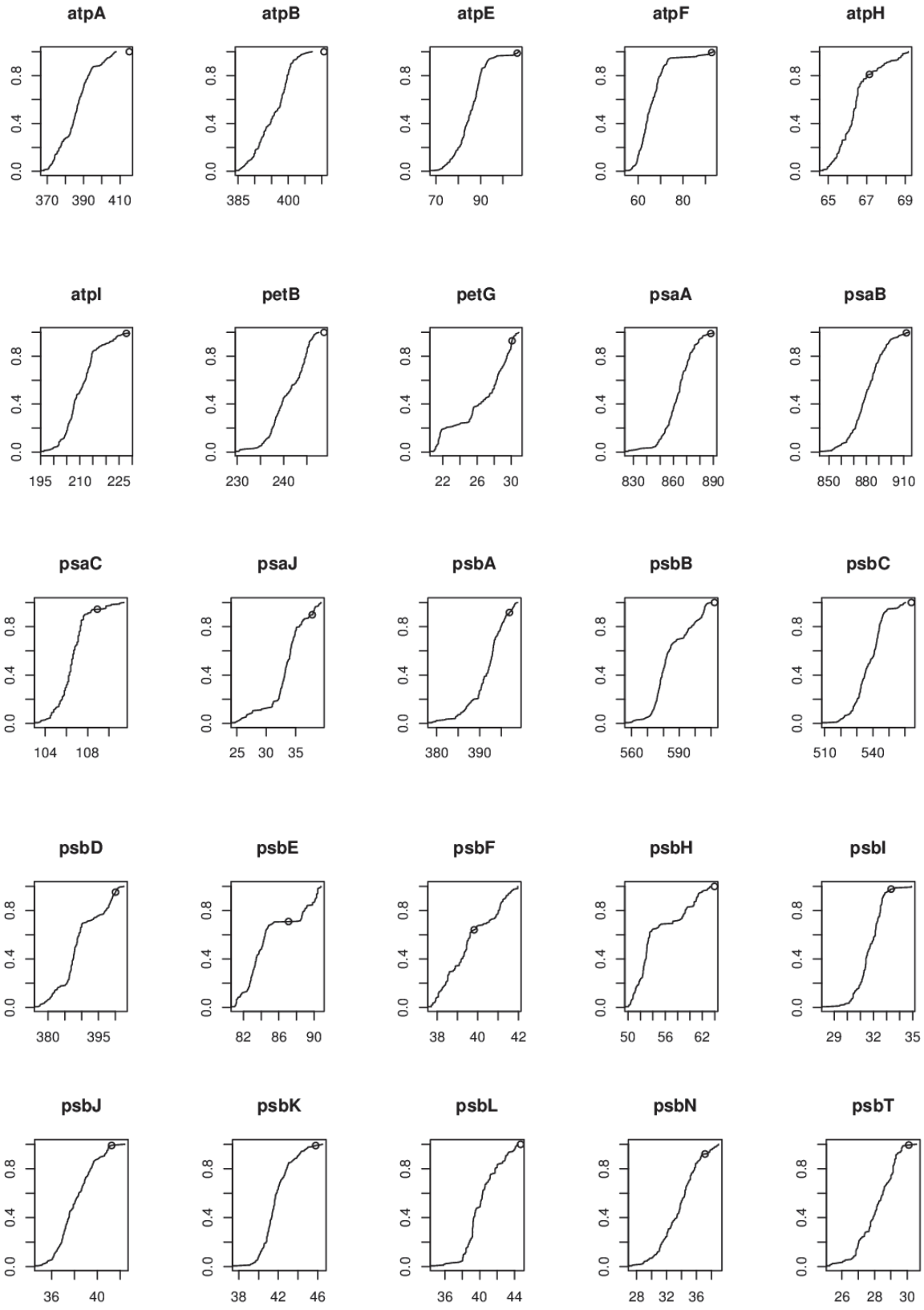
Red Algal Chloroplast Genes compared to Green Algal Chloroplast Genes

Table 3: Individual gene results in tabular form

Gene	Ancestral alignment score	Number of higher alignment scores	Total number of alignment scores	Proportion of higher alignment scores	Chi-squared term
atpA	415.1	0	170	0.005882	10.2716
atpB	410.7	0	170	0.005882	10.2716
atpE	106.4	3	170	0.017647	8.074372
atpF	92.82	2	170	0.011765	8.885303
atpH	67.14	33	170	0.194118	3.278582
atpI	227.8	2	170	0.011765	8.885303
petB	248.7	0	170	0.005882	10.2716
petG	30.08	13	170	0.076471	5.141698
psaA	888.1	2	170	0.011765	8.885303
psaB	912.1	1	170	0.005882	10.2716
psaC	108.9	10	170	0.058824	5.666427
psaJ	37.85	18	170	0.105882	4.490853
psbA	396.9	15	170	0.088235	4.855496
psbB	612.2	0	170	0.005882	10.2716
psbC	564	0	170	0.005882	10.2716
psbD	400.1	9	170	0.052941	5.877148
psbE	87.11	50	170	0.294118	2.447551
psbF	39.82	62	170	0.364706	2.017328
psbH	63.96	0	170	0.005882	10.2716
psbI	33.34	4	170	0.023529	7.499008
psbJ	41.24	2	170	0.011765	8.885303
psbK	45.8	2	170	0.011765	8.885303
psbL	44.69	0	170	0.005882	10.2716
psbN	37.14	14	170	0.082353	4.993482
psbT	30.09	2	170	0.011765	8.885303
rbcL	500.2	0	170	0.005882	10.2716
rpl14	100.4	1	170	0.005882	10.2716
rpl16	137.4	0	170	0.005882	10.2716
rpl2	229.2	0	170	0.005882	10.2716
rpl20	88.75	0	170	0.005882	10.2716
rpl36	43.04	0	170	0.005882	10.2716
rps11	108.2	0	170	0.005882	10.2716
rps14	85.95	30	170	0.176471	3.469202
rps18	30.23	0	170	0.005882	10.2716

rps19	79.46	0	170	0.005882	10.2716
rps2	207.5	0	170	0.005882	10.2716
rps3	173.9	0	170	0.005882	10.2716
rps4	145.5	3	170	0.017647	8.074372
rps7	125.7	0	170	0.005882	10.2716
rps8	102.7	1	170	0.005882	10.2716
ycf4	138	8	170	0.047059	6.112714

Total chi-squared value:	341.0136
Degrees of freedom:	82
Overall p (Fisher's Method):	**2.65E-33**


atpA
0.0 0.4 0.8
370 390 410
atpB
0.0 0.4 0.8
385 400
atpE
0.0 0.4 0.8
70 90
atpF
0.0 0.4 0.8
60 80
atpH
0.0 0.4 0.8
65 67 69
atpI
0.0 0.4 0.8
195 210 225
petB
0.0 0.4 0.8
230 240
petG
0.0 0.4 0.8
22 26 30
psaA
0.0 0.4 0.8
830 860 890
psaB
0.0 0.4 0.8
850 880 910
psaC
0.0 0.4 0.8
104 108
psaJ
0.0 0.4 0.8
25 30 35
psbA
0.0 0.4 0.8
380 390
psbB
0.0 0.4 0.8
560 590
psbC
0.0 0.4 0.8
510 540
psbD
0.0 0.4 0.8
380 395
psbE
0.0 0.4 0.8
82 86 90
psbF
0.0 0.4 0.8
38 40 42
psbH
0.0 0.4 0.8
50 56 62
psbI
0.0 0.4 0.8
29 32 35
psbJ
0.0 0.4 0.8
36 40
psbK
0.0 0.4 0.8
38 42 46
psbL
0.0 0.4 0.8
36 40 44
psbN
0.0 0.4 0.8
28 32 36
psbT
0.0 0.4 0.8
26 28 30

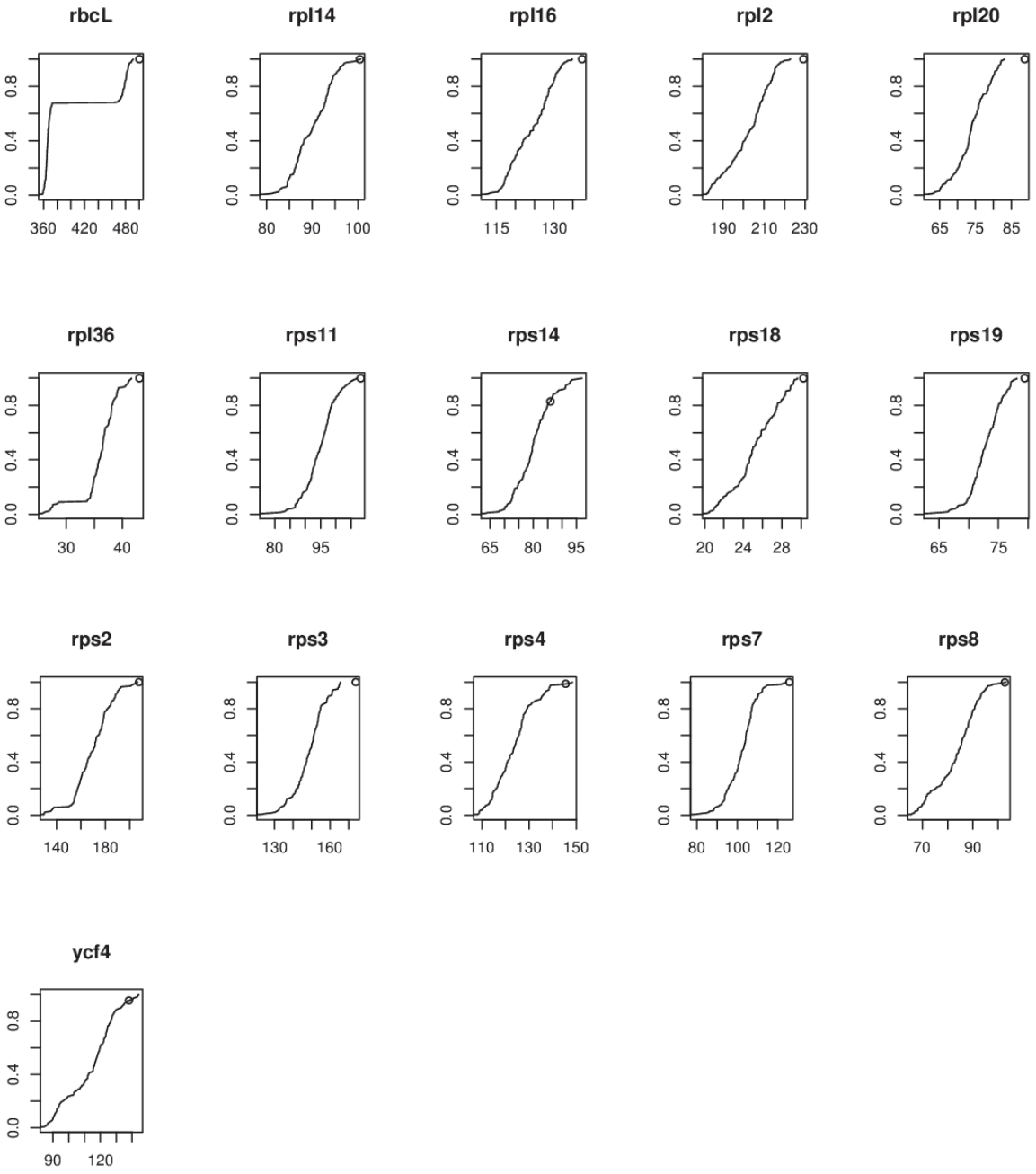


Figure 3: Individual genes in graphic form

Chapter 6. Summary, Conclusions and Future Directions

Summary

This thesis has been about looking for the bacterial signature of endosymbionts in eukaryotic genomes in the hope of finding the bacterial partner in the fusion/endosymbioses that led to both LECA and to the origin of the plastid. I had expected that with the use of ASR, previously uncovered homologies would be discovered that would clarify the bacterial source of the operational genes in the eukaryotic genome.

Starting initially with nucleotides, ASR did not appear to uncover any deeper pattern of single bacterial homology for plastids. Amino acids were explored next, and here I found some very interesting results to that lead to the conclusion that there is a difference between genes retained in the chloroplast (plastid) and genes transferred to the nucleus. The latter become more like 'eukaryotic genes', and gain an exon/intron structure and appear to have more extensive loop regions – making alignment ambiguous. Affecting this process were –

- the different transcriptional environments in prokaryotes and eukaryotes
- tree-building artefacts, particularly the depth at which HMM can reliably be used to model evolution
- the presence of multiple copies (isozymes) of the same enzyme in the same organism.

Adding to these factors was the failure to excise the cTP (chloroplast transit peptides) for the sequences analysed in chapters two and three. While the second analysis in chapter four showed that the lack of cTP excision had an effect on the results, this result was relatively minor for the two enzymes analysed.

The lack of manual curation of the MSA (multiple sequence alignments) also undoubtedly had an effect on the "negative results" in chapters two and three. Whether these initial methodological errors were the sole reasons behind this result is debateable. Results from the two analyses in chapter four, where these errors were corrected for, did not provide a definitive positive result – that of ASR sequences being phylogenetically closer to bacteria than *A.thaliana*- across the enzymes analysed.

The data mining conducted in chapter two, for potential cyanobacterial homologs in the pathways of essential amino acids and vitamins, shortlisted a dataset of the enzymes with the strongest potential cyanobacterial homologies for further analysis in chapter 3. The chapter 3 eukaryote datasets had the most potential for the lack of cTP excision to affect the results. That there remained enough signal for the chapter five analyses to prove that there is ancestral convergence between chloroplast

and cyanobacterial homologs and between eukaryote nuclear genes and their cyanobacterial counterparts, shows that there was enough of a signal in the data for ASR to work despite the early methodological errors.

What has been shown is that ASR does work very well when there are comparisons between enzymes from the same transcriptional environment- chloroplast to chloroplast, chloroplast to cyanobacteria - but less well, though still significant – cyanobacteria to nuclear encoded enzymes.

Where next

One avenue for future study is to use iTASSER (Zhang, 2008) to predict the 3-dimensional structure of the enzymes in this study. There is the suggestion that it is the structure of the enzymes that is conserved rather than the genetic code itself when enzymes are recruited (Caetano-Anollés *et al.*, 2009) and that 3D structures may be especially useful when genes are of a short length (Kurland and Harish, 2013, unpublished). This process was initiated about six months ago through the NeSi multi-computer site at the University of Auckland. iTASSER has large data and computing requirements and it is this issue, as well as multiple technical issues, that has meant that I have no results yet from this line of enquiry. But it is, however, a promising approach.

Another avenue is to research the evolution of metabolic pathways – it will be interesting to compare the position of the cyanobacterial homologs found in the pathways within archaeplastida with enzymes that hold the same position/function within non-plastid bearing eukaryotic genomes. This may help to establish a pattern of pathway evolution and shed some light on the question of the placement of cyanobacterial homologs in the pathways studied; this is discussed at more length later in this chapter.

Finally, there is the whole question of protein evolution, and change of function. With many complete genomes becoming available it is a good time to follow protein (particularly enzyme) evolution. Perhaps it has been assumed that a particular lineage of proteins retains its function “forever”, but it is now becoming obvious that protein evolution is a far more dynamic process with gene duplication, multiple reactions by a single enzyme, new proteins arising, etc., occurring much of the time.

Percentage of bacterial and archaeal genes in eukaryotes

Eukaryotes possess more bacterial-related genes than they possess archaeal -related genes. Of the 5,833 human proteins that have homologs in these prokaryotes at specified thresholds, 4,788 (80%) have greater sequence identity with bacterial homologs, whereas 877 (15%) are more similar to archaeal homologs (196 are ties) (Dagan and Martin, 2006). This is usually ‘taken at face value’, and

assumed to reflect some 'fusion' hypothesis (see LECA literature review, chapter 2). However, it is still possible that this observation arises from the new cofactors that are much more common in Archaea?

This is an interesting point. Many eukarygenesis theories posit some kind of fusion or endosymbiosis between an archaea (the source of informational genes) and a bacterium (the source of operational genes) with the origin of bacterial cells being with the mitochondrial endosymbiont most closely related to the Rickettsiales, an α-proteobacterium. If 80% of eukaryotic genes have a bacterial homology (assumes that this 80/20 relationship holds across most eukaryotes) and that the origin of these genes is from an endosymbiotic or fusion event with a single bacterium, then surely which bacterial family this individual came from should be more traceable.

I suggest that this thesis shows that there are not strong bacterial homologs from an individual bacterial family; that is, data shows homologs from all over the bacterial kingdom and this is usually explained away as the effect of lateral gene transfer between bacterial families. Maybe this lack of a traceable family points to eukarya being much older than currently accepted and may add weight to eukarya being the ancestral condition of all Life as discussed in Chapter 2/Another Possibility.

Future work in trying to isolate a set of bacterial genes that were partner in the formation of the eukaryotic lineage (if indeed this is what happened) is at this stage beyond our current capabilities. The vast array of bacterial families, listed in the appendix, Table 1, indicates that no one bacterial lineage stands out as the likely partner in the fusion/endosymbiotic event that might lead to the eukaryotes or to the algae and plant lineage.

The Scientific Process

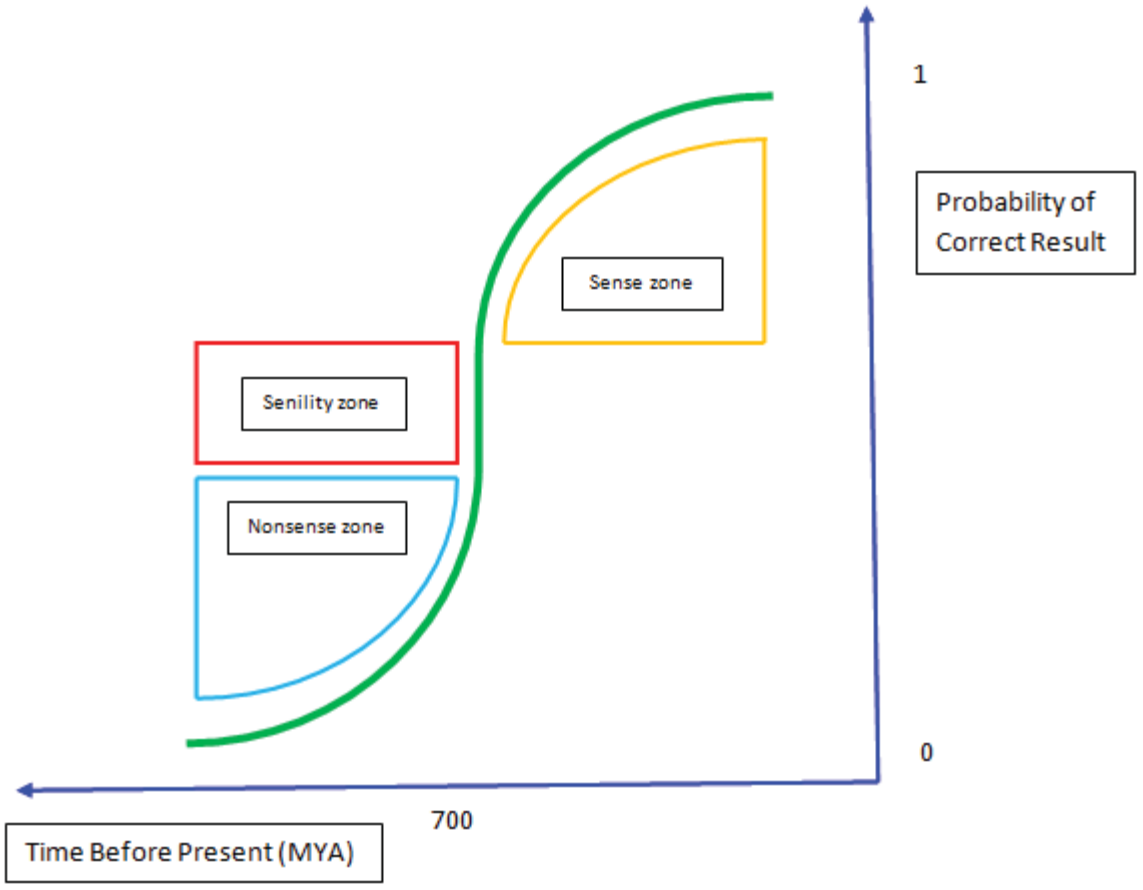
Of course, the tools we use to study phylogenies may also be part of the reason that no individual bacterial family has been discovered as the source of bacterial homologies in eukaryotes. The discussion about Hidden Markov Models (HMM) discussed in Chapter 3 /Markov Models and the depth at which these models can reliably predict ancestral states is also pertinent. We are knocking at the boundaries of the "senility zone" with the work that I have completed as shown in Figure 6.1 below. I am sure that in time researchers will develop the tools to be able to reliably look further back; after all, this has been the process in the past with current scientists standing on the shoulders of those that have gone before.

This process has been well documented in the phylogenetic field with more information and new techniques both leading to fully resolved phylogenetic trees. In the case of the placement of passerines (perching and/or song birds) in the avian tree (Harrison *et al.*, 2004), some initial analyses

placed this group as basal; with the addition of four new genomes and an improved method of phylogenetic analysis specific to vertebrate mitochondrial genomes the tree was resolved. Similar events happened for the resolution of the placental mammalian tree (Lin *et al.*, 2002) and probably the most famous case of Microsporidia being placed as basal in the eukaryote tree (Embley and Hirt, 1998) until discovery was made of the remnants of mitochondria within these organisms.

Future work in this line of scholarship would depend on new models being developed that extend the time period that ancestral sequences may be reliably produced. There is a lot of potential in this aspect, but the next step could be using 3D structures of proteins to help with selecting the most highly conserved regions of extant genes to be aligned.

Figure 6-1- Illustrating the decrease in the probability of finding a correct result at deeper times when using HMM to model evolutionary mutation rates. The "senility zone" represents time periods when HMM no longer works effectively.



Location of Cyanobacterial homologs in biosynthetic pathways

The original question that drove the funding of this thesis is the nature of the cell (either predator or prey) that took up the cyanobacterium to form the plastid. This can be addressed by the location of cyanobacterial homologs in the biosynthetic pathways analysed in this thesis. 56% of the homologs

are found in early or initial positions in the pathways (see Table 6.4 below); this percentage rises to 72% when homologs found in the late or final positions of these metabolic pathways are added. These positions in the pathways could suggest that pre LECA was more autotrophic (prey) than heterotrophic (predator) as the pathways studied are not full of cyanobacterial homologs but it does raise questions as to the process of pathway degradation when metabolites are not needed to be constructed *de novo*; presumably, degradation starts at the ends of the pathway working back to the core. This is one avenue for future study.

Some points need to be discussed about enzymes and metabolic pathways

- Metabolic pathways are thought to grow by recruitment of enzymes- either by gene duplication (Ohno, 1970), rearrangement of gene fragments (Vogel *et al.*, 2004) or by horizontal gene transfer (Pál *et al.*, 2005) or, as in this case, by gene transfer from the endosymbiont genome to the nucleus
- Enzymes may be either generalists or be highly specific(performing a single function) (Kacser and Beeby, 1984). Generalists are apparent from their inclusion in many pathways such as EC1.2.1.3 aldehyde dehydrogenase which has been discovered in 14 different pathways (see table 2.3)
- The pathways studied in this thesis are likely to be some of the most ancient and therefore highly conserved pathways within the three kingdoms; however the enzymes that make up these pathways can be much more plastic (Caetano-Anollés *et al.*, 2009). There is considerable diversity in metabolic pathways at the level of organisms with the enzymes being highly variable while their function is not (Forst and Schulten, 2001).

The location of these cyanobacterial homologs in the pathways studied could be

- replacements for a partially degraded pathway caused by redundancy due to a change in diet (indicates a predator lifestyle for LECA)
- or they may have been recruited over enzymes displaying a non-cyanobacterial homology due to superior fitness
- or they may be there by chance as evidence shows the enzymes are highly variable while their function is not.

Table 6-1- Position of cyanobacterial homologs in the biosynthetic pathways analysed for this thesis

Pathway	Enzyme	Position	Notes
Ascorbate VitC	1.13.99.1	Mid/side path	side path- Myo-inositol to D-Glucuronate
Folate B9	2.6.1.85	First/side path	1st step on side path
	3.5.4.16	Early	4 out of the first 6 enzymes in pathway
Histidine	2.4.2.17	Early	1st step in main path
	4.2.1.19	Midway	
	4.1.1.2	End	one of 2 enzymes(only plant one) that convert histamine to L-histidine
	1.2.1.3	End	last enzyme in plant pathway- past Histidine synthesis
Isoleucine/Leucine/Valine	4.2.1.35	First	special circumstances-faint homology
	LeuC	First	special circumstances-faint homology
	LeuD	First	special circumstances-faint homology
	LeuB-1.1.1.85	First	special circumstances-faint homology
	4.3.1.19	First	special circumstances-strong homology
	1.2.4.1	First	special circumstances-strong homologies- see diagram of pathway
Leucine	2.3.3.13	Early	1st/4
	4.2.1.33	Early	2nd/4
Lysine	1.2.1.11	Early	
	5.1.1.7	Midway	
Methionine	1.2.1.11	Midway	
	2.1.1.14	End	one of two for plants - final step in L-Methionine synthesis
Pantothene(B5)	2.2.1.6	First	
Threonine	4.2.3.1	End	5th/5
	1.2.1.11	Early	2nd/5
Tryptophan	4.2.1.20	Midway	all mid but there is an enzyme missing from pathway linking Prephate to Tyrosine
	4.1.3.27	Midway	
	4.1.1.48	Midway	

There appears to be a pattern of some kind (72% at early or late positions) that presents the pathway location of these cyanobacterial homologs. Further study and experimentation is needed in order to discover the nature of this pattern.

Rising enzyme lengths and multiple enzymes

There is a general trend within this study of increasing enzyme lengths within the archaeplastida. Generally, the prokaryote lengths are similar but within the archaeplastida there is a trend for the shortest length enzymes to be from the algal species with a steady increase in length up to the dicotyledons. From the literature, this may be explained by eukaryotic complexity, compared with

prokaryotes, in terms of expression regulation, protein length and protein domain structure (He and Zhang, 2005).

In the current analysis there has been a pattern of multiple protein domains being present within a single enzyme, which may be one of the factors confounding direct analysis of these enzymes by ASR. Recent work postulates that after a gene duplication event (either whole or localised), genes are retained in the nucleus if the copied gene:

1. Divided the original genes workload (i.e. - expressed to a lesser degree or in a different tissue) - known as subfunctionalization. This has been current thought until recently when the time taken for a surplus gene to reach fixation in the genome was questioned especially given the tendency for natural selection to eliminate unnecessary genes.
2. Had potential to carry out more than one function i.e. carry out the main function but also have the capacity to carry out a secondary, non-essential function. If conditions changed and the non-essential function became essential then it would be beneficial for more of the genes product to be produced. If mutations developed that increased the efficiency of the gene, then fixation would be assured. As it turns out this process has been documented in bacteria (Näsvall *et al.*, 2012).

This may help explain the presence of multiple protein families discovered in some of the enzymes in this study. Not only is there a pattern of increasing length in archaeplastida but also an increasing number of protein families and number of isozymes; maybe these genes have been neo-subfunctionalized and have been fixed into the nucleus. Whatever the mechanisms, there is a large amount of scope here for further research!

We are reminded of the proverbial Chinese curse, 'may you live in interesting times' – these are very interesting times for following protein evolution.

References

Adey, N. B., Tollefsbol, T. O., Sparks, A. B., Edgell, M. H. & Hutchison, C. A. 1994. Molecular resurrection of an extinct ancestral promoter for mouse L1. *Proceedings of the National Academy of Sciences, U.S.A.,* 91**,** 1569-1573. Available: DOI 10.1073/pnas.91.4.1569.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology,* 215**,** 403-410. Available: DOI 10.1016/s0022-2836(05)80360-2.

Archibald, J. M. 2008. The eocyte hypothesis and the origin of eukaryotic cells. *Proceedings of the National Academy of Sciences, U.S.A.,* 105**,** 20049-20050. Available: DOI 10.1073/pnas.0811118106.

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O. & Pupko, T. 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research,* 40**,** W580-W584. Available: DOI 10.1093/nar/gks498.

Bairoch, A., Boeckmann, B., Ferro, S. & Gasteiger, E. 2004. Swiss-Prot: juggling between evolution and stability. *Briefings in bioinformatics,* 5**,** 39-55.

Ball, S., Colleoni, C., Cenci, U., Raj, J. N. & Tirtiaux, C. 2011. The evolution of glycogen and starch metabolism in eukaryotes gives molecular clues to understand the establishment of plastid endosymbiosis. *Journal of experimental botany,* 62**,** 1775-1801.

Bapteste, E., Charlebois, R., Macleod, D. & Brochier, C. 2005. The two tempos of nuclear pore complex evolution: highly adapting proteins in an ancient frozen structure. *Genome Biology,* 6**,** R85.

Blanc, G., Ogata, H., Robert, C., Audic, S., Suhre, K., Vestris, G., Claverie, J.-M. & Raoult, D. 2007. Reductive genome evolution from the mother of Rickettsia. *PLoS Genetics,* 3**,** e14.

Bridgham, J. T., Carroll, S. M. & Thornton, J. W. 2006. Evolution of Hormone-Receptor Complexity by Molecular Exploitation. *Science,* 312**,** 97-101. Available: DOI 10.1126/science.1123348.

Brinkmann, H. & Philippe, H. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Molecular Biology and Evolution,* 16**,** 817-825.

Brocchieri, L. & Karlin, S. 2005. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research,* 33**,** 3390-3400.

Brown, J. R. & Doolittle, W. F. 1997. Archaea and the prokaryote-to-eukaryote transition. *Microbiology and Molecular Biology Reviews,* 61**,** 456-502.

Caetano-Anollés, G. 2002. Evolved RNA secondary structure and the rooting of the universal tree of life. *Journal of Molecular Evolution,* 54**,** 333-345.

Caetano-Anollés, G., Kim, K. M. & Caetano-Anollés, D. 2012. The phylogenomic roots of modern biochemistry: origins of proteins, cofactors and protein biosynthesis. *Journal of Molecular Evolution,* 74**,** 1-34.

Caetano-Anollés, G., Yafremava, L. S., Gee, H., Caetano-Anollés, D., Kim, H. S. & Mittenthal, J. E. 2009. The origin and evolution of modern metabolism. *The International Journal of Biochemistry & Cell Biology,* 41**,** 285-297. Available: DOI DOI: 10.1016/j.biocel.2008.08.022.

Cary, S. C., Mcdonald, I. R., Barrett, J. E. & Cowan, D. A. 2010. On the rocks: the microbiology of Antarctic Dry Valley soils. *Nature Reviews Microbiology,* 8**,** 129-138.

Cavalier-Smith, T. 2002a. Origins of the machinery of recombination and sex. *Heredity,* 88**,** 125-141.

Cavalier-Smith, T. 2002b. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *International Journal of Systematic and Evolutionary Microbiology,* 52**,** 297-354.

Chatton, E. 1925. *Pansporella perplexa: amœbien à spores protégées parasite des daphnies: réflexions sur la biologie et la phylogénie des protozoaires*, Masson.

Ciccarelli, F. D., Doerks, T., Von Mering, C., Creevey, C. J., Snel, B. & Bork, P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science,* 311**,** 1283-1287.

Collins, L. & Penny, D. 2005. Complex Spliceosomal Organization Ancestral to Extant Eukaryotes. *Molecular Biology and Evolution,* 22**,** 1053-1066. Available: DOI 10.1093/molbev/msi091.

Collins, L. J., Kurland, C. G., Biggs, P. & Penny, D. 2009. The Modern RNP World of Eukaryotes. *Journal of Heredity,* 100**,** 597-604. Available: DOI 10.1093/jhered/esp064.

Collins, L. J. & Lockhart, P. J. 2007. Evolutionary properties of sequences and ancestral state reconstruction. *In:* Liberles, D. A. (ed.) *Ancestral Sequence Reconstruction.* Oxford University Press.

Collins, L. J., Poole, A. M. & Penny, D. 2003. Using ancestral sequences to uncover potential gene homologues. *Applied Bioinformatics* 2**,** S85-S95.

Consortium, T. U. 2015. UniProt: a hub for protein information. *Nucleic Acids Research,* 43**,** D204-D212. Available: DOI 10.1093/nar/gku989.

Corradi, N. & Slamovits, C. H. 2011. The intriguing nature of microsporidian genomes. *Briefings in Functional Genomics,* 10**,** 115-124.

Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R. & Embley, T. M. 2008. The archaebacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences, U.S.A.,* 105**,** 20356-20361. Available: DOI 10.1073/pnas.0810647105.

Csuros, M., Rogozin, I. B. & Koonin, E. V. 2011. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS computational biology,* 7**,** e1002150.

Curtis, B. A. & Archibald, J. M. 2010. Problems and Progress in Understanding the Origins of Mitochondria and Plastids. *In:* Seckbach, J. & Grube, M. (eds.) *Symbioses and Stress.* Springer Netherlands. Available: DOI 10.1007/978-90-481-9449-0_3.

Dacks, J. B., Davis, L. a. M., Sjögren, Å. M., Andersson, J. O., Roger, A. J. & Doolittle, W. F. 2003. Evidence for Golgi bodies in proposed 'Golgi-lacking' lineages. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 270**,** S168-S171. Available: DOI 10.1098/rsbl.2003.0058.

Dacks, J. B. & Field, M. C. 2007. Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode. *Journal of Cell Science,* 120**,** 2977-2985. Available: DOI 10.1242/jcs.013250.

Dagan, T. & Martin, W. 2006. The tree of one percent. *Genome Biol,* 7**,** 118. Available: DOI 10.1186/gb-2006-7-118.

Dagan, T. & Martin, W. 2007. Testing hypotheses without considering predictions. *BioEssays,* 29**,** 500-503.

Daly, T. K., Sutherland-Smith, A. J. & Penny, D. 2013. In silico resurrection of the Major Vault Protein suggests it is ancestral in modern eukaryotes. *Genome Biology and Evolution,* 5**,** 1567-1583.

De Bodt, S., Maere, S. & Van De Peer, Y. 2005. Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution,* 20**,** 591-597. Available: DOI 10.1016/j.tree.2005.07.008.

De La Torre, F., Cañas, R. A., Pascual, M. B., Avila, C. & Cánovas, F. M. 2014. Plastidic aspartate aminotransferases and the biosynthesis of essential amino acids in plants. *Journal of experimental botany,* 65**,** 5527-5534.

De Nooijer, S., Holland, B. R. & Penny, D. 2009. The Emergence of Predators in Early Life: There was No Garden of Eden. *PLoS ONE,* 4**,** e5507.

Devos, D., Dokudovskaya, S., Alber, F., Williams, R., Chait, B. T., Sali, A. & Rout, M. P. 2004. Components of Coated Vesicles and Nuclear Pore Complexes Share a Common Molecular Architecture. *PLoS Biol,* 2**,** e380. Available: DOI 10.1371/journal.pbio.0020380.

Devos, D., Dokudovskaya, S., Williams, R., Alber, F., Eswar, N., Chait, B. T., Rout, M. P. & Sali, A. 2006. Simple fold composition and modular architecture of the nuclear pore complex. *Proceedings of the National Academy of Sciences, U.S.A. ,* 103**,** 2172-2177.

Diallo, A. B., Makarenkov, V. & Blanchette, M. 2010. Ancestors 1.0: a web server for ancestral sequence reconstruction. *Bioinformatics,* 26**,** 130-131.

Dorrell, R. G. & Howe, C. J. 2012. What makes a chloroplast? Reconstructing the establishment of photosynthetic symbioses. *Journal of Cell Science,* 125**,** 1865-1875.

Drevland, R. M., Waheed, A. & Graham, D. E. 2007. Enzymology and evolution of the pyruvate pathway to 2-oxobutyrate in Methanocaldococcus jannaschii. *Journal of bacteriology,* 189**,** 4391-4400.

Ebenhoh, O., Handorf, T. & Kahn, D. 2006. Evolutionary changes of metabolic networks and their biosynthetic capacities. *IEE Proceedings - Systems Biology,* 153**,** 354-358.

Edgar, R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics,* 5**,** 113.

Egel, R. & Penny, D. 2007. On the origin of meiosis in eukaryotic evolution: Coevolution of meiosis and mitosis from feeble beginnings. *In:* Egel, R. & Lankenau, D. H. (eds.) *Recombination and Meiosis.* Springer. Available: DOI 10.100717050_2007_036.

Emanuelsson, O., Nielsen, H. & Von Heijne, G. 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science,* 8**,** 978-984.

Embley, T. M. & Hirt, R. P. 1998. Early branching eukaryotes? *Current Opinion in Genetics & Development,* 8**,** 624-629.

Embley, T. M. & Martin, W. 2006. Eukaryotic evolution, changes and challenges. *Nature,* 440**,** 623-630.

Eme, L., Moreira, D., Talla, E. & Brochier-Armanet, C. 2009. A Complex Cell Division Machinery Was Present in the Last Common Ancestor of Eukaryotes. *PLoS ONE,* 4**,** e5021. Available: DOI 10.1371/journal.pone.0005021.

Field, M. C. & Dacks, J. B. 2009. First and last ancestors: reconstructing evolution of the endomembrane system with ESCRTs, vesicle coat proteins, and nuclear pore complexes. *Current Opinion in Cell Biology,* 21**,** 4-13. Available: DOI 10.1016/j.ceb.2008.12.004.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R. & Bateman, A. 2010. The Pfam protein families database. *Nucleic Acids Research,* 38**,** D211-D222. Available: DOI 10.1093/nar/gkp985.

Finnigan, G. C., Hanson-Smith, V., Stevens, T. H. & Thornton, J. W. 2012. Evolution of increased complexity in a molecular machine. *Nature,* 481**,** 360-364.

Fitzpatrick, D. & Tovar, E. M. M. 2011. Fungal Geneomics. *In:* Kavanagh, K. (ed.) *Fungi:Biology and Applications.* John Wiley & Sons, Ltd.

Forst, C. V. & Schulten, K. 2001. Phylogenetic Analysis of Metabolic Pathways. *Journal of Molecular Evolution,* 52**,** 471-489. Available: DOI 10.1007/s002390010178.

Forterre, P. & Philippe, H. 1999. Where is the root of the universal tree of life? *BioEssays,* 21**,** 871-879.

Foster, P. G. 2004. Modeling compositional heterogeneity. *Systematic biology,* 53**,** 485-495.

Fournier, G. P. & Alm, E. J. 2015. Ancestral Reconstruction of a Pre-LUCA Aminoacyl-tRNA Synthetase Ancestor Supports the Late Addition of Trp to the Genetic Code. *Journal of Molecular Evolution***,** 1-15. Available: DOI 10.1007/s00239-015-9672-1.

Fuerst, J. A. 2005. Intracellular Compartmentation in Planctomycetes. *Annual Review of Microbiology,* 59**,** 299-328. Available: DOI doi:10.1146/annurev.micro.59.030804.121258.

Fürst, P. & Stehle, P. 2004. What Are the Essential Elements Needed for the Determination of Amino Acid Requirements in Humans? *The Journal of Nutrition,* 134**,** 1558S-1565S.

Gabaldon, T. 2010. Peroxisome diversity and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 365**,** 765-773. Available: DOI 10.1098/rstb.2009.0240.

Gabaldon, T., Snel, B., Zimmeren, F. V., Hemrika, W., Tabak, H. & Huynen, M. 2006. Origin and evolution of the peroxisomal proteome. *Biology Direct,* 1**,** 8.

Gardner, P., Bateman, A. & Poole, A. 2010. SnoPatrol: how many snoRNA genes are there? *Journal of Biology,* 9**,** 4.

Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, M. F., Poole, R. J., Date, T. & Oshima, T. 1989. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proceedings of the National Academy of Sciences, U.S.A.,* 86**,** 6661-6665.

Goldfarb, K. C., Karaoz, U., Hanson, C. A., Santee, C. A., Bradford, M. A., Treseder, K. K., Wallenstein, M. D. & Brodie, E. L. 2011. Differential growth responses of soil bacterial taxa to carbon substrates of varying chemical recalcitrance. *Frontiers in microbiology,* 2.

Gribaldo, S. & Brochier-Armanet, C. 2006. The origin and evolution of Archaea: a state of the art. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 361**,** 1007-1022. Available: DOI 10.1098/rstb.2006.1841.

Gribaldo, S., Poole, A. M., Daubin, V., Forterre, P. & Brochier-Armanet, C. 2010. The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nat Rev Micro,* 8**,** 743-752.

Gruer, M. J., Artymiuk, P. J. & Guest, J. R. 1997. The aconitase family: three structural variations on a common theme. *Trends in Biochemical Sciences,* 22**,** 3-6. Available: DOI 10.1016/s0968-0004(96)10069-4.

Hahn, S. 2004. Structure and mechanism of the RNA polymerase II transcription machinery. *Nature Structual & Molecular Biology,* 11**,** 394-403. Available: DOI 10.1038/nsmb763.

Han, J., Collins, L. J., Biggs, P. J., White, W. T. & Penny, D. 2013. Are Signature Proteins the Key to Evaluating Eukaryotic Phylogeny? *Review of Bioinformatics and Biometrics,* 2.

Hanson-Smith, V., Kolaczkowski, B. & Thornton, J. W. 2010. Robustness of Ancestral Sequence Reconstruction to Phylogenetic Uncertainty. *Molecular Biology and Evolution,* 27**,** 1988-1999. Available: DOI 10.1093/molbev/msq081.

Harish, A., Tunlid, A. & Kurland, C. G. 2013. Rooted Phylogeny of the Three Superkingdoms. *Biochimie,* 95**,** 1593-1604.

Harms, M. J. & Thornton, J. W. 2010. Analyzing protein structure and function using ancestral gene reconstruction. *Current Opinion in Structural Biology,* 20**,** 360-366. Available: DOI http://dx.doi.org/10.1016/j.sbi.2010.03.005.

Harris, J. K., Kelley, S. T., Spiegelman, G. B. & Pace, N. R. 2003. The Genetic Core of the Universal Ancestor. *Genome Research,* 13**,** 407-412. Available: DOI 10.1101/gr.652803.

Harrison, G. A., Mclenachan, P. A., Phillips, M. J., Slack, K. E., Cooper, A. & Penny, D. 2004. Four new avian mitochondrial genomes help get to basic evolutionary questions in the late Cretaceous. *Molecular Biology and Evolution,* 21**,** 974-983.

Hart, K. M., Harms, M. J., Schmidt, B. H., Elya, C., Thornton, J. W. & Marqusee, S. 2014. Thermodynamic System Drift in Protein Evolution. *PLoS Biol,* 12**,** e1001994. Available: DOI 10.1371/journal.pbio.1001994.

He, X. & Zhang, J. 2005. Gene complexity and gene duplicability. *Current biology : CB,* 15**,** 1016-1021.

He, Y., Chen, B., Pang, Q., Strul, J. M. & Chen, S. 2010. Functional specification of Arabidopsis isopropylmalate isomerases in glucosinolate and leucine biosynthesis. *Plant and Cell Physiology,* 51**,** 1480-1487. Available: DOI 10.1093/pcp/pcq113.

Huson, D. H. & Bryant, D. 2006. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution,* 23**,** 254-267. Available: DOI 10.1093/molbev/msj030.

Ikeuchi, S. 2015. *RE: KEGG/GenomeNet.* Type to Cox, S. J. L.

Javaux, E. J., Marshall, C. P. & Bekker, A. 2010. Organic-walled microfossils in 3.2-billion-year-old shallow-marine siliciclastic deposits. *Nature,* 463**,** 934-938.

Jeffares, D. C., Mourier, T. & Penny, D. 2006. The biology of intron gain and loss. *Trends in Genetics,* 22**,** 16-22. Available: DOI 10.1016/j.tig.2005.10.006.

Jékely, G. 2003. Small GTPases and the evolution of the eukaryotic cell. *BioEssays,* 25**,** 1129-1138. Available: DOI 10.1002/bies.10353.

Jékely, G. 2007. Origin of Eukaryotic Endomembranes: A Critical Evaluation of Different Model Scenarios Eukaryotic Membranes and Cytoskeleton. Springer New York. Available: DOI 10.1007/978-0-387-74021-8_3.

Kacser, H. & Beeby, R. 1984. Evolution of catalytic proteins. *Journal of Molecular Evolution,* 20**,** 38-51.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. & Yamanishi, Y. 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Research,* 36**,** D480-D484. Available: DOI 10.1093/nar/gkm882.

Kanehisa, M. & Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research,* 28**,** 27-30. Available: DOI 10.1093/nar/28.1.27.

Kapp, L. D. & Lorsch, J. R. 2004. The molecular mechanics of eukaryotic translation. *Annual Review of Biochemistry,* 73**,** 657-704. Available: DOI 10.1146/annurev.biochem.73.030403.080419.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S. & Duran, C. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics,* 28**,** 1647-1649.

Knill, T., Reichelt, M., Paetz, C., Gershenzon, J. & Binder, S. 2009. Arabidopsis thaliana encodes a bacterial-type heterodimeric isopropylmalate isomerase involved in both Leu biosynthesis and the Met chain elongation pathway of glucosinolate formation. *Plant molecular biology,* 71**,** 227-239.

Knoll, A. H., Javaux, E. J., Hewitt, D. & Cohen, P. 2006. Eukaryotic organisms in Proterozoic oceans. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 361**,** 1023-1038.

Kopple, J. D. & Swendseid, M. E. 1975. Evidence that histidine is an essential amino acid in normal and chronically uremic man. *The Journal of clinical investigation,* 55**,** 881-91.

Kurland, C. G., Collins, L. J. & Penny, D. 2006. Genomics and the Irreducible Nature of Eukaryote Cells. *Science,* 312**,** 1011-1014. Available: DOI 10.1126/science.1121674.

Kurland, C. G. & Harish, A. 2013. Mayr versus Woese : Redefining Prokaryotes. Sweden: University of Lund.

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

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., Mcwilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. 2007a. Clustal W and Clustal X version 2.0. *Bioinformatics,* 23**,** 2947-2948.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., Mcwilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. 2007b. Clustal W and Clustal X version 2.0. *Bioinformatics,* 23**,** 2947-2948. Available: DOI 10.1093/bioinformatics/btm404.

Larsson, J., Nylander, J. A. & Bergman, B. 2011. Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. *BMC Evolutionary Biology,* 11**,** 187.

Leliaert, F., Verbruggen, H. & Zechman, F. W. 2011. Into the deep: New discoveries at the base of the green plant phylogeny. *BioEssays,* 33**,** 683-692. Available: DOI 10.1002/bies.201100035.

Lin, Y.-H., Waddell, P. J. & Penny, D. 2002. Pika and vole mitochondrial genomes increase support for both rodent monophyly and glires. *Gene,* 294**,** 119-129.

Lonhienne, T. G., Sagulenko, E., Webb, R. I., Lee, K.-C., Franke, J., Devos, D. P., Nouwens, A., Carroll, B. J. & Fuerst, J. A. 2010. Endocytosis-like protein uptake in the bacterium Gemmata obscuriglobus. *Proceedings of the National Academy of Sciences,* 107**,** 12883-12888.

Ludwig, W. & Schleifer, K.-H. 2005. Molecular Phylogeny of Bacteria based on Comparative Sequence Analysis of Conserved Genes. *In:* Sapp, J. (ed.) *Microbial phylogeny and evolution: concepts and controversies.* New York: Oxford University Press.

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., Pavlov, A., Pavlova, N., Karamychev, V. & Polouchine, N. 2006. Comparative genomics of the lactic acid bacteria. *Proceedings of the National Academy of Sciences,* 103**,** 15611-15616.

Makarova, K. S. & Koonin, E. V. 2007. Evolutionary genomics of lactic acid bacteria. *Journal of Bacteriology,* 189**,** 1199-1208.

Mans, B., Anantharaman, V., Aravind, L. & Koonin, E. V. 2004. Comparative Genomics, Evolution and Origins of the Nuclear Envelope and Nuclear Pore Complex. *Cell Cycle,* 3**,** 1625-1650.

Martin, W. 1999. A briefly argued case that mitochondria and plastids are descendants of endosymbionts, but that the nuclear compartment is not. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 266**,** 1387-1395. Available: DOI 10.1098/rspb.1999.0792.

Martin, W. 2003. Gene transfer from organelles to the nucleus: Frequent and in big chunks. *Proceedings of the National Academy of Sciences,* 100**,** 8612-8614. Available: DOI 10.1073/pnas.1633606100.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M. & Penny, D. 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 12246-12251. Available: DOI 10.1073/pnas.182432999.

Mayr, E. 1998. Two empires or three? *Proceedings of the National Academy of Sciences,* 95**,** 9720-9723. Available: DOI 10.1073/pnas.95.17.9720.

Merhej, V. & Raoult, D. 2011. Rickettsial evolution in the light of comparative genomics. *Biological Reviews,* 86**,** 379-405.

Morrison, D. A. 2009. Why would phylogeneticists ignore computerized sequence alignment? *Systematic biology,* 58**,** 150-158.

Mossel, E. & Steel, M. 2004. A phase transition for a random cluster model on phylogenetic trees. *Mathematical Biosciences,* 187**,** 189-203. Available: DOI 10.1016/j.mbs.2003.10.004.

Mossel, E. & Steel, M. 2005a. How much can evolved characters tell us about the tree that generated them? *In:* Gascuel, O. (ed.) *Mathematics of Evolution and Phylogeny.* Oxford University Press.

Mossel, E. & Steel, M. 2005b. Random biochemical networks: the probability of self-sustaining autocatalysis. *Journal of Theoretical Biology,* 233**,** 327-336. Available: DOI 10.1016/j.jtbi.2004.10.011.

Mount, D. W. 2008. Choosing a Method for Phylogenetic Prediction. *Cold Spring Harbor Protocols,* 2008**,** pdb.ip49. Available: DOI 10.1101/pdb.ip49.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology,* 247**,** 536-540.

Nakamura, T. M. & Cech, T. R. 1998. Reversing Time: Origin of Telomerase. *Cell,* 92**,** 587-590. Available: DOI 10.1016/s0092-8674(00)81123-x.

Näsvall, J., Sun, L., Roth, J. R. & Andersson, D. I. 2012. Real-Time Evolution of New Genes by Innovation, Amplification, and Divergence. *Science,* 338**,** 384-387. Available: DOI 10.1126/science.1226521.

Neumann, N., Lundin, D. & Poole, A. M. 2010. Comparative Genomic Evidence for a Complete Nuclear Pore Complex in the Last Eukaryotic Common Ancestor. *PLoS ONE,* 5**,** e13241. Available: DOI 10.1371/journal.pone.0013241.

Nevins, J. R. 1983. The Pathway of Eukaryotic mRNA Formation. *Annual Review of Biochemistry,* 52**,** 441-466. Available: DOI doi:10.1146/annurev.bi.52.070183.002301.

Ohno, S. 1970. *Evolution by gene duplication*, London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.

Okonechnikov, K., Golosova, O., Fursov, M. & Team, T. U. 2012. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics,* 28**,** 1166-1167. Available: DOI 10.1093/bioinformatics/bts091.

Olsen, G. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harbor Symposia on Quantitative Biology*, 1987. Cold Spring Harbor Laboratory Press, 825-837.

Ortlund, E. A., Bridgham, J. T., Redinbo, M. R. & Thornton, J. W. 2007. Crystal Structure of an Ancient Protein: Evolution by Conformational Epistasis. *Science,* 317**,** 1544-1548. Available: DOI 10.1126/science.1142819.

Oxford Dictionaries. 2010. *"Vitamin"* [Online]. Oxford University Press. Available: http://oxforddictionaries.com/definition/vitamin.

Pál, C., Papp, B. & Lercher, M. J. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature genetics,* 37**,** 1372-1375.

Pennisi, E. 2013. Ever-Bigger Viruses Shake Tree of Life. *Science,* 341**,** 226-227.

Penny, D., Collins, L. J., Daly, T. K. & Cox, S. J. 2014. The relative ages of Eukaryotes and Akaryotes. *Journal of Molecular Evolution,* 79**,** 228-239.

Petersen, T. N., Brunak, S., Von Heijne, G. & Nielsen, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods,* 8**,** 785-786.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G. & Baurain, D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS biology,* 9**,** e1000602.

Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L. & Bruley, C. 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science,* 341**,** 281-286.

Pisani, D., Cotton, J. A. & Mcinerney, J. O. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Molecular Biology and Evolution,* 24**,** 1752-1760.

Pond, S. L. K. & Frost, S. D. W. 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics,* 21**,** 2531-2533. Available: DOI 10.1093/bioinformatics/bti320.

Poole, A. M. & Neumann, N. 2011. Reconciling an archaeal origin of eukaryotes with engulfment: a biologically plausible update of the Eocyte hypothesis. *Research in Microbiology,* 162**,** 71-76. Available: DOI 10.1016/j.resmic.2010.10.002.

Poole, A. M. & Penny, D. 2001. Does endosymbiosis explain the origin of the nucleus? *Nature Cell Biology* 3**,** E173-174.

Poole, A. M. & Penny, D. 2007. Evaluating hypotheses for the origin of eukaryotes. *BioEssays,* 29**,** 74-84. Available: DOI 10.1002/bies.20516.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A. & Finn, R. D. 2011. The Pfam protein families database. *Nucleic Acids Research,* 40**,** D290-D301. Available: DOI 10.1093/nar/gkr1065.

Ramesh, M. A., Malik, S.-B. & Logsdon Jr, J. M. 2005. A Phylogenomic Inventory of Meiotic Genes: Evidence for Sex in Giardia and an Early Eukaryotic Origin of Meiosis. *Current Biology,* 15**,** 185-191. Available: DOI 10.1016/j.cub.2005.01.003.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M. & Claverie, J.-M. 2004. The 1.2-megabase genome sequence of Mimivirus. *Science,* 306**,** 1344-1350.

Reeds, P. J. 2000. Dispensable and Indispensable Amino Acids for Humans. *The Journal of Nutrition,* 130**,** 1835S-1840S.

Ribeiro, S. & Golding, G. B. 1998. The mosaic nature of the eukaryotic nucleus. *Molecular Biology and Evolution,* 15**,** 779-788.

Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. 1998. Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences, U.S.A.,* 95**,** 6239-6244.

Rogozin, I. B., Carmel, L., Csuros, M. & Koonin, E. V. 2012. Origin and evolution of spliceosomal introns. *Biol Direct,* 7.

Ronquist, F. & Huelsenbeck, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics,* 19**,** 1572-1574. Available: DOI 10.1093/bioinformatics/btg180.

Rose, W. C., Haines, W. J., Warner, D. T. & Johnson, J. E. 1951. The Amino acid requirements of Man: II. The role of Threonine and Histidine. *Journal of Biological Chemistry,* 188**,** 49-58.

Roy, S. W. & Gilbert, W. 2005. Complex early genes. *Proceedings of the National Academy of Sciences, U.S.A.,* 102**,** 1986-1991. Available: DOI 10.1073/pnas.0408355101.

Roy, S. W. & Irimia, M. 2009a. Mystery of intron gain: new data and new models. *Trends in Genetics,* 25**,** 67-73. Available: DOI 10.1016/j.tig.2008.11.004.

Roy, S. W. & Irimia, M. 2009b. Splicing in the eukaryotic ancestor: form, function and dysfunction. *Trends in Ecology & Evolution,* 24**,** 447-455. Available: DOI 10.1016/j.tree.2009.04.005.

Saw, J. H., Mountain, B. W., Feng, L., Omelchenko, M. V., Hou, S., Saito, J. A., Stott, M. B., Li, D., Zhao, G. & Wu, J. 2008. Encapsulated in silica: genome, proteome and physiology of the thermophilic bacterium Anoxybacillus flavithermus WK1. *Genome Biol,* 9**,** R161.

Smith, A. G., Croft, M. T., Moulin, M. & Webb, M. E. 2007. Plants need their vitamins too. *Current Opinion in Plant Biology,* 10**,** 266-275. Available: DOI DOI: 10.1016/j.pbi.2007.04.009.

Sober, E. 2004. The contest between parsimony and likelihood. *Systematic biology,* 53**,** 644-653.

Sober, E. & Steel, M. 2002. Testing the Hypothesis of Common Ancestry. *Journal of Theoretical Biology,* 218**,** 395-408. Available: DOI http://dx.doi.org/10.1006/jtbi.2002.3086.

Sonnhammer, E. L. L., Eddy, S. R. & Durbin, R. 1997. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins: Structure, Function, and Bioinformatics,*

28**,** 405-420. Available: DOI 10.1002/(sici)1097-0134(199707)28:3<405::aid-prot10>3.0.co;2-l.

Stanier, R. Y. & Niel, C. B. 1962. The concept of a bacterium. *Archives of Microbiology,* 42**,** 17-35.

Steel, M. 2002. Some statistical aspects of the maximum parsimony method. *Molecular systematics and evolution: theory and practice.* Springer.

Steel, M. & Penny, D. 2005. Maximum parsimony and the phylogenetic information in multistate characters. *Parsimony, phylogeny and genomics***,** 163-178.

Steel, M. & Penny, D. pers.comm. 2013. *RE: Exponential drop-off in the reliability of Markov models used in phylogenetics.* Type to Cox, S. J. L.

Sun, H., Merugu, S., Gu, X., Kang, Y. Y., Dickinson, D. P., Callaerts, P. & Li, W.-H. 2002. Identification of Essential Amino Acid Changes in Paired Domain Evolution Using a Novel Combination of Evolutionary Analysis and In Vitro and In Vivo Studies. *Molecular Biology and Evolution,* 19**,** 1490-1500.

Swofford, D. & Begle, D. P. 1993. *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1, March 1993*, Center for Biodiversity, Illinois Natural History Survey.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution,* 30**,** 2725-2729.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research,* 25**,** 4876-4882.

Thornton, J. W. 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat Rev Genet,* 5**,** 366-375.

Thornton, J. W., Need, E. & Crews, D. 2003. Resurrecting the Ancestral Steroid Receptor: Ancient Origin of Estrogen Signaling. *Science,* 301**,** 1714-1717. Available: DOI 10.1126/science.1086185.

Tourasse, N. J. & Gouy, M. 1999. Accounting for evolutionary rate variation among sequence sites consistently changes universal phylogenies deduced from rRNA and protein-coding genes. *Molecular phylogenetics and evolution,* 13**,** 159-168.

Van Der Giezen, M. & Tovar, J. 2005. Degenerate mitochondria. *EMBO Report,* 6**,** 525-530.

Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C. & Teichmann, S. A. 2004. Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology,* 14**,** 208-216.

Voordeckers, K., Brown, C. A., Vanneste, K., Van Der Zande, E., Voet, A., Maere, S. & Verstrepen, K. J. 2012. Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS biology,* 10**,** e1001446.

Wang, M., Kurland, C. G. & Caetano-Anollés, G. 2011. Reductive evolution of proteomes and protein structures. *Proceedings of the National Academy of Sciences,* 108**,** 11954-11958.

Whelan, S. & Goldman, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution,* 18**,** 691-699.

White, W. T. 2011. Ancestral sequence reconstruction - a PERL script

White, W. T. J., Zhong, B. & Penny, D. 2013. Beyond Reasonable Doubt: Evolution from DNA Sequences. *PLoS ONE,* 8**,** e69924.

Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature,* 504**,** 231-236.

Woese, C. R. & Fox, G. E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences, U.S.A.,* 74**,** 5088-5090.

Woese, C. R., Kandler, O. & Wheelis, M. L. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences, U.S.A.,* 87**,** 4576-4579.

Wolf, Y. I. & Koonin, E. V. 2013. Genome reduction as the dominant mode of evolution. *BioEssays,* 35**,** 829-837. Available: DOI 10.1002/bies.201300037.

Wolf, Y. I., Makarova, K. S., Yutin, N. & Koonin, E. V. 2012. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. *Biol Direct,* 14**,** 7-36.

Yang, Z. 2006. *Computational molecular evolution*, Oxford University Press Oxford.

Yang, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution,* 24**,** 1586-1591. Available: DOI 10.1093/molbev/msm088.

Yang, Z. 2012. PAML4 Users Guide : Phylogenetic Analysis by Maximum Likelihood. London. Available: http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf.

Yang, Z., Kumar, S. & Nei, M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics,* 141**,** 1641-1650.

Yasutake, Y., Yao, M., Sakai, N., Kirita, T. & Tanaka, I. 2004. Crystal structure of the Pyrococcus horikoshii isopropylmalate isomerase small subunit provides insight into the dual substrate specificity of the enzyme. *Journal of molecular biology,* 344**,** 325-333.

Yutin, N., Wolf, M., Wolf, Y. & Koonin, E. 2009. The origins of phagocytosis and eukaryogenesis. *Biology Direct,* 4**,** 9.

Zhang, J. 2000. Protein-length distributions for the three domains of life. *Trends in Genetics,* 16**,** 107-109.

Zhang, J. & Nei, M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *Journal of Molecular Evolution,* 44**,** S139-S146.

Zhang, Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC bioinformatics,* 9**,** 40.

Zmasek, C. M. & Godzik, A. 2011. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. *Genome Biol,* 12**,** R4.

Appendix 1

ASR Perl Script-

Available upon request- contact Simon Cox – sicox66@gmail.com

```
#!/usr/bin/perl

# asr.pl
# ======
# Created by: WTJW
# Created on: 3/06/2009

use strict;
use warnings;
use FindBin;
use lib "$FindBin::Bin/perllibs";              # Access our modules regardless of where we're installed
use PAML::AncestralStateReconstruction ':all';

# Automatically generated prototype list
sub writePamlAlignment($$$);
sub writePhylipAlignment($$$);
sub writeFastaAlignment($$$);

my $syntax = <<THE_END;
asr.pl: Conveniently extract ancestral sequences reconstructed using PAML

SYNTAX
  perl asr.pl [-c ctlfile -a analysistype] [options] [seqname [seqname...]]
  perl asr.pl [-c ctlfile -d dirname]      [options] [seqname [seqname...]]

OPTIONS
  -c ctlfile
    Names the PAML control file containing a description of the analysis
    to be (or which has been) performed.

  -a analysistype
    Can be either "codeml" or "baseml".  This names the PAML program that
    will be run.  Specifying this option causes a new PAML run to be
    performed; omitting it means that results from a previous PAML run should
    be analysed.  In the latter case, the -d option must be used to identify
    the directory containing the preexisting PAML output.

  -d dirname
    The directory that contains, or is to contain, the results from a PAML
    run.  (PAML produces numerous small output files.)  If the -a option
    is not specified, -d must be specified -- in this case dirname must name
    an existing directory.

    If the -a option is specified, the -d option can be omitted, in which
    case a unique subdirectory will be chosen based on the seqfile parameter
```

in the control file. If -d is specified, dirname must not already exist.

-r
Requests the root sequence of the tree. This option can be used with or without specifying additional sequences by name or number -- see the Notes section.

-R aligntype
Specify the type of reconstruction desired. This can be "marginal", "joint" or "either". (See the PAML documentation for details.) If "either" is specified and both reconstruction types are present in the PAML output, the marginal sequences will be preferred (marginal reconstruction works with more model types). The default is "either".

-f outputformat
Specify the output format. One of: "raw" (the default -- just unnamed sequences, one per line), "paml", "phylip", "fasta". "paml" and "phylip" are similar but different. A root sequence requested with -r will always be named "root", even if another node on the tree has been labelled with this name.

-q
Quiet mode: only output the sequences, plus any error messages.

-h
Output this help text.

--
Indicate the end of command-line options. Any subsequent arguments will be treated as literal sequence names, even if they look like options.

seqname
The name or number of the sequence whose data should be extracted. Multiple sequences can be specified.

NOTES
Sequences will be output in the order that their seqnames appear on the command line. The -r option can optionally be combined with other sequences requested by seqname -- in that case, the position of the -r switch determines where the root sequence will appear in the output.

Although trees provided to PAML using the treefile parameter cannot ordinarily contain node labels that do not correspond to an entry in the alignment, trees supplied to asr.pl can contain such node labels -- they will be stripped out before PAML sees the tree. This enables you to name particular ancestral nodes/subtrees, which you can then specify with the -s option.

EXAMPLES
perl asr.pl -c myseqs.ctl -a codeml -d results Baboon -r Rat Fish

Performs a PAML analysis, storing the output in directory "results", and

outputs four sequences in the following order: Baboon, the root sequence, Rat
and Fish.

  perl asr.pl -c myseqs.ctl -d results Monkey Ape Human Hominids

Accesses the previous run in the directory "results", and outputs four
sequences in the the following order: Monkey, Ape, Human, Hominids. Presumably
Hominids is the name of an ancestral node in the tree.
THE_END

```
# Process command-line arguments
my $ctlFName;
my $analType;
my $dirName;
my @seqNames;
my $quiet = 0;
my $reconsType = "either";
my $outputFormat = "raw";

while (@ARGV) {
	local $_ = shift;
	if (/\A-c\z/) {
		$ctlFName = shift;
	} elsif (/\A-a\z/) {
		$analType = shift;			# runPaml() will validate this
	} elsif (/\A-d\z/) {
		$dirName = shift;
	} elsif (/\A-r\z/) {
		push @seqNames, undef;		# Will be changed to the name assigned by PAML to
the root sequence after parseRstFile() is called
	} elsif (/\A-R\z/) {
		$reconsType = shift;	# Let extractAncestralSequence() find nothing if invalid
	} elsif (/\A-f\z/) {
		$outputFormat = shift;
	} elsif (/\A-q\z/) {
		$quiet = 1;
	} elsif (/\A-h\z/) {
		print $syntax;
		exit;
	} elsif (/\A--\z/) {
		last;
	} elsif (/\A-/) {
		die "Unrecognised command-line argument '$_'.\n\n$syntax";
	} else {
		push @seqNames, $_;
	}
}

# There may be arguments remaining if -- was seen.  In this case, "-r" will
# be interpreted literally, as the name of a sequence.
push @seqNames, @ARGV;
```

```
if (!defined($analType) && !defined($dirName)) {
	die "You must specify an existing results directory using -d when -a is not
specified.\n\n$syntax";
}

my %outputFormatFunc = (
	raw => \&writeRawAlignment,
	paml => \&writePamlAlignment,
	phylip => \&writePhylipAlignment,
	fasta => \&writeFastaAlignment
);

die "You must specify a control file with -c.\n\n$syntax" if !defined $ctlFName;
die "Output format specified with -f must be one of the following: " . join(", ", sort keys
%outputFormatFunc) . ".\n\n$syntax" if !exists $outputFormatFunc{$outputFormat}!Unexpected
End of Formula;

my $asr;
if (defined $analType) {
	print STDERR "Will run PAML analysis type '$analType'.\n" unless $quiet;
	$asr = runPaml $analType, $ctlFName, $dirName, $quiet;
} else {
	print STDERR "Will process existing PAML analysis run at path '$dirName'.\n" unless $quiet;
	$asr = { ctl => parseControlFile($ctlFName, $quiet), path => $dirName };
}

parseRstFile $asr, $quiet;

# It's only after parsing the rst file that we learn the name PAML has assigned
# to the root sequence.
foreach (@seqNames) {
	$_ = $asr->{tree}{seqNo} if !defined;
}

my %seqData = map { ($_ => extractAncestralSequence($asr, $_, $reconsType)) } @seqNames;

# Call the appropriate output function.
&{$outputFormatFunc{$outputFormat}}(\@seqNames, \%seqData, \*STDOUT);



# Doesn't output sequence names, just sequences.
sub writeRawAlignment($$$) {
	my ($seqNames, $seqData, $fh) = @_;
	for (my $i = 0; $i < @$seqNames; ++$i) {
		print $fh $seqData->{$seqNames->[$i]}, "\n";
	}
}

sub writePamlAlignment($$$) {
```

```
	my ($seqNames, $seqData, $fh) = @_;

	print $fh scalar(@$seqNames), " ", length($seqData->{$seqNames->[0]}), "\n";
	for (my $i = 0; $i < @$seqNames; ++$i) {
		print $fh $seqNames->[$i], "  ", $seqData->{$seqNames->[$i]}, "\n";
	}

	return 1;
}

sub writePhylipAlignment($$$) {
	my ($seqNames, $seqData, $fh) = @_;

	# Just trim names to 10 characters, pad them, and use the PAML routine.
	my @trimmedSeqNames = map { sprintf "%-10.10s", $_ } @$seqNames;
	my %trimmedSeqData = map { (sprintf("%-10.10s", $_) => $seqData->{$_}) } @$seqNames;

	if (keys %trimmedSeqData < @$seqNames) {
		die "Only " . scalar(keys %trimmedSeqData) . " distinct names remain after
contracting " . scalar(@$seqNames) . " names to 10 characters!";
	}

	return writePamlAlignment(\@trimmedSeqNames, \%trimmedSeqData, $fh);
}

sub writeFastaAlignment($$$) {
	my ($seqNames, $seqData, $fh) = @_;

	for (my $i = 0; $i < @$seqNames; ++$i) {
		print $fh ">", $seqNames->[$i], "\n";
		for (my $col = 0; $col < length $seqData->{$seqNames->[$i]}; $col += 60) {
			print $fh substr($seqData->{$seqNames->[$i]}, $col, 60), "\n";
		}
	}
}
```

Control File-

```
    seqfile = 10sp.paml
   treefile = 10sp.tre

    outfile = mlb      * main result file
      noisy = 9   * 0,1,2,3: how much rubbish on the screen
    verbose = 0   * 1: detailed output, 0: concise output
    runmode = 0   * 0: user tree;  1: semi-automatic;  2: automatic
              * 3: StepwiseAddition; (4,5):PerturbationNNI

      model = 4   * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
              * 5:T92, 6:TN93, 7:REV, 8:UNREST, 9:REVu; 10:UNRESTu

      Mgene = 0   * 0:rates, 1:separate; 2:diff pi, 3:diff kapa, 4:all diff
```

```
*       ndata = 5
        clock = 0   * 0:no clock, 1:clock; 2:local clock; 3:CombinedAnalysis
    fix_kappa = 0   * 0: estimate kappa; 1: fix kappa at value below
        kappa = 5  * initial or fixed kappa

    fix_alpha = 0   * 0: estimate alpha; 1: fix alpha at value below
        alpha = 0.5   * initial or fixed alpha, 0:infinity (constant rate)
       Malpha = 0   * 1: different alpha's for genes, 0: one alpha
        ncatG = 5   * # of categories in the dG, AdG, or nparK models of rates
        nparK = 0   * rate-class models. 1:rK, 2:rK&fK, 3:rK&MK(1/K), 4:rK&MK

        nhomo = 0   * 0 & 1: homogeneous, 2: kappa for branches, 3: N1, 4: N2
        getSE = 0   * 0: don't want them, 1: want S.E.s of estimates
 RateAncestor = 1   * (0,1,2): rates (alpha>0) or ancestral states

   Small_Diff = 7e-6
    cleandata = 1  * remove sites with ambiguity data (1:yes, 0:no)?
*       icode = 0  * (with RateAncestor=1. try "GC" in data,model=4,Mgene=4)
*  fix_blength = -1  * 0: ignore, -1: random, 1: initial, 2: fixed
       method = 0  * Optimization method 0: simultaneous; 1: one branch a time
```

Table 1

Key

act= Actinobacteria
alpha= Alphaproteobacteria
bac= Bacteria
beta= Betaproteobacteria
cyano= Cyanobacteria
delta= Deltaproteobacteria
detes= Bacteroidetes
firm= Firmicutes
gamma= Gammaproteobacteria
GNS= Green Non-sulphur bacteria
GS= Green-sulphur bacteria
hyp= Hyperthermophilic bacteria
plancto= Planctomycetes
proteo= Proteobacteria
spiro= Spirochaetes

NB-In the CyanoKEGG column, scores are out of 20 as this was a default setting. In the first row the 13/20 indicates that 13 of the homologues were from cyanobacteria, the other 7 were from homologues from the archaeplastida and were therefore of no more interest in this analysis.

Vit/AA	EC No	CyanoKEGG	Closest Phylogenetic Neighbours
Riboflavin (B2)	1.1.1.93	13/20	all cyano
Riboflavin (B2)	2.5.1.9	0/20	Mainly Bacteria, 2 Protists, 1 Fungi results scattered
Riboflavin (B2)	2.5.1.78	0/20	Mainly Firmicutes
Riboflavin (B2)	2.7.1.26	0/20	ProteoBac & Animals
Riboflavin (B2)	3.5.4.25 & 4.1.99.15	14/20	all cyano
Thiamine(B1)	ThiC	0/20	Mostly Actinobacteria
Thiamine(B1)	3.6.1.15	0/20	all animals
Thiamine(B1)	2.8.1.7	10/20	all cyano
Thiamine(B1)	2.7.6.2	0/20	all animals
Thiamine(B1)	2.7.4.7	0/20	5 families
Thiamine(B1)	2.7.1.50	0/20	mainly alphaproteobacteria
Thiamine(B1)	2.5.1.3	0/20	5 families
Niacin(B3)	2.7.1.23	0/20	Mainly insects
Niacin(B3)	2.7.7.1	0/20	Animals, Protists and fungi
Niacin(B3)	3.1.3.5	0/20	8 different bacterial classes
Niacin(B3)	3.6.1.22	0/20	Mainly mammals
Niacin(B3)	6.3.5.1.	0/20	Predominantly Fungi
Niacin(B3)	2.4.2.11	0/20	Mainly insects
Niacin(B3)	2.4.2.19	0/20	All Bacteroidetes
Niacin(B3)	NadA	4/20	Nests with Rice
Niacin(B3)	1.4.3.16	0/20	spirocheates, bacteroidetes, 2 GNS
Pyridoxine(B6)	2.6.1.52	0/20	Mostly gamma and beta proteo bac, 1 GS & 1 GNS
Pyridoxine(B6)	2.7.1.35	0/20	All Animals
Pyridoxine(B6)	4.2.3.1	1/20	Mainly Green non sulphur & proteo
Pyridoxine(B6)	YaaD	0/20	bacteria, protists and fungi
Pyridoxine(B6)	YaaE	0/20	bacteria, protists and fungi
Folates (B9)	MOCS2	0/20	all animals
Folates (B9)	MOCS1	0/20	all animals
Folates (B9)	3.5.4.16	1/20	animals, protists & Bacteria
Folates (B9)	3.1.3.1	0/20	animals, protists & Bacteria
Folates (B9)	4.1.2.25	0/20	Mainly Firmicutes
Folates (B9)	1.5.1.3	0/20	all protists
Folates (B9)	2.5.1.15	0/20	mainly Fungi
Folates (B9)	2.6.1.85	6/20	1/2 cyano
Folates (B9)	2.7.6.3	0/20	mainly Fungi
Folates (B9)	3.4.19.9	0/20	2 Protists, rest animals
Folates (B9)	6.3.2.17	0/20	all animals
Ascorbate(C)	1.1.1.22	0/20	animals, 1 protist, 2 bac
Ascorbate(C)	5.1.3.18	0/20	Protists and assorted bacteria, 1 GNS
Ascorbate(C)	2.7.7.69	0/20	all animals

Ascorbate(C)	VTC4	0/20	all animals
Ascorbate(C)	1.1.1.122	0/20	2 animals, assorted bac
Ascorbate(C)	1.13.99.1	1/20	6 fungi, 1 protist, 5 bacteroidetes, 1 cyano
Ascorbate(C)	1.3.2.3	0/20	5 animals, 4 protists
Ascorbate(C)	1.6.5.4	0/20	1 protist, assorted bac
Ascorbate(C)	1.10.3.3	0/20	all fungi
Ascorbate(C)	1.11.1.11	0/20	10 fungi 2 protists
Biotin(B7)	2.3.1.47	0/20	assorted bacteria
Biotin(B7)	2.6.1.62	0/20	mainly fungi, 1 protist
Biotin(B7)	2.8.1.6	0/20	mainly fungi, 3 bac
Pantothenic acid(B5)	2.2.1.6	8/20	all cyano
Pantothenic acid(B5)	1.1.1.86	0/20	3 protists, 4 assorted bac
Pantothenic acid(B5)	4.2.1.9	0/20	1 Fungi, 1 protist, assorted bac
Pantothenic acid(B5)	2.1.2.11	0/20	Even spread-Fungi, protists & bac
Pantothenic acid(B5)	2.6.1.42	0/20	All firmicutes
Pantothenic acid(B5)	6.3.2.1	0/20	all Bac, Mainly hyperthermophillic
Co-enzyme A	2.7.1.33	0/20	all animals
Co-enzyme A	6.3.2.5	0/20	animals, 1 fungi
Co-enzyme A	4.1.1.36	0/20	all animals
Co-enzyme A	1.3.1.2	0/20	9 proteo, i bacteroidetes
Co-enzyme A	3.5.2.2	0/20	Animals protists & assorted Bac
Co-enzyme A	3.5.1.6	0/20	1 Protist, rest animals
Co-enzyme A	2.7.7.3	0/20	all animals
Co-enzyme A	2.7.8.-	0/20	all assorted bac
Co-enzyme A	2.7.1.24	0/20	1 fungi, rest firmicutes
Histidine	2.4.2.17	1/20	Green non sulphur and 1 cyano
Histidine	3.6.1.31	0/20	all assorted bac
Histidine	3.5.4.19	0/20	all assorted bac
Histidine	5.3.1.16	0/20	3 protists, rest assorted bac
Histidine	HisF	0/20	bacteria, 3 protists, 1 fungi
Histidine	HisH	0/20	bacteria, 3 protists, 1 fungi
Histidine	4.2.1.19	11/20	all cyano
Histidine	2.6.1.9	0/20	Mainly Green-nonsulfur bac
Histidine	1.1.1.23	0/20	all fungi
Histidine	2.1.1.-	0/20	mainly GS
Histidine	4.1.1.2	3/20	3 cyano, 3 protists, rest firm gamma detes
Histidine	1.2.1.3	4/20	4 cyano, 1 gamma, rest animals
Histidine	4.1.1.28	0/20	all animals
Isoleucine	4.2.1.35	1/20	1 cyano, 4 GS, rest asorted bac
Isoleucine	LeuC	1/20	2 cyano, 4 GS, rest asorted bac
Isoleucine	LeuD	1/20	3 cyano, 4 GS, rest asorted bac
Isoleucine	LeuB	12/20	all cyano
Isoleucine	4.3.1.19	5/20	5 cyano , 9 beta
Isoleucine	1.2.4.1	10/20	all cyano
Isoleucine	2.2.1.6	0/20	1 animal, rest plancto

Isoleucine	1.1.1.86	0/20	red algae in middle of bac, 3 protists
Isoleucine	4.2.1.9	0/20	1 protists assorted bac
Isoleucine	2.6.1.42	0/20	assorted bacteria, mainly firm
Leucine	2.3.3.13	2/20	2 cyano, 2 protist, 4 GNS, delta
Leucine	4.2.1.33	1/20	1 cyano, Gs detes & assorted
Leucine	1.1.1.85	12/20	all cyano
Lysine	2.7.2.4	0/20	assorted bac- gamma beta spiro, detes
Lysine	1.2.1.11	5/20	1 protist, 2 alpha, 1 hyper, 5 cyano
Lysine	1.1.1.3	0/20	assorted bac
Lysine	4.2.1.52	0/20	assorted bac
Lysine	1.3.1.26	0/20	assorted bac, 2 protists
Lysine	2.6.1.83	0/20	assorted bac, 2 protists
Lysine	5.1.1.7	11/20	all cyano
Lysine	4.1.1.20	0/20	Mainly hyperthermophilic bacteria
Lysine	1.2.1.31	0/20	2 Protists, rest animals
Lysine	LysZ	10/20	all cyano
Lysine	LysY	0/20	all bac, odd groups
Methionine	2.3.1.20	0/20	gamma alpha
Methionine	4.4.1.8	0/20	mainly Fungi, 1 protist
Methionine	2.5.1.47	0/20	mainly Fungi, 1 protist
Methionine	2.6.1.1	0/20	all animals
Methionine	2.5.1.48	0/20	delta planct GNS
Methionine	2.8.1.2	0/20	all alpha
Methionine	2.7.2.4	0/20	detes, gamma, delta spiro
Methionine	1.2.1.11	5/20	2 alpha, 1 protist, 5 cyano, 1 hyper
Methionine	1.1.1.3	0/20	detes, gamma, delta spiro
Methionine	1.1.1.27	0/20	all animals
Methionine	3.3.1.1	0/20	6 protists, actino
Methionine	2.1.1.10	0/20	Mixed Bac GNS
Methionine	2.1.1.14	1/20	1 Cyano, assorted Bac- alpha, firm, hyp act
Methionine	2.1.1.37	0/20	1 protist rest animals
Methionine	2.5.1.6	0/20	5 protists rest animals
Methionine	4.1.1.50	0/20	3 protists rest animals
Methionine	2.5.1.16	0/20	animals fungi protists
Methionine	4.4.1.14	0/20	All Animals
Methionine	3.2.2.16	0/20	3 protists- firm gamma spiro
Methionine	4.4.1.11	0/20	All Bacteroidetes
Methionine	11.4.17.4	0/20	protists bac & 1 fungi
Methionine	1.13.11.53	0/20	All Animals
Methionine	3.1.3.77	0/20	Mainly protists & animals, 1 bac
Methionine	5.3.1.23	0/20	all euks
Methionine	1.13.11.54	0/20	all animals
Phenyalanine	2.5.1.4	0/20	Proteobacteria however 1 Diatom
Phenyalanine	4.2.3.4	0/20	gamma, 2 protists
Phenyalanine	4.2.1.10	0/20	all gamma

Phenyalanine	1.1.1.25	0/20	all gamma
Phenyalanine	2.7.1.71	11/20	all cyano
Phenyalanine	2.5.1.19	0/20	gamma, 2 protists
Phenyalanine	4.2.3.5	10/20	all cyano
Phenyalanine	5.4.99.5	0/20	fungi, 1 protist
Phenyalanine	2.6.1.1	0/20	all animals
Phenyalanine	2.6.1.5	0/20	all animals
Phenyalanine	2.6.1.9	0/20	mainly GS
Threonine	4.1.2.5	0/20	firm, delta, GNS, 1 animal
Threonine	1.4.3.21	3/20	3 cyano, actino, GNS,
Threonine	4.3.1.19	5/20	5 Cyano, rest Betaproteobac
Threonine	4.2.3.1	1/20	1 cyano, GNS, delta, plancto
Threonine	2.7.1.39	0/20	GNS oidetes
Threonine	1.1.1.3	0/20	detes, gamma,beta, spiro
Threonine	2.7.2.4	0/20	detes, gamma,beta, spiro
Threonine	1.2.1.11	5/20	2 alpha, 1 protist, 5 cyano, 1 hyper
Tryptophan	4.2.1.20	10/20	all cyano
Tryptophan	4.1.3.27	7/20	all cyano
Tryptophan	2.4.2.18	0/20	1 archaea, 1 diatom, chlamy
Tryptophan	5.3.1.24	0/20	1 fungi, chlamy
Tryptophan	4.1.1.48	9/20	all cyano
Valine	4.2.1.35	1/20	1 cyano, 4 GS, rest asorted bac
same as Isoleucine	LeuC	1/20	2 cyano, 4 GS, rest asorted bac
Valine	LeuD	1/20	3 cyano, 4 GS, rest asorted bac
same as Isoleucine	LeuB	12/20	all cyano
Valine	4.3.1.19	5/20	5 cyano , 9 beta
same as Isoleucine	1.2.4.1	10/20	all cyano
Valine	2.2.1.6	0/20	1 animal, rest plancto
same as Isoleucine	1.1.1.86	0/20	red algae in middle of bac, 3 protists
Valine	4.2.1.9	0/20	1 protists assorted bac
same as Isoleucine	2.6.1.42	0/20	assorted bacteria,mainly firm

Table 2

Column 1 contains all the ancestral sequences calculated for the biochemical pathway and the E.C. number, for example, “Asc1.1.1.22Anc_translation” means that this is the ancestral sequence for the Ascorbate pathway for the enzyme that catalyses the reaction 1.1.1.22(see key below Table). Column 2 gives the family name for the protein family that the AS has a homology to; families are groups of proteins that share a similar function. PFAM links to reference sources, structural sites and many other resources about each family group; family and clan names are searchable at the PFAM website- http://pfam.sanger.ac.uk/ . Column 3 gives the E-value of this homology and column 5 whether this is significant or not. Column 6 shows the PFAM clan that the AS belongs to; clans are sets of related Pfam-A families.

Enzyme	Family Name	E Value	Significance	Clan
Asc1.1.1.22Anc_translation	UDPG_MGDP_dh_N	4.40E-20	1	CL0063
Asc1.1.1.22Anc_translation	UDPG_MGDP_dh_N	3.80E-09	1	CL0063
Asc1.1.1.22Anc_translation	UDPG_MGDP_dh_C	1.90E-08	1	No_clan
Asc5.1.3.18Anc_translation	Epimerase	4.20E-46	1	CL0063
AscVTC4Anc_translation	Inositol_P	3.60E-16	1	CL0171
Asc1.13.99.1Anc_translation	DUF706	2.60E-05	1	CL0237
Asc1.3.2.3Anc_translation	FAD_binding_4	2.30E-05	1	CL0077
Asc1.6.5.4Anc_translation	Pyr_redox	0.00065	0	CL0063
Asc1.10.3.3Anc_translation	Cu-oxidase_3	6.50E-05	1	CL0026
B62.6.1.52Anc_translation	Aminotran_5	0.027	0	CL0061
B6YaaEAnc_translation	SNO	5.90E-19	1	CL0014
Bio2.6.1.62_translation	Aminotran_3	3.30E-07	1	CL0061
Bio2.8.1.6Anc_translation	Radical_SAM	1.10E-08	1	No_clan
Bio2.8.1.6Anc_translation	BATS	3.90E-05	1	No_clan
CoA2.7.1.24Anc_translation	CoaE	6.00E-06	1	CL0023
CoA2.7.1.33Anc_translation	Fumble	0.0011	0	CL0108
CoA3.5.1.6Anc_translation	CN_hydrolase	3.30E-25	1	No_clan
CoA3.5.2.2.Anc_translation	Amidohydro_1	1.60E-06	1	CL0034
CoA4.1.1.36Anc_translation	Flavoprotein	3.70E-24	1	No_clan
CoA6.3.2.5Anc_translation	DFP	3.30E-18	1	CL0063
FolancMOCS1_translation	MoaC	1.10E-05	1	No_clan
Folanc4.1.2.25_translation	FolB	0.016	0	CL0334
Folanc2.5.1.15_translation	HPPK	0.29	0	No_clan
Folanc2.5.1.15_translation	Pterin_bind	2.10E-11	1	CL0036
Folanc2.6.1.85_translation	GATase	7.10E-09	1	CL0014
Folanc2.6.1.85_translation	GATase	5.60E-05	1	CL0014
Folanc2.6.1.85_translation	Anth_synt_I_N	0.0072	0	No_clan
Folanc2.6.1.85_translation	Chorismate_bind	4.90E-69	1	No_clan
Folanc2.7.6.3_translation	HPPK	0.29	0	No_clan
Folanc2.7.6.3_translation	Pterin_bind	2.10E-11	1	CL0036
Hisanc2.4.2.17_translation	HisG	7.40E-13	1	CL0177
Hisanc2.4.2.17_translation	HisG_C	8.00E-10	1	CL0089
HisancHisH&F_translation	His_biosynth	3.20E-24	1	CL0036
Hisanc4.2.1.19_translation	IGPD	1.20E-24	1	CL0329
Hisanc2.6.1.9_translation	Aminotran_1_2	4.80E-06	1	CL0061
Hisanc1.1.1.23_translation	Histidinol_dh	2.50E-33	1	CL0099
Isoanc4.2.1.35_translation	Aconitase	2.50E-18	1	No_clan
Isoanc4.2.1.35_translation	Aconitase	2.70E-11	1	No_clan
Isoanc4.2.1.35_translation	Aconitase	2.70E-20	1	No_clan
IsoancLeuB_translation	Iso_dh	3.90E-09	1	CL0270
Isoanc4.3.1.19_translation	PALP	9.70E-11	1	No_clan
Isoanc4.3.1.19_translation	WCCH	0.038	0	No_clan
Isoanc4.2.1.9_translation	ILVD_EDD	1.00E-163	1	No_clan

Leuanc2.3.3.13_translation	HMGL-like	0.012	0	CL0036
Leuanc2.3.3.13_translation	LeuA_dimer	0.024	0	No_clan
Leuanc4.2.1.35_translation	Aconitase	2.50E-18	1	No_clan
Leuanc4.2.1.35_translation	Aconitase	2.70E-11	1	No_clan
Leuanc4.2.1.35_translation	Aconitase	2.70E-20	1	No_clan
Lys1.1.1.3Anc_translation	ACT	1.70E-06	1	CL0070
Lys1.1.1.3Anc_translation	NAD_binding_3	2.80E-09	1	CL0063
Lys1.1.1.3Anc_translation	Homoserine_dh	3.80E-09	1	No_clan
Lys1.1.1.3Anc_translation	Homoserine_dh	0.00051	0	No_clan
Lys1.2.1.1Anc_translation	Semialdhyde_dh	2.70E-09	1	CL0063
Lys1.2.1.31Anc_translation	Aldedh	1.30E-06	1	CL0099
Lys2.7.2.4anc_translation	ACT	1.70E-06	1	CL0070
Lys2.7.2.4anc_translation	NAD_binding_3	2.80E-09	1	CL0063
Lys2.7.2.4anc_translation	Homoserine_dh	3.80E-09	1	No_clan
Lys2.7.2.4anc_translation	Homoserine_dh	0.00051	0	No_clan
Lys4.1.1.20Anc_translation	Orn_DAP_Arg_deC	8.10E-06	1	No_clan
Lys4.2.1.52Anc_translation	DHDPS	0.0064	0	CL0036
Lys5.1.1.7Anc_translation	DAP_epimerase	0.68	0	CL0288
Meth2.1.1.10Anc_translation	S-methyl_trans	1.30E-07	1	No_clan
Meth2.1.1.14Anc_translation	Meth_synt_1	1.30E-13	1	CL0160
Meth2.1.1.14Anc_translation	Meth_synt_2	3.80E-30	1	CL0160
Meth2.5.1.6Anc_translation	S-AdoMet_synt_N	3.70E-14	1	No_clan
Meth2.5.1.6Anc_translation	S-AdoMet_synt_M	3.40E-50	1	No_clan
Meth2.5.1.6Anc_translation	S-AdoMet_synt_C	1.20E-50	1	No_clan
Meth3.1.3.77Anc_translation	Aldolase_II	2.20E-29	1	No_clan
Meth3.1.3.77Anc_translation	Pfam-B_575	6.10E-21	NA	NA
Meth4.4.1.11Anc_translation	Cys_Met_Meta_PP	2.20E-16	1	CL0061
Meth4.4.1.11Anc_translation	Cys_Met_Meta_PP	5.40E-05	1	CL0061
Meth5.3.1.23Anc_translation	IF-2B	3.90E-45	1	CL0246
Niaanc3.1.3.5_translation	SurE	0.0088	0	No_clan
Pan1.1.1.86Anc_translation	IlvC	0.015	0	CL0106
Pan4.2.1.9Anc_translation	ILVD_EDD	2.30E-22	1	No_clan
Pan2.1.2.11Anc_translation	Pantoate_transf	2.30E-23	1	CL0151
Pan6.3.2.1Anc_translation	Pantoate_ligase	1.90E-07	1	CL0119
Phenanc2.6.1.9_translation	Aminotran_1_2	4.80E-06	1	CL0061
Tyranc2.6.1.9_translation	Aminotran_1_2	4.80E-06	1	CL0061
Tyranc1.3.1.78_translation	F420_oxidored	0.0062	0	CL0063
Tyranc1.3.1.78_translation	Sporozoite_P67	0.73	0	No_clan
Phen&Tyranc1.1.1.25_translation	DHquinase_I	4.90E-12	1	CL0036
Phen&Tyranc1.1.1.25_translation	Shikimate_dh_N	0.0031	0	No_clan
P&Tanc2.5.1.19_translation	EPSP_synthase	0.012	0	CL0290
P&Tanc2.5.1.19_translation	EPSP_synthase	2.30E-33	1	CL0290
P&Tanc2.7.1.71_translation	SKI	1.40E-08	1	CL0023
P&Tanc4.2.1.10_translation	Shikimate_dh_N	0.00014	1	No_clan
P&Tanc4.2.3.4_translation	DHQ_synthase	4.30E-11	1	CL0224

P&Tanc5.4.99.5_translation	Pfam-B_2031	2.20E-25	NA	NA
Rib2.7.1.26Anc_translation	Flavokinase	2.70E-12	1	No_clan
Rib3.5.4.25_translation	DHBP_synthase	8.80E-16	1	No_clan
Thi3.6.1.15Anc_translation	NTPase_1	9.00E-06	1	CL0023
Thi3.6.1.15Anc_translation	Pfam-B_12239	0.00087	NA	NA
Thi2.8.1.7Anc_translation	Aminotran_5	0.00071	0	CL0061
Thi2.8.1.7Anc_translation	Aminotran_5	0.36	0	CL0061
Thi2.7.6.2Anc_translation	TPK_catalytic	0.0071	0	No_clan
Thi2.7.1.50Anc_translation	HK	3.30E-06	1	CL0118
Thi2.7.1.50Anc_translation	HK	5.00E-05	1	CL0118
Thi2.5.1.3Anc_translation	TMP-TENI	2.20E-05	1	CL0036
Thi2.5.1.3Anc_translation	TMP-TENI	0.52	0	CL0036
Thr1.1.1.3Anc_translation	ACT	1.70E-06	1	CL0070
Thr1.1.1.3Anc_translation	NAD_binding_3	2.80E-09	1	CL0063
Thr1.1.1.3Anc_translation	Homoserine_dh	3.80E-09	1	No_clan
Thr1.1.1.3Anc_translation	Homoserine_dh	0.00051	0	No_clan
Thr1.2.1.1Anc_translation	Semialdhyde_dh	2.70E-09	1	CL0063
Thr1.4.3.21Anc_translation	Cu_amine_oxid	6.40E-10	1	No_clan
Thr1.4.3.21Anc_translation	Cu_amine_oxid	8.70E-12	1	No_clan
Thr2.7.1.39Anc_translation	DUF3248	0.28	0	No_clan
Thr2.7.1.39Anc_translation	GHMP_kinases_C	0.0079	0	No_clan
Thr2.7.2.4anc_translation	ACT	1.70E-06	1	CL0070
Thr2.7.2.4anc_translation	NAD_binding_3	2.80E-09	1	CL0063
Thr2.7.2.4anc_translation	Homoserine_dh	3.80E-09	1	No_clan
Thr2.7.2.4anc_translation	Homoserine_dh	0.00051	0	No_clan
Thr4.1.2.5Anc_translation	Beta_elim_lyase	4.90E-17	1	CL0061
Tryanc2.4.2.18_translation	Glycos_transf_3	3.30E-24	1	No_clan
Tryanc2.4.2.18_translation	Glycos_transf_3	0.02	0	No_clan
Try4.1.1.48_translation	RE_Eco29kI	0.058	0	No_clan
Tryanc4.2.1.20_translation	Trp_syntA	0.026	0	CL0036
Tryanc4.2.1.20_translation	Trp_syntA	4.40E-12	1	CL0036

Key to abbreviations above

Symbol	Meaning
Asc	Ascorbate (C)
B6	Pyridoxine (B6)
Bio	Biotin (B7)
CoA	Co-enzyme A
Fol	Folate (B9)
His	Histidine
Iso	Isoleucine
Leu	Leucine
Lys	Lysine
Meth	Methionine

Nia	Niacin (B3)
Pan	Pantothenic Acid (B5)
Phen	Phenyalanine
Try	Tryptophan
P&T	Phenyalanine & Tryptophan- shared pathway
Rib	Riboflavin (B2)
Thi	Thiamone (B1)
Thr	Threonine

Table 3

Organism	Lysine						Ether Lipids			
	1.2.1.11	1.1.1.3	4.3.3.7	1.17.1.8	2.6.1.83	2.7.8.1	3.1.1.4	3.1.4.4	3.1.4.3	LPCAT
ath	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
aly	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
crb	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
eus	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
brp	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
cit	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
cic	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
tcc	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
egr	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
gmx	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
pvu	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
mtr	Seq	Seq	Seq	>0.001	no hit	Seq	Seq	Seq	Seq	> 0.001
cam	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
fve	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
pper	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
pmum	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
mdm	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
pxb	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
csv	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
cmo	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
rcu	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
jcu	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
pop	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
vvi	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
sly	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
sot	Seq	Seq	Seq	>0.001	Seq	Seq	Seq	Seq	Seq	Seq
bvg	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
osa	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
dosa	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
obr	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
bdi	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq

sbi	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
zma	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
sita	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
pda	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
egu	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
mus	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
atr	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
smo	Seq	Seq	Seq	Seq	no hit	Seq	Seq	Seq	Seq	> 0.001
ppp	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq
cre	Seq	Seq	Seq	no hit	Seq	no hit	Seq	no hit	no hit	> 0.001
vcn	Seq	Seq	Seq	no hit	Seq	no hit	Seq	no hit	no hit	> 0.001
olu	Seq	Seq	Seq	no hit	no hit	no hit	Seq	no hit	no hit	no hit
ota	Seq	Seq	Seq	Seq	<0.001	no hit	Seq	no hit	no hit	no hit
bpg	Seq	Seq	Seq	no hit	Seq	no hit	Seq	no hit	no hit	no hit
mis	Seq	no hit	no hit	no hit	Seq	no hit	Seq	no hit	no hit	no hit
mpp	Seq	no hit	Seq	> 0.001	Seq	no hit	Seq	no hit	>0.001	no hit
csl	Seq	Seq	Seq	Seq	Seq	> 0.001	Seq	> 0.001	no hit	no hit
cvr	Seq	Seq	Seq	Seq	no hit	no hit	Seq	no hit	no hit	no hit
cme	Seq	Seq	no hit	no hit	Seq	no hit	Seq	no hit	no hit	no hit
gsl	Seq	>0.001	>0.001	> 0.001	Seq	Seq	Seq	no hit	no hit	no hit
ccp	Seq	>0.001	no hit	Seq	Seq	> 0.001	Seq	> 0.001	no hit	> 0.001

Table 4

Organism	Type	Coverage	Assembly
Arabidopsis thaliana (thale cress)	PacBio	140	ASM83594v1
Arabidopsis lyrata (lyrate rockcress)	Illumina HiSeq2000	279.3	Alyr_1.0
Capsella rubella	ABI 3739; Roche 454FLX	22.35	Caprub1_0
Eutrema salsugineum	Illumina GAIIx	134	TsV2-8
Brassica rapa (field mustard)	Illumina GA	72	Brapa_1.0
Citrus sinensis (Valencia orange)	454 GS-FLX Titanium; 454 FLX Standard; ABI 3739	35.1	Citrus_sinensis_v1.0
Citrus clementina (mandarin orange)	ABI 3739	6.9	Citrus_clementina_v1.0
Theobroma cacao (cacao)	454; Illumina; Sanger	15.58	Theobroma_cacao_20110822
Eucalyptus grandis (rose gum)	ABI 3739	6.73	Egrandis1_0
Glycine max (soybean)	Sanger	8	V1.1
Phaseolus vulgaris (common bean)	ABI 3730; Roche 454 FLX; Illumina GAII	21.2	PhaVolg1_0
Medicago truncatula (barrel medic)	Sanger; 454; Illumina	90	MedtrA17_4.0
Cicer arietinum (chickpea)	Illumina Hiseq 2000	120	ASM33114v1
Fragaria vesca (woodland strawberry)	454; Solexa	49	FraVesHawaii_1.0
Prunus persica (peach)	ABI 3739	8.47	Prupe1_0
Prunus mume (Japanese apricot)	Illumina Hiseq2000	180	P.mume_V1.0
Malus domestica (apple)	Sanger; 454		MalDomGD1.0

Pyrus x bretschneideri (Chinese white pear)	Illumina HS2000	76	Pbr_v1.0
Cucumis sativus (cucumber)	Illumina GAII	86	ASM407v2
Cucumis melo (muskmelon)	Sanger; 454	13.52	ASM31304v1
Ricinus communis (castor bean)	AB3730xl	4.5	JCVI_RCG_1.1
Jatropha curcas	Illumina GAII; Illumina HS	189	JatCur_1.0
Populus trichocarpa (black cottonwood)	ABI 3739	7.45	Poptr2_0
Vitis vinifera (wine grape)	Sanger	12	12X
Solanum lycopersicum (tomato)	454; Sanger; Illumina; SOLiD	27	SL2.50
Solanum tuberosum (potato)	Illumina GA2	114	SolTub_3.0
Beta vulgaris (sugar beet)	454; Illumina GAIIx; Illumina HiSeq; Sanger	30	RefBeet-1.2.1
Oryza sativa japonica (Japanese rice) (RefSeq)	Illumina HiSeq	220	HEG4v1.0
Oryza brachyantha (malo sina)	Illumina GA II	104	Oryza_brachyantha.v1.4b
Brachypodium distachyon	Sanger	9.43	v1.0
Sorghum bicolor (sorghum)	Illumina Solexa	12	ASM23674v2
Zea mays (maize)	454 GS20; 454 Titanium; Sanger	4.2	B73 RefGen_v3
Setaria italica (foxtail millet)	Sanger	8.29	Setaria V1
Phoenix dactylifera (date palm)	454; SOLiD; ABI3730	139	DPV01
Elaeis guineensis (African oil palm)	454 LS	16	EG5
Musa acuminata (wild Malaysian banana)	454; Sanger; Illumina	20.5	ASM31385v1
Amborella trichopoda	454 LS	30	AMTR1.0
Selaginella moellendorffii	Sanger	7	v1.0
Physcomitrella patens subsp. patens	Sanger	8.1	v1.1
Chlamydomonas reinhardtii	Sanger	12	v3.0
Volvox carteri f. nagariensis	Sanger	8	v1.0
Ostreococcus lucimarinus	WGS		ASM9206v1
Ostreococcus tauri	WGS		version 050606
Bathycoccus prasinos	no record of sequencing		
Micromonas sp. RCC299	WGS		ASM9098v2
Micromonas pusilla	WGS		Micromonas pusilla CCMP1545 v2.0
Coccomyxa subellipsoidea	ABI 3730	8	Coccomyxa subellipsoidae v2.0
Chlorella variabilis	Sanger	8.9	v 1.0
Cyanidioschyzon merolae	WGS		ASM9120v1
Galdieria sulphuraria	Sanger; 454	10	ASM34128v1
Chondrus crispus (carragheen)	ABI3730xl	14	ASM35022v2