Abstract

Amino acids at conserved sites in the residue sequence of 10 ancient proteins, from 844 phylogenetically diverse sources, were used to specify their time of origin in the interval before species divergence from the last common ancestor (LCA). The order of amino acid addition to the genetic code, based on biosynthesis path length and other molecular evidence, provided a reference for evaluating the ‘code age’ of each residue profile examined. Significantly earlier estimates were obtained for conserved amino acid residues in these proteins than non-conserved residues. Evidence from the primary structure of ‘fossil’ proteins thus corroborated the biosynthetic order of amino acid addition to the code.

Low potential ferredoxin (Fdxn) had the earliest residue profile among the proteins in this study. A phylogenetic tree for 82 prokaryote Fdxn sequences was rooted midway between bacteria and archaea branches. LCA Fdxn had a 23-residue antecedent whose residue profile matched mid-expansion phase codon assignments and included an amide residue. It contained a highly acidic N-terminal region and a non-charged C-terminal region, with all four cysteine residues. This small protein apparently anchored a [4Fe-4S] cluster, ligated by C-terminal cysteines, to a positively charged mineral surface, consistent with mediating e\(^{-}\) transfer in a primordial surface system before cells appeared. Its negatively charged N-terminal ‘attachment site’ was highly mutable during evolution of ancestral Fdxn for Bacteria and Archaea, consistent with a loss of function after cell formation. An initial glutamate to lysine substitution may link ‘attachment site’ removal to early post-expansion phase entry of basic amino acids to the code. As proteins evidently anchored non-charged amide residues initially, surface attachment of cofactors and other functional groups emerges as a general function of pre-cell proteins.
A phylogenetic tree of 107 proteolipid (PL) helix-1 sequences from H⁺-ATPase of bacteria, archaea and eukaryotes had its root between prokaryote branches. LCA PL h1 residue profile optimally fit a late expansion phase codon array. Sequence repeats in transmembrane PL helices h1 and h2 indicated formation of the archetypal PL hairpin structure involved successive tandem duplications, initiated within the gene for an 11-residue (or 4-residue) hydrophobic peptide. Ancestral PL h1 lacked acidic residues, in a fundamental departure from the prototype pre-cell protein. By this stage, proteins with a hydrophobic domain had evolved. Its non-polar, late expansion phase residue profile point to ancestral PL being a component of an early permeable cell membrane. Other indicators of cell formation about this stage of code evolution include phospholipid biosynthesis path length, FtsZ residue profile, and late entry of basic amino acids into the genetic code.

Estimates based on conserved residues in prokaryote cell septation protein, FtsZ, and proteins involved with synthesis, transcription and replication of DNA revealed FtsZ, ribonucleotide reductase, RNA polymerase core subunits and 5'→3' flap exonuclease, FEN-1, originated soon after cells putatively evolved. While reverse transcriptase and topoisomerase I, Topo I, appeared late in the pre-divergence era, when the genetic code was essentially complete. The transition from RNA genes to a DNA genome seemingly proceeded via formation of a DNA–RNA heteroduplex. These results suggest formation of DNA awaited evolution of a catalyst with a hydrophobic domain, capable of sequestering radical bearing intermediates in its synthesis from ribonucleotide precursors. Late formation of topology altering protein, Topo I, further suggests consolidation of genes into chromosomes followed synthesis of comparatively thermostable DNA strands.

© 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Pre-divergence proteins; Residue profile; Code evolution; Cofactor anchor; Hydrophobic domain; Cell formation; Transitional genome

Contents

1. Introduction .......................................................... 79
2. Low-potential ferredoxin as a prototype of pre-cell proteins ........................................ 81
3. H⁺-ATPase proteolipid is a marker for cell membrane formation ........................................ 93
4. Origin of cell division and the DNA genome ..................................................................... 106
5. Discussion .............................................................. 117
   5.1. Early ‘code age’ of residue profile at conserved sites .................................................. 117
   5.2. Earliest protein sequence contained an amide residue ............................................. 118
   5.3. [4Fe-4S] Pro-ferredoxin-[5] has a surface attachment segment .................................. 119
   5.4. Mutability of acidic attachment segment and appearance of cells ............................ 119
   5.5. 5'-Initiation of translation before code expansion .................................................... 120
   5.6. Sequence extension by tandem duplication ............................................................. 121
   5.7. Protein hydrophobic domain and the origin of cells ................................................. 122
   5.8. Retarded origin of DNA and the RNA–DNA heteroduplex ....................................... 123
   5.9. Gene consolidation .................................................................................................... 124
   5.10. Overview and the final frontier ................................................................................ 124

References .......................................................................................................................... 129
1. Introduction

Quantitative expressions of change underlie the ordered chain of events represented by a phylogenetic tree constructed from species differences in a protein or polynucleotide sequence. Macromolecules can also display qualitative change, reflected in the type of monomers substituted during evolution. In the event the genetic code evolved from a smaller code, with few amino acids, an ancient protein that formed before completion of the code would conserve a smaller and earlier set of amino acids than in the standard code. Qualitative expressions of change may then prove significant in ordering events during protein evolution before divergence of Bacteria, Archaea and Eukarya. Studies on pre-divergence protein evolution could provide insights into the structure and function of the first proteins, their role in the evolution of cells, and in the transition to a DNA genome.

Protein and polynucleotide sequences broadly converge to homology at a chemolithoauto-trophic thermophilic ancestor (Fox et al., 1980; Woese, 1987, 2000; Kandler, 1998) from the pre-Isuan era (Mojzsis et al., 1996; Eiler et al., 1997). Species differences between protein sequences that evolved from the same gene in the last common ancestor (LCA) are extinguished at this node. There is evidence, however, that it marks the point at which amino acid species forming an ancestral protein sequence acquire significance. Eck and Dayhoff (1966) were prompted to suggest, soon after the code became known, that the small number of simple amino acids conserved within low-potential ferredoxin (Fdxn) from a non-photosynthetic anaerobe, Clostridium pasteurianum, indicated Fdxn predated completion of the code. Amino acid frequencies in the 26-residue internal repeat in Fdxn, from this source, were subsequently shown to match codon assignments at stage 5 of code evolution (Davis, 1999a, Section 11.4). While tandem duplication of the ancestral Fdxn gene had plainly followed formation of the stage-5 code, this event seemingly occurred well before development of a stage-14 code, with a standard set of 20 amino acids.

The scope of this approach for elucidating pre-divergence protein evolution clearly rests on the availability of archaic residue sequences. Koonin and Galperin (1997) and Kyrpides et al. (1999) have established that a substantial fraction (39–70 per cent) of proteins from bacteria and archaeal sources contain ancient conserved regions. Complete sequences for the whole genome of a number of archaeons (Bult et al., 1996; Klenk et al., 1997; Smith et al., 1997), bacteria (Fleischmann et al., 1995; Fraser et al., 1995; Himmelreich et al., 1996; Kaneko et al., 1996; Blattner et al., 1997; Tomb et al., 1997) and eukaryotes (Goffeau et al., 1996) were available to these investigators. They also benefited from the availability of sensitive algorithms for locating statistically detectable levels of sequence homology (Altschul et al., 1997), a development that was also significant in the present investigation.

A code-imposed restriction on the number of amino acid species is solely a feature of protein synthesis in the pre-divergence era. Freezing of codon assignments before emergence of the LCA accounts for this. The virtually universal distribution of the standard code requires that codon assignments were complete when species diverged, apart from the context-dependent assignment of AUG to N-formyl-methionine (f-Met) in the Bacteria code (Marcker and Sanger, 1964). Furthermore, they remained almost unchanged in all later organisms (O'Sullivan et al., 2001). Addition of the final amino acid (f-Met) to the Bacteria code following divergence from the LCA, suggests that evolution of the ‘universal’ code continued until the end of the pre-divergence era.
Biosynthetic evidence indicates the first code (stage 2 of code evolution) contained the four N-fixer amino acids, Asp\(^1\), Glu\(^1\), Asn\(^2\), Gln\(^2\) (Davis, 1999a). They form on paths extending only 1 or 2 reaction steps from the citrate cycle. The earliest proteins, consequently, had polyanionic quaternary residue sequences. The ‘before and after’ order of codon sets assigned to 11 distinct precursor/product amino acid pairs revealed that 5 transitions emanated from the NAN set, with no back transitions. They were accordingly identified as the earliest set of codons in the code \((p = 7.8 \times 10^{-3})\). Expansion from the 16 triplets of the NAN set, in the N-fixers code, to all 64 possible triplets (stage 4–7) added 10 simple amino acids, with increasingly long hydrophobic side chains, synthesized on progressively longer pathways. Synthesis of polyanionic and non-polar residue sequences, with as many as 14 kinds of amino acids, then became possible. On conclusion of the final phase of code evolution (stage 9–14), a full set of 20 amino acids, including residues with basic groups and ring structures, were incorporated into proteins.

Several previously identified regularities in the code (Woese, 1965; Nirenberg et al., 1966; Dunnill, 1966; Perlwitz et al., 1988; Taylor and Coates, 1989; Davis, 1999a; Garrett and Grisham, 1999) were compatible with the biosynthetic path of code evolution (cf. Table 5). The model also revealed (Davis, 2002) why some precursor–product amino acid pairs do not have nearest-neighbour codons, in contrast to an early biosynthetic model of code evolution (Wong, 1975; Amirnovin, 1997; Di Giulio and Medugno, 2000; Ronnenberg et al., 2000). Evidence corroborating the biosynthetic order of amino addition to the code (Davis, 1999a) will be given in this paper that is derived from the primary structure of ‘fossil’ proteins.

Residue sequences in 10 pre-divergence proteins from over 800 phylogenetically distinct sources were examined in this survey. Their time of origin, in relation to stages identified in code evolution, has been specified from the residues conserved in each ancestral sequence. Consistent with earlier results (Davis, 1999a, b), low-potential Fdxn and proteolipid (PL) subunit of H\(^+\)-ATPase will be shown to predate prokaryote septation protein, FtsZ, RNA polymerase (RNAP) core subunits, ribonucleotide reductase (RNR), and 5'→3' flap exonuclease, FEN-1. Reverse transcriptase (RT) and topoisomerase I (Topo-I) formed last, after all 20 amino acids in the standard set had entered the code. A phylogenetic analysis of Fdxn residue sequences from Bacteria and Archaea domains is presented in the following section. It reveals the LCA Fdxn residue sequence contained an internal repeat and this enabled reconstruction of an early, pre-duplication Fdxn antecedent. The structure of this small protein shows adaptations consistent with anchoring an inorganic cofactor to a positively charged surface in a primordial mineral system. In Section 3, a parallel analysis is presented on transmembrane PL helices. Ancestral PL appeared representative of a later stage of pre-divergence evolution, when proteins had developed a hydrophobic domain. Formation of a permeable cell membrane was found to coevolve with this development. Synthesis of deoxyribonucleotides from ribonucleotides involves formation of radical intermediates sequestered within the hydrophobic domain of a reductase (Reichard, 1993). Synthesis of DNA seemed likely, therefore, to occur comparatively late in code evolution. To clarify when, and how, the transition to a DNA genome took place, residue sequences for RNR and a number of replication proteins were examined, and the results appear in Section 4. The estimated time of origin of conserved and non-conserved regions were compared among all 10 ‘fossil’ proteins (Section 4), to test of the validity of the biosynthetic order of amino acid addition to the code.
2. Low-potential ferredoxin as a prototype of pre-cell proteins

Clostridium pasterianum \([4\text{Fe-4S}]\) Fdxn is representative of a group of small proteins involved in e^− transfer. Its sequence contains only 55 amino acids and has an internal repeat of around 26 residues. There are two \([4\text{Fe-4S}]\) clusters and the protein has a reduction potential of \(-417\) mV at neutral pH (Aron, 1965). This is close to the reduction potential of \(\text{H}_2\) and about 100 mV more electronegative than pyridine nucleotides. Their low reduction potential and occurrence in bacteria, archaea, and plant chloroplast photosystem I complex (Sticht and Rosch, 1998)

![Fig. 1. Sequences of low potential Fdxn from prokaryote species. Each of these 55-residue sequences showed homology (identity, 28–89 per cent) with Clostridium pasterianum Fdxn, in a database search using gapped-BLAST and PSI-BLAST sequence-matching algorithms. They include 22 Fdxn from Archaea and 60 from Bacteria. The biological source and database access number of each sequence is given. At left of each sequence is the location of its first residue in the whole protein. Single-letter amino acid designations and protein contractions appear under Abbreviations. A pair of italic letters marks the location of a gap. Sites resolved to not less than three residues in the most parsimonious ancestral sequence are designated as: a, for the set \{E,P,N,A,I,T,K,F,W\}; b, \{D,N,I\}; c, \{V,T,R\}; d, \{E,L,M\}; e, \{E,P,S,G\}; f, \{I,V,T,L\}; g, \{D,N,S,Y\}; and h, \{G,D,I,R,Y\}. Highlighted letters mark amino acids identical with that at a site displaced by 29-residues in the ancestral sequence. Twenty-two amino acid residues, between site-1 and -23, matched residues at site-30 to -52 of Fdxn in LCA (internal identity, 95.7 per cent, 22/23). Two segments of unmatched residues occur; a central 6-residue segment at site-24 to -29, and a C-terminus 3-residue segment at site-53 to 55. These Fdxn sequences were obtained in a search of European Molecular Biology, GenBank, Swiss Protein and Tokyo databases.](image-url)
are consistent with an early origin, at a time when ambient conditions on Earth were still reducing.

Sequences homologous with \emph{C. pasterianum} Fdxn were identified in a search of the European Molecular Biology, GenBank, Swiss Protein and Tokyo database. Gapped-BLAST and PSI-BLAST algorithms (Altschul et al., 1997) performed with the aid of the GenomeNet BLAST 2 Server, Tokyo Centre facilitated this search. Eighty-two Fdxn sequences were obtained (Fig. 1), 22 from 10 archaea species and 60 from 52 species of bacteria. Twenty-eight to 89 per cent of residues in these sequences were identical with residues in the aligned reference sequence. The primary structure of ancestral Fdxn was established initially. Amino acids conserved by ancestral Fdxn for each species domain were attributed to LCA Fdxn. An internal homology in its sequence allowed reconstruction of a pre-duplication Fdxn antecedent. Phylogenetic trees for Archaea and Bacteria Fdxn were constructed by the neighbour-joining method (Saitou and Nei, 1987). Branch lengths correspond to the aggregate number of codon base differences between respective template sequences (Jukes and Cantor, 1969). Outgroup sequences from an alien domain helped to locate the LCA node. The phylogenetic depth of this node was checked by assessing the dependence of internal identity on distance from outgroup sequences.

Fig. 2 shows the Archaea Fdxn tree. \emph{C. pasterianum}, \emph{Peptostreptococcus asaccharolyticus}-1 and \emph{Rhodospirillum rubrum}-1 sequences (Fig. 1) formed an outgroup of non-archaeon Fdxn. They cluster on a branch, in the upper half of this tree, that has a node with sequence-3 and -4 of the anaerobic methanogen, \emph{Methanobacterium thermoautotrophicum}.
Given that the internal repeat in Fdxn resulted from a pre-divergence error in replication (Eck and Dayhoff, 1966; Schwarz and Dayhoff, 1978a, b), sequences near the tree root are likely to show elevated internal homology. Archaea Fdxn displays the anticipated dependence on phylogenetic distance. The number of identical amino acids at equivalent sites in each repeat decreased linearly with distance from the tree root (Fig. 3a),

\[ n = 22 - 38d, \]

where \( n \) is the number of identical amino acid residues and \( d \) is the sequence distance from the root. The fall-off of \(-38\) identical residues/unit distance had a 98 per cent confidence interval, \([-63.2, -2.8]\), that excluded a non-negative slope. Sequence distances were measured from the average mid-point between outgroup species; mid-point (root) = \( v^{-1}(\sum d_i/2) = 1/2(\sum d_i/v) \), where \( d_i \) is the distance between the \( i \)th pair of outgroup sequences and \( v \) refers to the number of pairs (Fig. 3a). The resulting distances are to a fixed point proximal to the base of the outgroup cluster and tree root. Sequences aligned to \( C.\ pasterianum \) Fdxn show a repeat at a 29-residue span (Fig. 1), attributed by Eck and Dayhoff (1966) to 26-residue repeats with a 3-residue linker segment.

Bacteria Fdxn separate into two nearly equal deeply branching phylogenetic groups (Fig. 4). Anaerobes broadly form one group and aerobes the other. Of 33 sequences in the upper group, 21 were from anaerobes. Between \( Chl\ limicola\)-1 and \( Ch\ vinosum\)-2, anaerobes accounted for 20 of 24 sequences. Only seven of 27 sequences in the lower group were from anaerobes. The
Fig. 3. Change in internal homology with distance in the phylogenetic tree for low potential Fdxn from each prokaryote domain. (a) Twenty-two Archaea Fdxn show a significant fall-off in number of identical amino acids with distance. Internal homology between 26-residue Fdxn repeats, 29-residues apart, thus diminished as species diverged. A least-squares equation, \( n = 22.01 - 38.03d \) \( (r^2 = 0.34) \), relates, \( n \), number of identical amino acids between duplicate segments and \( d \), distance. Its slope had 98 per cent confidence limits of \([-63.2, -2.8]\). (b) Internal homology vs. distance among 60 Bacteria Fdxn. Least-squares regression gave a linear equation, \( n = 9.75 - 4.16d \) \( (r^2 = 0.009) \), whose 90 per cent confidence interval on the slope, \([-13.63, 5.31]\), spanned a zero decline in identity with distance. The extreme position of anaerobe Peptostreptococcus elsdenii (no. repeats, 13; distance to root, 0.49) contributed to damping the fall-off in internal homology with distance. All distances were measured from the mid-point between outgroup sequences. Open circles, Fdxn from aerobic species; closed circles, anaerobic species. Number of identical amino acids in duplicate segments is from sequences in Fig. 1. Distances are from trees in Figs. 2 and 4.

Fig. 4. Phylogenetic tree of low potential Fdxn from Bacteria. The tree contains 60 bacterial and 3 archaeal Fdxn and was constructed by the neighbour-joining method. Branch lengths depend on the estimated number of nucleotide differences separating templates for the proteins examined and reflect the time elapsed since their divergence. Two deeply branching groups of Fdxn occur. The three archaeal Fdxn (outgroup) cluster on a mid-tree branch, together with three sulphur-reducing bacteria of the lower group. Most anaerobes (marked by *) are in upper group. Parentheses contain the number of identical amino acids between 26-residue duplicate segments. Bold type indicates Fdxn from an outgroup source. A small circle specifies the mid-point between outgroup sequences. A bar below the tree provides a reference distance for branch length. Species abbreviations are from Fig. 1.
departure from monophyly probably involved horizontal gene transfer between aerobic and anaerobic bacteria (Doolittle, 2000).

The antiquity of anaerobes would suggest the root of the Bacteria Fdxn tree is in the upper group. Archaea Fdxn, *Archaeoglobus fulgidus*-2, *Methanosarcina thermophila*-2 and *Sulfolobus* strain 7 branched at mid-tree, with Fdxn from three sulphate-reducing anaerobic bacteria of the lower, predominantly aerobic group (Fig. 4). A fall-off occurred, nonetheless, in bacterial Fdxn internal homology (Fig. 3b),

\[ n = 9.75 - 4.16d, \]

where \( n \) and \( d \) retain their previous meaning. Sequence distance was again measured from the midpoint between outgroup sequences (Fig. 3b). The slope of \(-4.16\) identical residues/unit distance, however, had a 90 per cent confidence interval of \([-13.63, 5.31]\) and it is seen to be too shallow to exclude ‘no fall-off’ among the 60 bacteria Fdxn. Even so, sequences combining high internal homology and proximity to tree root (Fig. 3b) were from anaerobes. A statistically significant fall-off in internal homology with distance does occur, when the Bacteria tree root is located at the average mid-point between the 28 anaerobic species. At this location, \( n = 10.42 - 7.14d, \) with 90 per cent confidence limits on the slope of \([-13.39, -0.89]\). Notwithstanding a weak fall-off in internal identity with distance to the Bacteria Fdxn tree root (Eq. (2)), ancestral Fdxn for Bacteria showed high internal homology—identity index, 78.3 per cent (cf. Fig. 6).

Species differences occur in length of the linker sequence between Fdxn repeats (Wakabayashi et al., 1983). There are also species differences within repeat segments (George et al., 1985). A quantitative evaluation of internal homology at different spans among Fdxn residue sequences was undertaken and the results are given in Fig. 5. It shows the spectrum of sequence periodicities for non-overlapping repeats displaced by an indicated number of residues (span), among 1000 sequences bootstrapped from amino acid frequencies for Fdxn sequences in Fig. 1. The fraction of sites with identical amino acids was independent of span length over all spans,

\[ n = 7.74 - 0.073d \]

as the fall-off in identical sites (per cent) per unit increase in span length had a 95 per cent confidence interval of \([-0.157, 0.011]\). An identity index of 30.15 per cent occurred at span-29, and this far exceeded other spans. It also conspicuously exceeded the 99.8 per cent confidence interval on mean identity (mean, 5.62 per cent).

Occurrence of two 26-residue segments, separated by a 3-residue insert, akin to *C. pasterianum* Fdxn (Eck and Dayhoff, 1966), thus fits periodicity in these 55-residue Fdxn sequences (Fig. 1). Ancestral Fdxn based on phylogenetic analysis of 82 Fdxn sequences in Fig. 1 fit more closely 23-residue repeats with a 6-residue insert (cf. Fig. 6). Thirteen additional span lengths (3, 4, 14, 19, 22, 23, 25, 26, (29), 32, 33, 35, 36, 39) exceeded the upper limit of the 99.8 per cent confidence interval on expected identity, after omitting the span-29 value. Peaks in identity were spaced at \( 2.8 \pm 0.7 \) residues (\( \bar{X} \pm \text{SEM} \)). This repeat conceivably resulted from successive repeats of an oligomer antecedent (Eck and Dayhoff, 1966). Alternatively, secondary structure has constrained the primary structure of this early protein. The bootstrap procedure (Efron, 1992) was applied here to a statistically significant global feature of phylogenetically related sequences. Since Felsenstein (1985), bootstrapping has generally been used to appraise branching at each tree node, albeit is seldom significant.
Fig. 1 shows the most parsimonious (Fitch, 1971) LCA Fdxn sequence based on the phylogenetic distribution of Fdxn sequences in Figs. 2 and 4. Thirty-two of 55 sites in this sequence were resolved to a single amino acid. Doublets arose at 15 sites, and 8 sites were

Ancestral ferredoxin - Archaea:

ADKD
V
T
H

WI
PN
KE
AF

ESCIGCG
K
L
G

EA
ASP
ER
G

CVIVCPRAEG
IN
SA

EL
MV
WN

KDIA
V
P

VE
VE

VNEEKCVGC
S
Y

AC
VA

NVPVRALI
YE

VK
E

Fig. 5. Spectrum of internal homologies in low potential Fdxn residue sequences. Identity index values specify homology between non-overlapping segments, in a 55-residue sequence, displaced by an indicated number of residues (span). Each point is the mean identity among 1000 sequences. They were generated by random resampling (with replacement) of the amino acid frequency distribution at each site observed in 82 low potential Fdxn sequences. A conspicuous maximum in mean identity index occurred at a 29 residue span. Some neighbouring spans also had elevated internal homology; span-22, -26, -32 and -33 values exceeded the upper limit of the 99.8 per cent confidence interval. Identity between sequence segments and span length were related by the least-squares regression equation; $\hat{h} = 7.74 - 0.074s (r^2 = 0.06)$, where $\hat{h}$ is the statistical estimator for identity index (per cent) and $s$ is span length. The slope had a 95 per cent confidence interval of \{\(-0.157, 0.011\)\}, consistent with no significant change in identity index with span. Residual errors in the least-squares regression were gaussian, $N(0, \sigma)$, on omission of the extreme identity at span-29. Residual errors were also independent of span length, according to the Durbin–Watson test. Segment identities at span-3, -4, -14, -19, -22, -23, -25, -26, -32, -33, -35, -36, and -39 exceeded the upper limit of the 99.8 per cent confidence interval, on exclusion of the span-29 value. The range of spans with elevated homology had an underlying repeat with an observed mean of $2.8 \pm 0.7$ (\(\bar{x} \pm \text{SEM}\)). Amino acid frequency distributions in this analysis were based on aligned sequences in Fig. 1.
Fig. 6. Evolution of low potential Fdxn in prokaryote domains. A 23-residue protein, pro-Fdxn-[5], is shown to have formed around stage-5 in code evolution. Three G:A transitions in the pro-Fdxn-[5] gene form pro-Fdxn-[10]. The pro-Fdxn-[10] gene doubled in size to code for a 46-residue protein with internal symmetry. This event appears to have occurred soon after formation of the stage-10 code, judging from the pro-Fdxn-[10] residue profile. Insertion of a central 6-residue segment and a 3-residue addition at the C-terminus produced a 55-residue molecule in the last common ancestor, LCA. Seven single and two double base changes then formed ancestral Fdxn for Bacteria. The path to Archaea ancestral Fdxn shows three single and four double base changes in the LCA Fdxn gene. Amino acids at mutated sites in pro-Fdxn-[5] (site-7, -9), and in both intermediates between LCA and prokaryote ancestral Fdxn, arose at the equivalent sites in the most parsimonious ancestral Fdxn sequence of either Archaea or Bacteria. At site-23, pro-Fdxn-[5] contains an amino acid from the immediate descendant of an ancestral sequence. Each sequence minimized the number of base changes required for each transition. [A], Archaea ancestral Fdxn; [B], Bacteria ancestral Fdxn. Bold letter, amino acid residue identical to that at equivalent site in LCA Fdxn sequence. The LCA Fdxn sequence appears in Fig. 1. Ancestral Fdxn for Archaea and Bacteria was the most parsimonious sequence at the root of the respective phylogenetic trees depicted in Figs. 2 and 4.
resolved to 3+ residues. Ancestral Fdxn for Archaea can be seen to have 29, 4, and 22 sites, respectively, resolved to 1, 2, and 3+ amino acids. While ancestral Fdxn for Bacteria had 32, 10, and 13 sites with 1, 2 and 3+ amino acid residues, respectively. Enclosed amino acids indicate sites not resolved beyond 5 amino acids. Amino acids conserved between prokaryote domains are underlined.

High internal homology characterized the LCA Fdxn residue sequence (Fig. 1). Twenty-two identical amino acids occurred between two 23-residue segments, separated by a 6-residue linker sequence. Nineteen of these internally conserved amino acids entered the code at stage 5, or earlier (Davis, 1999a). Ferredoxin has long been considered an ancient protein that originated before the genetic code was complete (Eck and Dayhoff, 1966; Schwarz and Dayhoff, 1978a, b). A stage-7 amino acid (Ile) was internally conserved at two sites and a stage-10 amino acid (Lys) at one site in LCA Fdxn (Fig. 1). Tandem duplication of the ancestral Fdxn gene thus occurred no earlier than stage-10 of code evolution.

Fig. 6 depicts a reconstruction of Fdxn evolution up to prokaryote divergence. Amino acids conserved between each repeat segment in LCA Fdxn (Fig. 1) form a pre-duplication stage-10 molecule, pro-Fdxn-[10]. Nineteen of its 23 residue sites were resolved to a single amino acid. Site-1, -5, -6 and -16 were resolved to amino acid pairs (Val, Ala), (Glu, Pro), (Glu, Asp), and (Asn, Asp), respectively. The first amino acid in each pair was assigned to pro-Ferredoxin-[10], minimizing the number of base changes at site-5, -6, and -16 (Fig. 6). Valine was placed at site-1, because it includes an initiation codon (GUG) in its codon set (GUN). A centrally located 6-residue insert (site 24–29) occurs in LCA Fdxn resulting in symmetry at span-29 (Fig. 5). Insert residues either showed no identity with sites at span-29 (site 24–26 vs. 53–55), or had no counterpart (site 27–29). The chance of 6 consecutive unmatched residues has a probability of only $3.61 \times 10^{-5}$, or $(10/55)^6$, given that 10 of 55 residues were unmatched over the whole sequence. A 3-residue addition at the C-terminus of LCA Fdxn (Fig. 6) contained Ile, Val, Glu in both Archaea and Bacteria.

Table 1 shows mutation distances separating LCA Fdxn from ancestral Fdxn for Archaea and Bacteria. The minimum number of base differences between templates for these proteins reveals the LCA gene to be positioned midway between ancestral Fdxn genes for each prokaryote domain. Seven single-base and 2 double-base changes in the LCA Fdxn gene led to ancestral Fdxn for Bacteria (Fig. 6). Formation of ancestral Fdxn for Archaea required 3 single-base and 4 double-base changes. With 11 base changes attributed to each path, ancestral Fdxn genes place Archaea and Bacteria equidistant from the LCA (Table 1). Diversification from the LCA

<table>
<thead>
<tr>
<th>Source</th>
<th>Archaea</th>
<th>LCA</th>
<th>Bacteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archaea</td>
<td>0</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>LCA</td>
<td>0.07</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Bacteria</td>
<td>0.13</td>
<td>0.07</td>
<td>0</td>
</tr>
</tbody>
</table>

*a Upper right half of table lists the minimum number of base differences between their templates. Lower left half gives the corresponding Jukes-Cantor distance. Based on results in Fig. 6.
occurred at a similar rate for both species domains, according to these results. This is likely to be at the upper end of estimates for Archaea formation. Kollman and Doolittle (2000) examined eight sets of genes and they found a significantly slower rate of change for Archaea vs. Bacteria and Eukarya.

Eight of 11 base changes in the ancestral Bacteria gene were expressed at amino acid residue site 2–6 and site 31–35 (equivalent sites in duplicate segment). While 3 of 11 base changes leading to ancestral Fdxn for Archaea were expressed at these sites. Combining mutations on both paths from the LCA Fdxn gene reveals that half (11/22) of all base substitutions occurred in only 18 per cent ((15 × 2)/165) of the bases in this gene. This fast evolving 15-base sequence coded for a highly acidic region (2 Asp + 2 Glu in a 5 residue segment, site 2–6) in pre-duplication Fdxn.

The acidic region spans 6 (site 2–7) of 23 residues in pro-Fdxn-[5] (Fig. 6). During formation of ancestral Fdxn for Archaea and Bacteria, 2 and 4 residues, respectively, were substituted in this 6-residue segment. This highly negative region is a putative attachment site for pro-Fdxn that anchored it to a positively charged mineral surface of a primordial system (cf. Fig. 9). Elevated mutability in this attachment segment conforms with a loss of function, following replacement of the primordial surface system when cells appeared.

Evidence for an earlier Fdxn antecedent, pro-Fdxn-[5], arises in the primary structure of pro-Fdxn-[10] (Fig. 6). Fig. 7 compares amino acid frequencies in pro-Fdxn-[10] with their frequencies in the 82 extant Fdxn sequences in Fig. 1. Both distributions show a statistically acceptable goodness of fit in a Kolmogorov–Smirnov test; $D_{\text{max}}(n : 23, 55) = 0.138, 0.20 < p < 0.30$. This suggests Fdxn has changed only slowly over a long interval. Consistent with its great antiquity, early amino acids are over-represented. Amino acids from stage-5, or earlier, form 87 per cent of pro-Fdxn-[10] residues. In extant Fdxn, 73.2 per cent of amino acids are of the same vintage. Their frequency in Fdxn, and its antecedent, substantially exceeds their codon representation (52 per cent) within the standard code. It also exceeds their frequency (59 per cent) in a sample of 37 monomeric proteins (Miller et al., 1987).

Only three sites in pro-Fdxn-[10] contain post-stage-5 amino acids (Fig. 6), Lys (stage-10) at site-7 and Ile (stage-7) at site-9 and -23. Their low frequency (13 per cent) is consistent with substitution at sites previously occupied by pre-stage-7 amino acids. Fig. 7 depicts amino acid frequencies in pro-Fdxn-[10] distributed by stage of entry into the code (Davis, 1999a). They demonstrate a maximum goodness of fit with the codon array at stage 6.470.53 (% X$^2$/SEM). The probability that a 23-residue sequence with this residue profile could arise from the standard code can be expressed as

$$p_{\text{seq}} = \frac{\phi_{\text{aa}}}{\phi_{\text{st}}},$$

$$\phi = \frac{n!}{\pi_{g} n_{g}^{\prime} \pi_{s} n_{s}^{\prime} \phi_{s}^{\prime \prime}},$$

where $p_{\text{seq}}$ is the conditional probability. It corresponds to a ratio of multinomials for the probability of the distribution of entry stages among amino acids, $\phi_{\text{aa}}$, and for the probability of the distribution of entry stages from the standard set of codons, $\phi_{\text{st}}$. $\phi_{s}$ is the probability of a codon assigned at stage-s in the standard code. $n$ and $n_{s}$ are, respectively, the total number of sequence residues and number with a stage-s entry to the code. Pro-Fdxn-[10] has a formation probability, $p_{\text{seq}}$, equal to $1.67 \times 10^{-3}$, or $5.12 \times 10^{-3}/3.08 \times 10^{-5}$. This shows that
the pro-Fdxn-[10] residue profile differs significantly from the residue profile representative of the standard set of codon assignments, consistent formation before the code was complete.

Although pro-Fdxn-[10] had a stage 6.4 residue profile (Figs. 6 and 7), it actually contained no stage-6 amino acids (Thr). Hence, Fdxn can be traced back to stage-5 of code evolution.

Reconstruction of pro-Fdxn-[5] requires replacement of Lys^{10} (site-7) and Ile^{7} (site-9, -23) with amino acids of a stage 5, or earlier, ‘code age’. Glutamate^{1} occurs at site-36 (equivalent to site-7 in duplicate segment) in the most parsimonious ancestral Fdxn sequence for Bacteria. Its codon (GAR) specifies Lys^{10} (pro-Fdxn-[10] residue), following a 5'-G:A substitution. Valine^{4} most closely matched Ile^{7}. It occurs at site-38 (equivalent to site-9 in duplicate segment) in ancestral Fdxn for Archaea. A 5'-G:A transition in Val^{4} codons (GUY, GUA) yields Ile^{7} (Fig. 6). Valine^{4} did not occur at site-23 and -52 (source of second Ile^{7} in pro-Fdxn-[10]) in ancestral prokaryote Fdxn. It does occur however in their immediate descendants. The pro-Fdxn-[5] sequence could be reconstructed, therefore, solely from early (pre-stage 6) amino acids in phylogenetically related sequences. Furthermore, its sequence minimized the mutation distance to pro-Fdxn-[10].
A conspicuous feature of pro-Fdxn-[5] primary structure is a 6-residue region with five acidic amino acids at the N-terminus. This region contains all acidic residues in pro-Fdxn-[5]. No other charged amino acid residues occur in this molecule; incorporation of basic amino acids did not occur until stage 9–10. Its remaining 16 residues (C-terminal segment) are mainly non-polar, apart from Asn (site-16). Cysteine residues occur at site-8, -11, -14 and -18. In low potential Fdxn, each Cys-S covalently bonds to an Fe atom, normally in a cubic [4Fe-4S] cluster. They show oxidized/reduced states of [4Fe-4S]$^{2+}$/+ and reduction potentials of $-250$ to $-600$ mV. Low-potential Fdxn with a single [4Fe-4S] cluster function in bacteria *Desulfovibrio africanus*, *D. gigas*, *Thermotoga maritima* and archaeons *Bacillus thermoproteolyticus*, *Pyrococcus furiosis* and *Thermococcus litoralis* (Sticht and Rosch, 1998).

The presence of two structurally distinct segments in pro-Fdxn-[5] suggests this small ancient protein combined two different functions. Fig. 9 shows an e$^{-}$ transfer centre in the C-terminal segment and a negatively charged attachment site at the N-terminus. All five carboxyl groups in the attachment segment would be ionized above pH 4 in an aqueous medium. They could thereby anchor this small protein to a positively charged surface. The proposed structure for pro-Fdxn-[5] provides compelling support for a chemolithoautotrophic model for the origin of life. Wachterhauser (1988, 1992, 1997, 1998) retrodicted biosynthetic pathways and evidence emerged of a primordial surface system that had preceded cell formation. It appeared to couple exergonic oxidation of FeS, in the presence of H$_2$S, with organosynthesis of surface-bound multianionic molecules (cf. Section 5.10).

---

1 A mini-Fdxn with only 31 residues and single cubane [4Fe-4S] cluster, liganded by Cys6, Cys9, Cys12, Cys25, had a redox potential of $-370$ mV (Sow et al., 1996).
3. **H⁺-ATPase proteolipid is a marker for cell membrane formation**

A- and F-ATP synthase couple dissipation of an electrochemical potential, resulting from a difference in proton concentration across a cell membrane, to the synthesis of ATP. Alternatively, A-, F- and V-ATPase hydrolyze ATP to pump protons against a concentration gradient. Core subunits of A₁A₀-ATPase (synthase/hydrolyase) from archaeons, F₁F₀-ATPase (synthase/hydrolyase) from bacteria, mitochondria and chloroplasts, and V₁V₀-ATPase (hydrolyase) from eukaryote endomembranes are structurally related. Each type of H⁺-ATPase thus evolved from a common ancestral gene (Muller et al., 1999) that predated species divergence (Gogarten and Taiz, 1992).

Denda et al. (1989) noted the ATPase operon in *Sulfolobus acidocaldarius* contained a gene, atpP, for a hydrophobic 10 kDa protein with a tandem 11-residue repeat. The repeat occurred in one of three transmembrane ₜ-helices of the ATPase PL subunit, judging from the hydropathy profile of its residue sequence. An uneven chromosome crossover during evolution of *S. acidocaldarius* could have produced the repeat. Alternatively, it could have arisen far earlier, through duplication and fusion of an antecedent of atpP that coded for a 1 kDa protein. *S. acidocaldarius* is a hyperthermophilic facultative autotroph, from one of the deepest phylogenetic branches of Archaea (Woese, 1987; Barnes et al., 1996), it is accordingly well placed to retain evidence of an early gene duplication. A resolution of this issue has significance for pre-divergence evolution.

Fig. 9. Molecular model of an early antecedent of low potential Fdxn (length, 23-residues) that appears to have formed at around stage-5 of code evolution. Its C-terminal region has four Cys that ligand a [4Fe-4S] cluster. A negatively charged segment occurs at the N-terminus. Pro-Fdxn-[5] could thereby anchor its inorganic cofactor to a positively charged surface and facilitate e⁻ transfer in redox reactions within a primordial surface system before cells evolved. Residues with a negative group, the carboxyl of Asp and Glu, and Cys-γ-S atoms, coordinately bonded to Fe atoms in Fe-S cluster, are indicated. The pro-Fdxn-[5] residue sequence is from Fig. 6. This model was constructed from the high resolution X-ray structure of 1[4Fe-4S] Fdxn from hyperthermophile *T. maritima* (Macedo-Ribeiro et al., 1996) using the Swiss-Pdb viewer (http://www.expasy.ch/spdbv/) of Guex et al. (1999).
S. acidocaldarius PL h1 residue profile is consistent with duplication of the PL gene in the pre-divergence era (Davis, 1999a), apart from a lack of acidic residues and their amides. They were either lost, or never incorporated into pre-duplication PL. The absence of negatively charged residues suggests direct attachment to a mineral surface was not mandatory when PL evolved. Phylogenetic evidence is presented that establishes it formed during the pre-divergence era, when formation of this small hydrophobic protein would be conditional on (externally negative) phospholipid bilayer formation. In view of this dependence, the time of origin of PL provides a marker for development of cells bounded by a fluid mosaic membrane. The estimate of when cells appeared, inferred from the ‘code age’ of the PL residue profile, agrees with estimates derived from other sources (Section 5.7).

Homologues of transmembrane regions in S. acidocaldarius PL were identified in the European Molecular Biology, GenBank, Swiss Protein and Tokyo database. As before, BLAST algorithms assisted the search. H+-ATPase PL from 10 archaeons, 42 eukaryotes, 24 eubacteria, 12 mitochondria, and 19 chloroplasts were obtained. They included 11 A-, 42 V- and 54 F-type ATPases.

The archetypal PL subunit of H+-ATPase forms an 8 kDa hairpin, whose arms are transmembrane α-helix-1 and -2. Fig. 10 lists 22-residue segments of PL h1 and h2 from 107 sources, with their biological source and accession number. The repeat identified by Denda and colleagues is in helix-1 and, as anticipated from their report, it has higher internal homology than helix-2 (Fig. 10). Nine of 11 (internal identity index, 81.8 per cent) amino acids were identical between repeats in LCA PL h1. While four residues were identical in helix-2 repeats (internal identity index, 36.4 per cent).

Fig. 11 portrays the phylogenetic kinship between PL h1 sequences of A-ATPase. Outgroup F-ATPase PL h1 sequences from eubacteria Aquifex aeolicus, Synechococcus PCC6301 and T. maritima branch with thermophilic methanogens, Methanobacterium thermoautotrophicum and Methanococcus jannaschii. A least-squares regression analysis established that the number of identical amino acids between A-ATPase PL h1 internal repeats decreased with sequence distance from outgroup sequences (Fig. 12);

\[ n = 8.32 - 10.99d. \]  

The slope had a 98 per cent confidence interval of \((-20.99, -0.99)\), indicating a significant fall-off in internal identity within A-ATPase PL h1 repeats with distance from outgroup sequences. \( n \) and \( d \) retain their previous meaning (Eq. (1)), distance being measured from the mid-point between outgroup residue sequences. Pyrococcus abyssi and P. horikoshii share an identical PL h1 sequence and consequently have the same coordinates (Fig. 12): repeats, 6, and distance, 0.265. Sulfolobus species PL h1 exhibit high internal homology (repeats, 9 and 8; internal identity index, 81.8 and 72.7 per cent), in accord with Denda et al. (1989).

Internal homology in Sulfolobus PL h1 exceeded the highest values in repeats within 54 F-type and 42 V-type enzymes. Phylogenetic trees for PL h1 of F- and V-type ATPase appear in Figs. 13 and 14, respectively.

Among F-type PL, helix-1 attained an internal identity of 54.5 per cent (repeats, 6) in A. aeolicus, Bacillus stearothermophilus and Helio bacter pylori. Equal homology occurred in V-ATPase PL h1 from Acetabularia acetabulum, Drosophila melanogaster, Heliothis virescens, Manduca sexta and Neurospora crassa. A higher internal identity (63.6 per cent; repeats, 7) was
Fig. 10. Sequences of transmembrane helix-1 and -2 in PL subunit of H+-ATPase were identified in a search for homologues of *Sulfolobus acidocaldarius* PL. One hundred and seven sequences were found, each representing two 22-residue segments. They included 11 PL from A-ATPase (archaeal ATP synthase), among them the *Thermus thermophilus* enzyme acquired by horizontal transfer, 54 from F-ATPase (bacteria ATP synthase) and 42 from V-ATPase (eukaryote ATP hydrolase). Amino acids at 15–42 sites in h1+h2 sequences were identical with residues at equivalent sites in the reference sequence: identity, 34.1–95.5 per cent. Biological source and database access number of each sequence is given. At the left of each helix segment is the location of its first residue in the whole protein. A pair of italic letters marks a gap introduced during sequence alignment. Deleted residues in the native sequence are indicated by a dash. Highlighted letters mark identical amino acids at sites separated by 11-residues in the ancestral sequence.
Fig. 10 (continued).

Fig. 11. Phylogenetic tree of PL h1 from A-ATPase. The tree relates PL h1 residue sequences (length, 22 residues) from 11 A-type and 3 F-type enzymes. Other features of this figure are as described under Figs. 2 and 10.
achieved in *Aedes aegypti*. Internal homology in PL sequences from both sources was dependent on their distance from outgroup sequences. The number of identical amino acids between repeats in F-ATPase PL h1 sequences was linearly dependent on sequence distance from outgroup sequences:

$$n = 8.41 - 18.44d.$$  

(6)
The fall-off had a 98 per cent confidence interval of \([-24.71, -12.17]\). This fall-off (\(-18.44\) identical amino acids/unit distance) is steeper than among archaeal sequences (Eq. (5)). In the V-ATPase tree (Fig. 14), a statistically significant fall-off also occurred in internal identity of PL h1 sequences with distance from outgroup sequences
\[
\lambda = 6.68 - 10.52d. \tag{7}
\]
The 98 per cent confidence interval on this slope was \([-16.42, -4.62]\).
Analysis of internal homology in PL h1 sequences (Fig. 10) supported the proposition that a predecessor of atpP duplicated and fused at some time before species divergence. LCA PL h1 sequence obtained by maximum parsimony (Fig. 10) displayed an internal homology of 81.8 per cent. Elevated internal homology in *Sulfolobus* PL h1 (Denda et al., 1989; Fig. 12) conforms with the antiquity of this genus (Fox et al., 1980; Woese, 1987; Barnes et al., 1996).

Results of a quantitative assessment of internal periodicity at different spans within sequences bootstrapped from PL h1 and h2 are given in Fig. 15. Internal identity (35.8 per cent) in PL h1 derived sequences was maximal at span 11. It far exceeded the upper limit of the 99.8 per cent confidence interval for mean repeat homology. Each identity value in Fig. 15 is the mean among 1000 sequences, generated through random sampling (with replacement) amino acids at observed frequencies per residue site among the 107 PL h1 sequences in this study (Fig. 10). Elevated identity values also occurred at span-4 (26.6 per cent), -7 (26.7 per cent), -11 (35.8 per cent) and -15 (26.8 per cent). Identity values at each of these spans exceeded the upper limit of the 95 per cent confidence interval. An underlying repeat of 3.8 ± 0.43 residues was observed. A tetramer repeat is discernible in the LCA PL h1 sequence (Fig. 10).

This result points to tandem duplication of a hydrophobic tetramer, such as Ala.Gly.Leu.Ala (cf. Fig. 16), in formation of PL antecedent, pro-PL-[7]. The Fdxn sequence also retained evidence of an oligomer repeat (Section 2). It has been suggested ancestral Fdxn consisted of successive repeats of the tetrapeptide, Ala.Asp.Ser.Gly (Eck and Dayhoff, 1966). It seems equally likely though that the low order repeat resulted from an early effect of secondary structure on primary structure.
A significant elevation in internal identity at span-11, likewise, occurred in PL h2 (Fig. 15). An identity index of 19.8 per cent existed between 11-residue repeats in sequences bootstrapped from PL h2 amino acid frequencies (Fig. 10). This exceeded the upper 99.8 per cent confidence limit on mean identity. As before, this analysis was based on mean internal identity obtained at different spans among 1000 sequences randomly generated from PL h2 sequences. An 11-residue repeat was also evident in the most parsimonious residue sequence for LCA PL h2 (unpublished results). The presence of span-11 repeats in both helix-1 and –2 sequences suggests both transmembrane \( \alpha \)-helices in the PL molecule had a common ancestor. The alternative possibility that bilayer symmetry induced these repeats can be discounted. PL h1 and h2 exhibit translational identity, while a phospholipid bilayer contains mirror symmetry.

Fig. 15. Spectrum of internal homologies in \( \text{H}^+ \)-ATPase PL transmembrane regions. Total number of identical amino acids found at sites separated by 1–18 residues are shown for sequences of 22-residues. Each point is the average identity index (per cent) in 1000 sequences produced by bootstrapping from amino acid frequencies observed at each site in helix-1 and -2 among 107 PL molecules. (a) Helix-1 had an identity index of 37.9 per cent at span-11. This far exceeded the 99.8 per cent confidence interval on expected identity. A periodicity of marginal significance also arose at span-7. No significant change in internal homology accompanied variations in span length; \( \hat{h} = 17.30 - 0.04s (r^2 = 0.0005) \), with a 90 per cent confidence interval of \( (-0.76, 0.69) \) about the slope. \( \hat{h} \) and \( s \) refer to the identity estimator and span length, respectively. (b) Helix-2 likewise had a significant periodicity at span-11. In addition, internal homology in this segment was stationary with variations in span length; \( \hat{h} = 9.19 - 0.46s \), where the 90 per cent confidence interval on this slope was \( (-0.06, 0.98) \). Residual errors in each least-squares regression analysis were gaussian, \( N(0, \sigma) \), and independent of span length in a Durbin–Watson test. These calculations used amino acid frequencies for the PL sequences listed in Fig. 10. Databases searched to obtain these sequences and alignment algorithms are given under Fig. 1.
Fig. 16. Evolution of H⁺-ATPase PL helix-1. Its origin is traced to a hydrophobic 11-residue polypeptide, pro-proteolipid h1-[7], formed soon after appearance of the stage-7 code. Before alterations to its sequence occurred, the polypeptide doubled in size. Substitutions at site-12 and -20 subsequently led to the 22-residue PL h1 sequence in the LCA ATPase subunit. A substitution at site-11 led to ancestral PL h1 of Archaea. Ancestral V-ATPase PL h1 (and h3) formation required five additional single-base substitutions, within codons for amino acids at site-11, -14, -17, -21, and -22, followed by doubling of the gene for the archetypal 8 kDa PL molecule. Formation of the ancestral F-ATPase subunit, likewise, followed five single-base substitutions in LCA PL h1 codons. Two of them were revertants to pro-proteolipid h1-[7] residues. Each sequence in this model minimized the number of base changes. Bold letters identify amino acids shared by duplicate 11-residue segments in LCA PL h1. This sequence was reconstructed by ‘majority rule’ from the most parsimonious sequence obtained for ancestral PL h1 of A-, F-, and V-ATPase. It is given in Fig. 10. Phylogenetic trees used to obtain ancestral PL h1 for each type of ATPase appear in Figs. 11, 13, 14.

Fig. 10 gives the LCA PL h1 sequence obtained by ‘majority rule’ among conserved amino acids for PL h1 sequences (listed below) at the root of A-, F- and V-ATPase tree (Figs. 11, 13 and 14).

Ancestral proteolipid h1 - A-type:

F-type:

V-type:

Between 15 and 17 sites (68.2–77.3 per cent) were uniquely defined in these sequences.
Conserved amino acids are underlined. The LCA PL h1 residue sequence displayed high internal homology (Fig. 10). Nine of 11 residues were identical between duplicate segments (internal identity index, 81.8 per cent). Duplication and fusion of the ancestral PL gene thus occurred at some time prior to appearance of the LCA.

Fig. 16 depicts the evolution of an 11-residue pre-duplication antecedent of PL h1. Pro-proteolipid h1-[7.11] has a hydrophobic 11-residue sequence with amino acids from a stage-7 code. Its sequence minimized the number of mutations required to produce ancestral PL h1 for F-, A- and V-ATPase. Tandem duplication of the pro-PL h1-[7.11] gene yields duplicate 11-residue segments, pro-PL h1-[7.22]. An Ile:Leu and Leu:Ile substitution at site-12 and -20, respectively, in this protein produced LCA PL h1 (Fig. 10). Sixteen sites (72.7 per cent) were resolved to a single amino acid in LCA PL h1. It was the consensus sequence among the most parsimonious sequence at the root of the F-, A- and V-ATPase PL h1 tree (Figs. 11, 13 and 14). Five sites in LCA PL h1 were constrained to two amino acids, while site-11 was resolved to three amino acids (Fig. 10). Fig. 16 shows ancestral V-type PL h1 evolved from A-type PL h1. The path to ancestral V-ATPase PL included five mutations and a tandem duplication. A double hairpin 16 kDa PL molecule thereby results (Muller et al., 1999).

Table 2 summarizes mutation distances separating LCA and ancestral F-, A- and V-type PL h1 sequences. LCA PL h1 is positioned between the F-type sequence, on one side, and A- and V-type sequences, on the other. This places the first node in the tree of life between Bacteria and Archaea, and the second between Eukarya and Archaea. Other genes gave a similar position for the root (Gogarten et al., 1989; Iwabe et al., 1989; Brown and Doolittle, 1995; Baldauf et al., 1996; Lawson et al., 1996; Gribaldo and Cammarano, 1998). Despite this consensus, Forterre and colleagues caution that locating the root between prokaryote domains might reflect long branch attraction (Philippe and Forterre, 1999; Lopez et al., 1999). Five residue changes occur on the path from LCA PL h1 to ancestral F-type PL h1. A single change separates ancestral A-ATPase PL h1 from its antecedent. This places Archaea nearer the root, indicative of a slow rate of evolution. This is in accord with the assessment of Kollman and Doolittle (2000). Consistent with the structural relationship established between catalytic and regulator subunits of extant ATPases (Gogarten and Taiz, 1992), ancestral V-type PL h1 exhibits closer homology to its A-type counterpart, than an F-type sequence (Table 2).

Table 2
Separation between ancestral sequences of proteolipid helix-1 in F-, A- and V-ATPase and last common ancestor*

<table>
<thead>
<tr>
<th>Source</th>
<th>F</th>
<th>LCA</th>
<th>A</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>0</td>
<td>5</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>LCA</td>
<td>0.08</td>
<td>0</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>A</td>
<td>0.1</td>
<td>0.02</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>V</td>
<td>0.17</td>
<td>0.1</td>
<td>0.08</td>
<td>0</td>
</tr>
</tbody>
</table>

*Upper right half of table lists the minimum number of base differences between their templates. Lower left half gives the corresponding Jukes-Cantor distance. From results in Fig. 16.
Proteolipid h1 has homology with h2, giving added interest to PL evolution. Each ancestral helix-2 sequence was resolved by the method of maximum parsimony.

Ancestral PL h2 - A-type:

```
LI FAVLGEGLAIYGLLIALILL
```

F-type:

```
LL ALALAE ALAIYGLLIALILL
```

V-type:

```
LL VLGLAE VAIFGLALLILI
```

Between 59 and 68 per cent of helix-2 sites were uniquely defined. This is slightly lower than for helix-1. Four residues repeat between 11-residue segments in LCA PL h2, giving an internal identity index of 36.4 per cent (Fig. 10); Gly at site-3 and -14 are included here, where the former site is resolved to \{G,A,V,F\} and the latter to G. Three residues conserved by helix-2 repeats were identical with equivalent residues conserved by helix-1 repeats. Here, residue \( n \) in helix-1 was matched with \( n + 1 \) in helix-2. The observed helix-1 and -2 identity had a probability of less than \( 1 \times 10^{-3} \); no match occurred in 1000 pairwise comparisons between two sets of randomly generated sequences. They were obtained on random resampling (with replacement) of the amino acid frequency distribution for each helix. Further evidence of homology between both helices is provided by a jointly elevated internal identity index at span-11 (Fig. 15).

The motif, - Gly - - Ala - - Leu -, occurs in each repeat within helix-1 and -2. This indicates they evolved from a common 11-residue antecedent. Successive tandem duplications subsequently produced the hairpin-shaped PL molecule, according to present evidence. Fig. 17 depicts formation of the PL molecule through three duplication + fusion events. An additional duplication + fusion is depicted which traces the origin of PL back to the gene for a hydrophobic tetrapeptide, GLAV\(_A\). Existence of a low-order periodicity detected in helix-1 supports this step (Fig. 15). Early structural constraints, as noted, could also account for this feature of PL primary structure. Tandem duplication of the gene for a 22-residue protein, together with insertion of a central polar segment, yields the standard hairpin PL structure. Formation of the 8 kDa PL molecule occurred before species divergence (Fig. 10). This is consistent with the LCA containing a functional H\(^{+}\)-ATPase (Gogarten and Taiz, 1992; Muller et al., 1999). Since the linker sequence between helix-1 and -2 conserves basic amino acid residues in both F- and A-type PL (from prokaryote domains), the PL hairpin molecule appeared late in code evolution. Positively charged residues in the linker sequence apparently bind PL to ATPase catalytic subunits, positioned in the aqueous phase external to a phospholipid bilayer (Hilario and Gogarten, 1998). Evolution of the
ATPase catalytic subunits, in the pre-divergence era (Gogarten and Taiz, 1992), parallels PL evolution outlined here. Fig. 18 compares amino acid frequencies representative of all PL sequences examined (Fig. 10) with their frequency in the 11 residue pre-duplication antecedent, pro-PL h1-[7.11]. Both frequency distributions are arranged according to the biosynthetic order of amino acid addition to the genetic code (Davis, 1999a). They exhibit close agreement and a goodness of fit test revealed no significant difference between them; $D_{\text{max}}(n : 11, 2354) = 0.08 \text{NS}$, $\chi^2 = 0.28, 0.80 < P < 0.9$, in a Kolmogorov–Smirnov test. Alanine$^4$, Gly$^5$, Leu$^7$, Ile$^7$ and Val$^4$ were the only amino acids contained by pro-PL h1-[7.11]. In addition, they are the five most frequent amino acids in PL h1 from extant sources (Fig. 10), accounting for 83.0 per cent of its residues. All these amino acids were from the expansion phase (stage-4 to -7) of code evolution. Pro-proteolipid h1-[7] consequently formed no earlier than the stage-7 code. Consistent with formation before the post-expansion phase (stage-9 to -14), pro-PL h1-[7] lacked aromatic amino acids.

Amino acid frequencies in the 11-residue antecedent, pro-PL h1-[7.11] (Fig. 16), are distributed in the order of amino acid addition to the code in Fig. 19. They were incorporated between stage-4 and -7 of code evolution (Fig. 18) and have optimum goodness of fit with the codon array at stage 7.0±0.83. Goodness of fit assessments at each stage of code evolution produced a peak
Fig. 18. Comparison between amino acid frequencies before and after formation of an internal repeat in PL h1 sequence. Pre-duplication frequencies are from an early 11-residue protein, pro-PL h1-[7]. Post-duplication frequencies reflect the amino acid composition in 107 PL h1 sequences examined in this survey. A Kolmogorov–Smirnov test for goodness of fit established no significant difference between both sets of frequencies; $D_{\text{max}} = 0.079$ NS, $\chi^2 = 0.279$ ($P = 0.87$) for $n = 11$ (11 sites $\times$ 1 sequence) and 2354 (22 sites $\times$ 107 sequences) for pre- and post-duplication distributions, respectively. All amino acids in pro-PL h1-[7] are from an expansion phase genetic code (stage-4 to -7). Consistent with the evolutionary stability of PL h1, expansion phase amino acids accounted for 93 per cent of residues in the 107 sequences examined. Amino acid frequencies for PL h1 and its pre-duplication antecedent, pro-PL h1-[7], were obtained from sequences in Figs. 10 and 16.

Fig. 19. Frequency of amino acids with a specified stage of entry into the genetic code among pro-PL h1-[7] residues vs. the frequency of codons assigned at each stage in a stage-7 code. Pro-PL h1-[7] amino acids are shown to have entered the code between stage-4 and -7 (expansion phase). Amino acid and codon profiles broadly agree over this interval, except for omission of stage-2 amino acids. Unlike expansion phase amino acids, stage-2 amino acids are hydrophilic. Hence, they are excluded from this membrane protein, although they account for 21 per cent (12/56) of codons in the stage-7 code. Goodness of fit is shown between the pro-PL h1-[7] amino acid distribution and codon distributions between stage-4 and -14 (omitting stage-2 codons). An optimum exists at stage $7.0 \pm 0.83$. The pro-PL h1-[7] amino acid profile was obtained from Fig. 16. Codon profiles at each stage of code evolution were deduced from previous results (Davis, 1999a).
probability at stage 7 of 0.865; no stage-2 amino acids occur in pro-PL h1 and they were omitted from these calculations. In a stage-7 code, 21.4 per cent (12 of 56) of assigned codons were specific for stage-2 amino acids and their omission from an 11-residue peptide chain by chance had a probability of only 0.07. For a 22-residue chain, this probability reduces to $4.96 \times 10^{-3}$. Strong selection forces evidently promoted formation of this highly hydrophobic protein. Formation of pro-PL h1-[7.11] by a stage-14 code, on the other hand, had a probability of formation ($\phi_{c}^{14} \approx \phi_{c}^{\text{std}}$, Eq. (4)) of only 0.092 ($7.22 \times 10^{-5}/7.82 \times 10^{-4}$).

According to these findings pro-PL h1-[7.11] formed no earlier than stage 7, but well before stage 14 of code evolution. Incorporation of basic and aromatic amino acids after stage 9, together with exclusion of negatively charged residues from pro-PL [7.11], combine to suggest replacement of the primordial surface system by membrane-bound cells occurred between stage-7 and -9.

### 4. Origin of cell division and the DNA genome

The time of origin of proteins involved in cell division and DNA synthesis, transcription and replication is appraised in this section. These estimates are based on the ‘code age’ (stage of code evolution) of the residue profile at conserved sites within consensus sequences representative of each protein in different species domains. Results obtained using this approach were validated by comparison with earlier estimates from the phylogenetically determined Fdxn and PL ancestral residue sequence. Cell division and some DNA related proteins could be traced to an antecedent in the post-expansion phase of code evolution. The collection of ‘code ages’ for each set of conserved amino acids in the 10 ancient proteins examined made a statistical evaluation possible of an advancement in ‘code age’ among the set of amino acids at non-conserved sites. Results obtained in this analysis corroborate the biosynthetic order of amino acid addition of the genetic code.

Residue sequences for 10 pre-divergence proteins appear in Fig. 20. They are representative of an indicated number of homologues from each species domain, identified in a database search. Each sequence contains the most frequent amino acid per site within a protein from a specified domain. Sites with an identical amino acid in all species domains are highlighted. These strongly conserved amino acids were used to characterize the LCA residue sequence. The accuracy of this procedure was established, by comparing the LCA sequence obtained for Fdxn and H$^{+}$-ATPase PL h1 with the sequence obtained through phylogenetic analysis in Sections 2 and 3. Its computational efficiency in extracting phylogenetic information, compared with standard phylogenetic methods, facilitated investigation of several pre-divergence proteins opening the way for a comparative assessment of shifts in ‘code age’ among amino acid sets.

Ferredoxin and PL h1 sequences in Fig. 20 are based on amino acid frequencies in sequences listed by Figs. 1 and 10. The search for homologues of *C. pasteurianum* Fdxn, as noted earlier, revealed related residue sequences from 22 archaea and 60 bacteria sources. Fifty of 55 sites were resolved to a single amino acid residue in archaea and bacteria consensus sequences (Fig. 20). Twenty-nine of them were identical between both prokaryote domains (identity index, 58 per cent—29/50). Significantly, 27 of 29 conserved amino acids were identical with amino acids in the
LCA Fdxn sequence identified phylogenetically (Fig. 1). The LCA Fdxn sequence reconstructed from amino acids at strongly conserved sites, in consensus sequence analysis, thus had 93 per cent (27/29) accuracy with respect to phylogenetic analysis. Amino acids at nearly half of LCA Fdxn residue sites (27 of 55 sites) were therefore shown to be strongly conserved. By comparison, phylogenetic analysis resolved 32 sites to a single residue, 15 sites to two residues and 8 sites to three or more (Fig. 1). Consensus sequence analysis, consequently, attained a comparative precision of 73 per cent, or (29/(32+(15/2))). These findings establish that conserved sites in consensus sequences from different species domains accurately identified amino acids at almost three-quarters of the most parsimonious LCA protein sequence.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Domain No. of sources</th>
<th>Ref. source</th>
<th>Sequence placement</th>
<th>N-terminus</th>
<th>Most frequent residue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ferredoxin</td>
<td>B</td>
<td>60</td>
<td>1</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proteolipid h1</td>
<td>B</td>
<td>55</td>
<td>37</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FtsZ</td>
<td>B</td>
<td>87</td>
<td>51</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEN-1-Pol-I</td>
<td>B</td>
<td>92</td>
<td>201</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNAP-α(N)</td>
<td>B</td>
<td>60</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNAP-γ</td>
<td>B</td>
<td>76</td>
<td>131</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>52</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 20.** Strongly conserved amino acids in consensus sequences of archaic proteins in Bacteria, Archaea and Eukarya. They include cell division protein, FtsZ, and proteins involved in DNA synthesis, replication and transcription. Sequences in this analysis were identified in a database search for homologues of the reference sequence. Biological source and access number of each reference sequence is given, together with the location (from N-terminus of reference sequence) of the first residue in each segment within the reference protein. FEN-1-Pol-I sequences include homologues of Pol-I (5'-3' exonuclease domain of bacterial DNA polymerase I) from *Rickettsia prowazekii* (access no., C71638). This protein was also a homologue of the reference sequence listed. RNA Polymerase-α(N) sequences include homologues of RNAP-D from archaeon *Halobacterium salinarum* (access no., T43940), a homologue of the N-terminal domain of RNAP-α from *Bacillus subtilis*. Single letter designations for amino acids and enzyme abbreviations are given in the text. Sites resolved to a single residue that had an identical amino acid in each domain consensus sequence are highlighted. Where two amino acids have highest frequency at a site, both are listed. - , indicates a most prevalent amino acid could not be determined. _, signifies a gap in consensus sequences. Ferredoxin and PL h1 consensus sequences were derived from sequences in Figs. 1 and 10.
Proteolipid h1 in *S. acidocaldarius* ATPase displayed homology with PL h1 from 55 bacteria, 41 eukaryote and 10 other archaea sources (Fig. 10). Amino acids were identical at nine of 22 sites among the PL h1 consensus sequence for each species domain (Fig. 20). All nine strongly conserved amino acid residues were identical to amino acids in the phylogenetically determined LCA PL h1 sequence (Fig. 10). An accuracy of 100 per cent therefore arose. Forty-one per cent (9/22) of PL h1 residue sites were strongly conserved between consensus sequences; it may be noted, 10 of the remaining 13 residue sites were identical in two of three domains. Phylogenetic methods resolved 15 residue sites in LCA PL h1 to a single amino acid, 5 sites to two amino acids, and 2 sites to not less than 3 amino acids. The comparative precision of the consensus sequence method was therefore 51 per cent, or (9/(15+(5/2))).

These results demonstrate strongly conserved amino acids, in consensus sequence analysis, had virtually complete accuracy and they identified half to three-quarters of sites resolved by phylogenetic analysis. The computational efficiency of consensus sequence analysis is thus achieved at virtually no cost in accuracy. On the other hand, precision was reduced between one quarter to one half. A loss in precision of this magnitude means consensus sequence analysis would not be the method of choice for sequence reconstruction, as undertaken in Sections 2 and 3. Its accuracy insures, however, that the set of amino acids identified reliably reflect the ‘code age’ of an ancient residue sequence. Agreement between the ‘code age’ obtained by both methods for the residue profile of LCA Fdxn and PL verifies this conclusion. Specifically, phylogenetic (Fig. 6) and consensus sequence analysis (cf. Table 3) yield a ‘code age’ for the residue profile of LCA

<table>
<thead>
<tr>
<th>Protein</th>
<th>Domain No. of sources</th>
<th>Reference</th>
<th>Sequence alignment</th>
<th>10</th>
<th>20</th>
<th>Most frequent residue</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENAP g-</td>
<td>B</td>
<td>72</td>
<td>E. coli [2001/2004/2006]</td>
<td>246</td>
<td>22</td>
<td>LGLGKRMKEMQSSYDS</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>14</td>
<td></td>
<td>372</td>
<td>45</td>
<td>LGLGKRMKEMQSSYDS</td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>13</td>
<td></td>
<td>469</td>
<td>91</td>
<td>LGLGKRMKEMQSSYDS</td>
</tr>
</tbody>
</table>

![Fig. 20 (continued).]
Table 3
Stage of code evolution attributed to the residue profile in conserved sequences of proteins that originated before the divergence of species domains

<table>
<thead>
<tr>
<th>Protein</th>
<th>No. sequences</th>
<th>Bacteria</th>
<th>Archaea</th>
<th>Eukarya</th>
<th>Total no. sites aligned&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Identity between domains&lt;sup&gt;b&lt;/sup&gt; (per cent)</th>
<th>Identity between random sequences (per cent)</th>
<th>Probability of difference in identity&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Conserved residues</th>
<th>Non-conserved residues</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ferredoxin</td>
<td>60</td>
<td>22</td>
<td>0</td>
<td>55</td>
<td>58.0***</td>
<td>10.8</td>
<td>1.87 × 10&lt;sup&gt;−8&lt;/sup&gt;</td>
<td>29</td>
<td>7.0±0.53</td>
<td>7</td>
</tr>
<tr>
<td>Proteolipid-helix 1</td>
<td>55</td>
<td>11</td>
<td>41</td>
<td>22</td>
<td>47.7***</td>
<td>13.6</td>
<td>5.35 × 10&lt;sup&gt;−3&lt;/sup&gt;</td>
<td>9</td>
<td>7.1±0.67</td>
<td>7</td>
</tr>
<tr>
<td>FtsZ</td>
<td>87</td>
<td>15</td>
<td>0</td>
<td>56</td>
<td>74.1***</td>
<td>31.6</td>
<td>3.65 × 10&lt;sup&gt;−6&lt;/sup&gt;</td>
<td>40</td>
<td>7.5±0.59</td>
<td>11</td>
</tr>
<tr>
<td>FEN-1/Pol I</td>
<td>52</td>
<td>7</td>
<td>9</td>
<td>54</td>
<td>22.2*</td>
<td>4.7</td>
<td>3.66 × 10&lt;sup&gt;−2&lt;/sup&gt;</td>
<td>7</td>
<td>9.1±1.39</td>
<td>10</td>
</tr>
<tr>
<td>RNAP-α(N)</td>
<td>76</td>
<td>11</td>
<td>12</td>
<td>221</td>
<td>12.4*</td>
<td>0</td>
<td>4.23 × 10&lt;sup&gt;−2&lt;/sup&gt;</td>
<td>22</td>
<td>9.3±1.39</td>
<td>11</td>
</tr>
<tr>
<td>RNAP-β</td>
<td>76</td>
<td>10</td>
<td>52</td>
<td>238</td>
<td>35.8***</td>
<td>1.4</td>
<td>1.83 × 10&lt;sup&gt;−8&lt;/sup&gt;</td>
<td>81</td>
<td>10.9±0.85</td>
<td>13</td>
</tr>
<tr>
<td>RNAP-β&lt;sup&gt;′&lt;/sup&gt;</td>
<td>72</td>
<td>14</td>
<td>33</td>
<td>228</td>
<td>39.1***</td>
<td>1.4</td>
<td>1.59 × 10&lt;sup&gt;−9&lt;/sup&gt;</td>
<td>84</td>
<td>10.9±0.98</td>
<td>13</td>
</tr>
<tr>
<td>Ribonucleotide reductase</td>
<td>39</td>
<td>9</td>
<td>19</td>
<td>230</td>
<td>28.8***</td>
<td>0</td>
<td>6.04 × 10&lt;sup&gt;−5&lt;/sup&gt;</td>
<td>46</td>
<td>11.0±0.95</td>
<td>14</td>
</tr>
<tr>
<td>Reverse transcriptase</td>
<td>12</td>
<td>0</td>
<td>11</td>
<td>155</td>
<td>40.2*</td>
<td>16.1</td>
<td>2.89 × 10&lt;sup&gt;−2&lt;/sup&gt;</td>
<td>37</td>
<td>13,14</td>
<td>11</td>
</tr>
<tr>
<td>Topoisomerase I</td>
<td>26</td>
<td>9</td>
<td>4</td>
<td>116</td>
<td>40.6***</td>
<td>0</td>
<td>6.54 × 10&lt;sup&gt;−3&lt;/sup&gt;</td>
<td>28</td>
<td>14</td>
<td>11</td>
</tr>
</tbody>
</table>

<sup>a</sup> Most frequent residue at each of these sites and their position in a reference protein are given in Fig. 20.

<sup>b</sup> Asterisks indicate significance of difference in identity, at unambiguous sites, between observed and random residue sequences, assessed by a χ² test; *, 0.05 ≥ P > 0.025; ***, 0.01 > P. The total number of aligned sites less number of conserved + non-conserved sites corresponds to the number of sites with an ambiguous residue assignment.

<sup>c</sup> Probabilities were evaluated using χ²; corrected for continuity, at one degree of freedom.

<sup>d</sup>The stage of code evolution that best fit residue profiles in each sequence was established by a Kolmogorov–Smirnov goodness of fit test. Conserved residues best fit a significantly lower stage than non-conserved residues, in a Mann–Whitney U-test; U = 75.5 ***, P = 1.30 × 10<sup>−2</sup>. 
Fdxn of 7.0 ± 0.37 and 7.0 ± 0.53, respectively. While they yield a ‘code age’ for the residue profile of LCA PL h1 of 6.9 ± 0.43 and 7.1 ± 0.67, respectively (Fig. 16, cf. Table 3). For all proteins considered, the number of strongly conserved sites significantly exceeded the number that could occur by chance (cf. Table 3).

Cell division protein, FtsZ (filamenting temperature-sensitive mutant Z protein), occurs in Archaea, Bacteria and chloroplasts (Margolin et al., 1996; Erickson et al., 1996). It has some sequence homology with tubulin, and X-ray crystallographs reveal structural similarities between both cytokinetic proteins (Lowe and Amos, 1998; Desai and Mitchison, 1998), raising the possibility that the prokaryote protein is a progenitor of tubulin in eukaryotes. Four conserved segments within FtsZ, containing 87 residue sites, were examined in 15 archaea and 87 bacteria sequences (Fig. 20). They included base and sugar recognition regions h1, s4/h3, s5/h4 and h5 in the N-terminal domain. Strand4/helix3 includes turn, t4, containing the ‘tubulin’ motif, GGGTGTG, which forms a loop that contacts the phosphates of a GTP molecule hydrolyzed by FtsZ (Lowe and Amos, 1998). Homologues of FtsZ from the hyperthermophilic archaeon, *Archaeoglobus fulgidus*, were identified in a search involving gapped-BLAST and PSI-BLAST algorithms (Altschul et al., 1997). FtsZ sequences in Fig. 20 contain the most frequent amino acid at each residue site among archaeal and bacteria homologues of the reference protein. Sites with identical amino acids in these sequences are highlighted. Amino acids conserved within FtsZ consensus sequences for different domains were attributed to LCA FtsZ.

Deoxyribonucleotides derive from ribonucleotides in a reaction catalyzed by one of four classes of RNR (Stubbe and van der Donck, 1998). Although these enzymes differ in distribution, type of cofactor, substrate phosphorylation and O2 sensitivity, each utilizes a basically similar reaction path, involving radical formation, and allosteric regulation of substrate specificity (Reichard, 1993). Sequence homology between distinct classes of the enzyme in species from different domains (Riera et al., 1997; Tauer and Benner, 1997) strengthens the view that RNR had a common ancestor. Amino acid frequencies were appraised at 230 sites, within two segments of this enzyme (residue, 140–(188)–192, and 223–(319)–401). ‘Majority rule’ residue sequences are given in Fig. 20. These sequences are representative of homologues of the class I (O2-dependent Fe cofactor) reductase of *E. coli* from 39 bacteria, 9 archaea and 19 eukaryote sources. Four of the 67 proteins were class II reductases (adenosylcobalamin cofactor). They included RNR from thermophilic archaeons *P. furiosus* (identity, 22 per cent) and *Thermoplasma acidophilum* (20 per cent) and bacteria *T. maritima* (23 per cent), also a thermophile, and *Streptomyces clavuligerus* (27 per cent).

Sequence homology in DNA dependent RNAP core subunits (α, β, β′) from archaea, bacteria and eukaryote sources (Gropp et al., 1986; Sweetser et al., 1987; Langer et al., 1995) shows them to be pre-divergence proteins (Benner et al., 1989; Leipe et al., 1999). RNAP core enzyme, α2ββ′, self-assembles, beginning with dimerization of the α subunit. The catalytic subunits, β and β′ (Jin and Zhou, 1996), subsequently bind to α2 sites located in the N-terminal region. This region spans two-thirds of the α residue sequence in bacteria (Kimura and Ishihama, 1995). Consensus sequences for the RNAP subunits are presented in Fig. 20, from sources in different species domains. α-subunit consensus sequences contain 222 sites from four segments (residue, 6–26, 31–59, 64–68, 71–235) of the *Bacillus subtilis* protein. They are representative of sequences from 76 bacteria, 12 archaea and 11 eukaryote sources that were identified in a search for homologues of the *B. subtilis* α subunit. Included among these proteins was the D subunit of *Halobacterium*
salarinum RNAP. It served as a reference sequence in a second database scan that enlarged the number archaea sequences investigated.

β and β' sequences in Fig. 20 show the most prevalent amino acids at 238 and 230 residue sites, respectively, for each species domain. A search revealed 76 bacteria, 10 archaea and 52 eukaryote sequences with significant homology to the β subunit of *E. coli* RNAP. Seventy-two bacteria, 14 archaea and 33 eukaryote homologues of its β subunit were identified. The reference sequence for β combined residue sites from four segments (residue, 793–828, 859–889, 1047–1118, 1198–1296) and the β' sequence combined five segments (246–253, 322–379, 419–508, 780–826a,b–838, 912–924). Note that two amino acids were inserted (826a and 826b) into the β' consensus sequence. This corrected a 2-residue deletion, between residues 826/827, in the *E. coli* β' subunit, used as reference sequence in the search.

RNA dependent DNA synthesis occurs in both eukaryotes and prokaryotes (Lim and Maas, 1989; Inouye and Inouye, 1991). A phylogenetic analysis of RT sequences from both sources, and organelle plasmids and diverse groups of viruses revealed an early bifurcation in the evolution of this enzyme (Xiong and Eickbush, 1990). Retroposons with long terminal repeats (LTR) are associated with one branch, while non-LTR retroposons and bacterial multicopy single-strand DNA (msDNA) are associated with the other branch. Ancestral RT was closer to the non-LTR enzyme, when RNA dependent RNAP was used as an outgroup to root the RT tree. Consensus sequences representative of eukaryote and bacteria RT sequences with homology to the msEc67 enzyme from *E. coli* strain B are presented in Fig. 20. The reference sequence combined six segments (108–143, 170–203, 215–249, 256–272, 283–308, 312–322) with 159 residue sites. Eighty-seven sites were resolved to a single amino acid, in both bacteria and eukaryote consensus sequences. Identical residues occurred at 37 of these sites.

RNAP and RT produce a DNA–RNA heteroduplex. Conversely, the accessory replication protein, FEN-1 (flap exonuclease), removes a DNA–RNA hybrid molecule. During lagging strand DNA synthesis, the structure-specific, 5'→3' exonuclease removes the Okazaki fragment. Consistent with evolution by direct descent from a gene in the LCA, FEN-1 retains detectable sequence and structural homology with proteins having similar specificity in phylogenetically diverse species (Edgell and Doolittle, 1997; Leipe et al., 1999). With FEN-1 from *P. furiosus* as reference, sequences in the 5'→3' exonuclease domain of 52 bacterial DNA polymerase-I molecules and in flap exonucleases of 7 archaeons and 9 eukaryotes (Fig. 20) were revealed to exhibit significant homology. They included bacterial DNA Pol-I homologues of *Rickettsia prowazekii* DNA Pol-I, whose 5'→3' exonuclease domain was homologous with *P. furiosus* FEN-1. Forty-six of 54 sites (residue 201–254 in *P. furiosus* FEN-1) were unambiguously specified. Amino acids at 10 sites, resolved to a single residue, were identical in consensus sequences representative of the exonuclease in Archaea, Bacteria and Eukarya.

DNA topoisomerase I is one of a group of topology-altering enzymes required for replication in Archaea, Bacteria and Eukarya. It is a monomeric protein containing 769 amino acids in the thermophilic archaeon *M. jannaschii*. Topoisomerase I from this source reduces the linking number of coiled DNA, after covalently bonding to the 5'-phosphate of a base at a break introduced into one strand of the duplex. Consensus sequences for Topo I of 26 bacteria, 9 archaea and 4 eukaryote with homology to *M. jannaschii* Topo-I are given in Fig. 20. The reference sequence contained 116 amino acid residues from five segments (residue, 12–20, 299–320, 600–630, 645–670, 680–707). It spans the active site Tyr11 (residue, 315) and three Zn2+...
fingers (600–630, 645–670, 680–708) (Kaltoum et al., 1997). Sixty-nine sites were resolved to a single amino acid, in all three consensus sequences. Amino acids at 28 resolved sites were identical.

Fig. 21 shows conserved amino acids distributed by their putative stage of entry into the genetic code, in eight of the foregoing proteins. They broadly form two groups: those that conserve amino acids up to stage 7, and those that include post-stage 7 amino acids. The latter participate in cell division and DNA synthesis, replication and transcription. Their residue profile reveals these protein sequences formed after stage 7 of code evolution. By contrast, conserved sites in Fdxn and

<table>
<thead>
<tr>
<th>Residue</th>
<th>Ferredoxin</th>
<th>Proteolipid h1</th>
<th>FtsZ</th>
<th>FEN-1</th>
<th>RNAP-β'</th>
<th>Reverse transcriptase</th>
<th>Topoisomerase-I</th>
<th>Ribonucleotide reductase (Fe)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glu</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Asp</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ala</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ile</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pro</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Asn</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ser</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cys</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Leu</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Val</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gln</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gly</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Thr</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Met</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Arg</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lys</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tyr</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>His</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Trp</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig. 21. Frequency of conserved amino acids, from a specified stage of code evolution, among archaic proteins. Ferredoxin and PL h1 do not conserve amino acids beyond stage-7. More advanced amino acids were conserved by proteins involved in cell septation and DNA synthesis, replication and transcription. The former also have a comparatively high frequency of early amino acids. Exclusion of stage-2 amino acids from PL h1 can be attributed to selection against polar amino acids in membrane proteins. At top, amino acids appear in groups from the same stage of code evolution. Amino acid frequencies were derived from consensus sequences in Fig. 20 and the stage of their addition to the genetic code corresponds to the biosynthetic order of amino acid addition to the code (Davis, 1999a).
PL h1 lack post-stage 7 amino acids. This is consistent with a time of origin, as early as, stage 7 (Sections 2 and 3). They appear from this to predate the former group of proteins.

Amino acid frequencies in extant proteins correlate with the number of codons assigned to them in the standard code (Osawa, 1995; Davis, 1999a). Generalizing this relationship to the pre-divergence era implies amino acid frequencies in an ancient residue sequence will correlate with codon assignments at the stage of code evolution when it formed. Fig. 22 shows that the distribution of strongly conserved amino acids in FtsZ and a number of replication proteins had an optimum goodness of fit with codon assignments at stage 8–11 in code evolution. Strongly conserved amino acids of a specified ‘code age’ in archaic proteins and amino acid frequency based on codon assignments at each stage of code evolution. Ferredoxin and PL h1 have optimal goodness of fit at stage 7. Prokaryote septation protein FtsZ, 5′ → 3′ exonuclease FEN-1, RNR, and RNAP subunits α, β and β′ show a peak in goodness of fit between stage 8 and 11. Topoisomerase I (Topo I) and reverse transcriptase (RT) had optimal goodness of fit at stage 13 or 14. Curves for all 10 proteins in this study are depicted in (a) and (b). A broken line is used to distinguish the Fdxn curve in (a). Goodness of fit between amino acid and codon frequency distributions were evaluated by a Kolmogorov–Smirnov test. Protein designations appear under Abbreviations. Amino acid frequencies were derived from sequences in Fig. 20. Codon frequencies at specified stages of code evolution were from Davis (1999a).
conserved amino acids in Fdxn and PL h1 exhibit optimal goodness of fit with a stage 7 codon array, consistent with evolving before proteins implicated in cell division and DNA formation. This conforms with an earlier ‘code age’ for the most advanced amino acid conserved by Fdxn and PL h1 (Fig. 21).

Consensus sequences for each protein contained significantly more strongly conserved sites than in randomized sequences. Background identity levels in Table 3 refer to homology among consensus sequences representative of an equivalent set of residue sequences, bootstrapped from the amino acid frequency distribution characterizing each site within proteins in each species domain. Differences between the number of conserved amino acids and background identity levels were statistically evaluated from $\chi^2(\text{df} = 1)$ values in contingency tables, corrected for continuity. $\chi^2$ values obtained corresponded to probabilities (Table 3) from $4.23 \times 10^{-2}$ (RNAP-α, N-domain) to $1.59 \times 10^{-9}$ (RNAP β′). This effectively excludes chance as the source of identity obtained between proteins from different species domains.

Strongly conserved amino acids in Fdxn and PL h1 optimally matched the codon assignment pattern at stage $7.0 \pm 0.53$ and $7.1 \pm 0.67$ ($\bar{X} \pm \text{SEM}$), respectively (Table 3). The mean and standard error of each estimate was determined from peak position and width in Fig. 22. In this connection, $\text{SEM} = (\text{half-width at } p = 0.606 \times p_{\text{max}})/\sqrt{n}$, where $n$ is the number of sample points in a peak with maximum probability $p_{\text{max}}$. As noted, the residue profile in both proteins had a cut-off at stage 7, since they conserved amino acids up to the ‘code age’ of Ile$^7$ in Fdxn and Leu$^7$ in PL h1 (Fig. 20). A stage-7 cut-off occurred also in the phylogenetically determined LCA PL h1 residue sequence (Fig. 10). Six of 22 amino acids (27 per cent) attributed to LCA PL h1 were stage 7 additions to the code; they included four Leu$^7$ and two Ile$^7$. Phylogenetically determined LCA Fdxn contained three stage-10 amino acids (3 Lys$^{10}$), in addition to eight stage-7 amino acids (6 Ile$^7$, 2 Leu$^7$), among 55 residues (Fig. 1). Two Lys$^{10}$ residues were separated by 29 residues (Fig. 10), indicating the pre-duplication pro-Fdxn gene encoded Lys$^{10}$ (Fig. 16) some time before the LCA evolved. The consensus sequence for archaea Fdxn contained Lys$^{10}$ at site-7 and -36, as in phylogenetically determined LCA Fdxn. However, the bacteria Fdxn consensus sequence did not (Fig. 20) and this led to omission of Lys$^{10}$ from strongly conserved amino acids in Fdxn.

Cell septation and replication proteins had a later ‘time of origin’ than Fdxn and PL h1. Conserved residues in FtsZ displayed an optimal goodness of fit with the codon distribution at stage $7.5 \pm 0.59$ in code evolution (Table 3). A stage-11 amino acid (Phe$^{11}$) was most advanced among its strongly conserved residues. Aromatic and basic residues occurred at 5 of 40 conserved sites (3 Phe$^{11}$, 2 Arg$^9$) in FtsZ (Fig. 20), consistent with a post-stage 7 origin.

Core subunits in RNAP (α, β, β′) are considered to be pre-divergence proteins, as noted previously, since they show homology between phylogenetically diverse sources (Gropp et al., 1986; Sweetser et al., 1987; Langer et al., 1995). Results in Table 3 now place their origin in the late stages of code evolution. The frequency distribution of strongly conserved amino acids in α, β and β′ subunits optimally matched codon assignments at stage $9.3 \pm 0.6$ to $10.9 \pm 0.9$ of code evolution. By comparison, non-conserved amino acids in these sequences form distributions with maximum goodness of fit between stage $11.1 \pm 0.6$ and 14. Consistent with this result, residues at non-conserved sites in all three subunits included Trp$^{14}$, whose synthesis is longest (14 steps) and ‘code age’ (stage 14) greatest among amino acids incorporated into protein. By contrast, Trp$^{14}$ did not occur at any site conserved by RNAP core subunits.
An early origin of RNAP implies an equally early pathway for deoxyribonucleotide synthesis. Results in Table 3 show 46 of 230 sites examined were strongly conserved in RNR. Amino acids at conserved sites matched the codon assignment pattern at stage 11.0±0.9. By comparison, amino acids at 114 non-conserved sites matched codon assignments at stage 13 to 14 of code evolution. Ribonucleotide reductase thus has a possible time of origin at the upper end of the range (stage 9.3±0.6–10.9±0.9) obtained for RNAP subunits. Homologues of the reference protein (E. coli, class I RNR) included class I and II reductases. Class I RNR is best characterized, and, as shown, it preserves evidence of a pre-divergence origin. It appears likely, however, that the ancestral enzyme had closer homology with the O2-sensitive, class III (Fe-S cofactor, glycyl radical) RNR (Reichard, 1993).

RNA dependent DNA polymerization catalyzed by RT seems necessary in the transition to a DNA genome (Darnell and Doolittle, 1986; Lazcano et al., 1992). The discovery that RT exists in bacteria and eukaryotes (Lim and Maas, 1989; Inouye and Inouye, 1991), and shows sequence homology between species domains (Xiong and Eickbush, 1990), strengthened the view that this enzyme was significant in the pre-divergence era. Table 3 shows that amino acids were conserved at 37 of 155 sites in consensus sequences representative of 12 bacteria and 11 eukaryote RT molecules identified with msEc67 RT as reference (Fig. 20). Conserved amino acids did not exceed stage 11 in ‘code age’ (residue 259, Tyr11). Their frequency distribution was more advanced, however. It displayed optimal goodness of fit with a stage 13 or 14 codon pattern. As might be expected, a similar residue profile characterized the 52 non-conserved sites in each RT consensus sequence. Existence of significant identity between RT from prokaryote and eukaryote sources suggests the enzyme originated in the pre-divergence era. The advanced ‘code age’ of its residue profile, however, conforms with an origin late in the pre-divergence era when the genetic code was essentially complete.

Results on RNAP core subunits provided evidence for synthesis of DNA–RNA hybrid molecules when the genetic code was still developing. Additional evidence comes from analysis of the FEN-1 residue profile. Table 3 shows that amino acids were identical at 10 of 45 sites, resolved to a single amino acid, in the FEN-1/Pol I consensus sequence (length, 54 residues). Each consensus sequence was representative of homologues of P. furiosus FEN-1 and R. prowazekii Pol I (5’→3’ exonuclease domain). Fifty-two bacteria, 7 archaea and 9 eukaryote FEN-1 sequences exhibited significant homology. Strongly conserved amino acids had a stage-10 cut-off (Fig. 21). In addition, their frequency distribution optimally matched stage 9.1±1.39 codon assignments (Table 3); a large standard error accompanies this estimate reflecting the broad goodness of fit peak for FEN-1 in Fig. 22. An elevated number of stage-2 residues seemingly distorted the FEN-1 residue profile, giving rise to this spread. They included Asp2 clusters at the catalytic site, implicated in divalent cation (Zn, Mn) binding, and at a site behind its helical arch (Ceska et al., 1996), linked with product release. Amino acids at non-conserved sites had a stage 14 cut-off and an advanced (stage 13 or 14) residue profile.

Evidence of DNA synthesis, transcription and polymerization in the pre-divergence era raises the question whether the pre-divergence genome was sufficiently advanced to require a topoisomerase. The presence of Topo I in all three species domains (Wang, 1996) suggests the ancestral protein was present in the LCA. Consistent with this, a search for homologues of Topo-I-5’ from M. jannaschii identified Bacteria, Archaea and Eukarya topoisomerase sequences (Fig. 20). Forty-one sites were resolved to a single amino acid in the consensus sequence (length,
116 residues) in each species domain (Table 3). Identical residues occurred at 28 sites. Amino acids at strongly conserved sites did not exceed stage 11 in ‘code age’. However, they had an optimal goodness of fit with a stage-14 codon assignment pattern. Residues up to stage-14 occurred at non-conserved sites and they had a distribution of similar vintage. These findings indicate that Topo-I originated in the pre-divergence era, albeit when the genetic code was virtually complete.

This investigation is based on the proposition that amino acids at conserved sites portray a protein as it was at an earlier stage of its existence than amino acids at variable sites. Strongly conserved amino acids in a protein of pre-divergence origin, in principle, could have occupied a residue site for an interval extending far back in time, to when the code was still developing. Mutations accounting for species variations in an ancient protein would be anticipated to replace early additions to the code with amino acids from a later code, notably the standard code. Given that the biosynthetic order of amino acid addition to the code (Davis, 1999a) is valid, amino acids at variable sites could be anticipated to show an advanced ‘code age’ (longer biosynthesis path length) vs. amino acids at conserved sites in a pre-divergence protein.

Fig. 23 compares the ‘code age’ of the set of amino acids at conserved and non-conserved sites in 10 pre-divergence proteins from 844 phylogenetically diverse sources. In all proteins conserved

![Graph](image)

Fig. 23. Relation between ‘code age’ of residues at strongly conserved and variable sites within amino acid sequences for 10 archaic proteins. Conserved residue profiles were from a significantly earlier stage of code evolution than non-conserved profiles. The disparity in ‘code age’ decreased with advancement in conserved residue ‘code age’. A linear least-squares regression equation, depicted by the continuous line, consequently has a positive slope, \( N = 7.81 + 0.46C \), where \( N \) and \( C \) refer to the ‘code age’ of the non-conserved and conserved residue profile, respectively. Residue profile ‘code age’ corresponds to the stage of code evolution whose codon assignment pattern showed optimal goodness of fit. This graph is based on results in Table 3 and Fig. 20.
residues had a ‘code age’ earlier, or equal, to residues at non-conserved sites. Among the 10 proteins examined, seven had an earlier set of residues at conserved sites, while the remaining three proteins had residue sets of equal ‘code age’. The ‘code age’ of amino acids at conserved sites was significantly earlier than for non-conserved sites (Table 3); Mann–Whitney U-test, \(U = 75.5\), \(p = 1.3 \times 10^{-2}\).

The set of residues at non-conserved sites shows the expected advancement in ‘code age’. Their ‘ages’ consequently occupy the upper sector of a ‘non-conserved age’ vs. ‘conserved age’ plot (Fig. 23). A linear least-squares equation fitting ‘code age’ for non-conserved vs. conserved sets of residues,

\[ N = 7.9 + 0.5C, 7 \leq C \leq 14, N \leq 14 \]  

had a positive intercept, as required. \(N\) and \(C\) refer to the stage of code evolution (‘code age’) whose codon assignment pattern optimally matched the set of residues at non-conserved and conserved sites, respectively. A positive slope of less than 1 in Eq. (8) implies: (i) ‘code age’ increased among residues at non-conserved sites as the ‘code age’ of conserved residues advanced, and (ii) the difference in ‘code age’ diminished between non-conserved and conserved residues with the advance of ‘code age’ in conserved residues. Stage 14 is the upper limit of ‘code age’ in both conserved and non-conserved residues and this accounts for the diminishing difference between them with advances in conserved amino acids ‘code age’. The advanced ‘code age’ of amino acid residues at non-conserved sites within the pre-divergence proteins examined (Fig. 23), therefore, corroborates the biosynthetic order of amino acid addition to the genetic code (Davis, 1999a).

5. Discussion

5.1. Early ‘code age’ of residue profile at conserved sites

By using the biosynthetic order of amino acid addition to the genetic code as a reference, it has become possible to chronicle the development of early proteins. This provided a window on evolution in the interval preceding species divergence. On the other hand, the primary structure of the ancient proteins examined here furnished evidence substantiating the biosynthetic path of code evolution (Section 4).

Amino acids at strongly conserved sites within the 10 pre-divergence proteins in this study were demonstrated to have a significantly earlier ‘code age’ than the set of amino acids at variable sites (Section 4). This revealed their ancestral protein generally contained an early set of amino acids, formed on comparatively short paths measured from the citrate cycle. The difference in ‘code age’ arose as a more advanced set of amino acids were subsequently introduced at mutable sites. Expression of this time dependent change would have been obscured had amino acid sets been incorrectly, or randomly, defined. The primary structure of pre-divergence proteins was therefore revealed to: (i) retain evidence on the order of amino acid addition to the genetic code, and (ii) corroborate the biosynthetic order of amino acid addition to the genetic code.

Present findings indicate that the standard code evolved from a small code, with few amino acids. Evidence regarding the advancement of ‘code age’ in amino acid sets within these ancient
proteins encompassed an interval from stage 7 to 14 of code evolution (Table 3). Observations on pre-stage 7 sequences would be relevant to the initial stages of code evolution. Consensus sequence analysis is constrained, however, by the ‘code age’ of residue sequences among available LCA proteins. Phylogenetic evidence leading to reconstruction of a LCA protein sequence, with an internal repeat, can also specify the pre-duplication antecedent. Internal sequence repeats occurred in both Fdxn and PL, making it possible to reconstruct the residue sequence of antecedents from stage 10 and 7 of code evolution, respectively (Figs. 6 and 16). A quantitative correspondence between the residue profile in pre-duplication Fdxn and stage-5 codon array led to reconstruction of an even earlier, stage 5, antecedent of Fdxn (Section 2).

The residue profile of Fdxn and PL h1 antecedents showed good agreement with early codon assignment patterns, derived previously from biosynthetic evidence (Davis, 1999a). This complements statistical evidence showing that strongly conserved consensus sequence residues conform with the biosynthetic order of amino acid addition to the code (Table 3, Fig. 23).

5.2. Earliest protein sequence contained an amide residue

Pro-ferredoxin-[5] residue profile optimally matched the codon array at stage 5.6 ± 0.4 of code evolution (Fig. 8), consistent with synthesis by a mid-expansion phase code. No other protein in this study had an earlier residue profile. In view of its antiquity, it is significant that this small Fdxn antecedent (length, 23 residues) included an Asn² residue (Fig. 6). Occurrence of Asn² at sites 16 and 45 in ancestral Fdxn for Archaea (length, 55 residues) placed it in the pre-duplication sequence; ancestral Fdxn for Bacteria lacked this residue. A spacing of 29 residues between Asn² sites conformed with doubling of a 23-residue sequence followed by insertion of a 6-residue segment at site 24–29 (Section 2). Chance substitution with Asn² at equivalent sites in the internal repeat within ancestral archaeal Fdxn can be excluded, since this event had a probability of only $7.74 \times 10^{-3}$, or $(23/(55 \times 54))$. Omission of Asn² from ancestral bacteria Fdxn can be attributed to mutation during evolution from the LCA (Fig. 6).

Incorporation of Asn² into pro-Fdxn-[5] necessitates addition of this amide residue to the genetic code at some time before stage 5. Biosynthetic evidence indicates the N-fixers—Asp¹, Glu¹, Asn², Gln²—were the first amino acids to enter the genetic code (Davis, 1999a). Early incorporation of Asn² and Gln² is taken as evidence (Davis, 1999a) for a chemo-autotrophic origin of life (Wachterhauser, 1988, 1992, 1997, 1998). Neither occurs in amino acid preparations from abiotic sources (Kvenvolden et al., 1970; Miller and Orgel, 1974). A heterotrophic model for the origin of life (Miller et al., 1997), accordingly, favours late addition of both amino acids to the code (Weber and Miller, 1981). Pre-divergence amino acid synthesis apparently circumvented the lability of free Asn and Gln, by combining their acidic precursors with a tRNA cofactor prior to amidation (Davis, 2002). The distribution of identity elements in tRNA of thermophilic archaea and related species (Saks and Sampson, 1995) support early involvement of tRNA cofactors in amino acid synthesis. Moreover, tRNA cofactors are still retained at some reaction steps in prokaryote synthesis of amino acids (Wilcox and Nirenberg, 1968; Schon et al., 1988; Curnow et al., 1996; Gagnon et al., 1996; Wilting et al., 1997).
5.3. \([4\text{Fe}-4\text{S}]\) Pro-ferredoxin-[5] has a surface attachment segment

Pro-ferredoxin-[5] shows unmistakable signs of adaptation to a primordial surface system. It contains five acidic residues and all are located within a six residue segment (site 2–7) at its N-terminus (Fig. 9). When viewed from the perspective of the chemolithoautotrophic model for the origin of life, the unusually negative sequence in this ancient protein can be recognized immediately as a potential attachment site.

An autocatalytic system of surface reactions spawned metabolism, preceding emergence of cells. The ubiquity of multianionic reactants at the root of metabolic pathways suggested metabolites were initially bound by charge attraction to a mineral surface (Wachterhauser, 1988, 1992). Existence of a primordial charge attraction principle clearly fits well with the presence of a highly acidic segment at the N-terminus of pro-Fdxn-[5]. This segment could anchor the small, pre-cell protein molecule, with its \([4\text{Fe}-4\text{S}]\) e\(^{–}\)/C\(_{0}\) transfer centre, to a positively charged mineral surface. The presence of all four Cys\(^5\) residues, which ligate the \([4\text{Fe}-4\text{S}]\) cluster, in a non-charged C-terminal region strengthens this interpretation. All pro-Fdxn-[5] residues had entered the genetic code by stage 5, consonant with its participation in a pre-cell surface system.

The residue profile of LCA PL h1 indicates this membrane protein originated as early as 7.01 ± 0.67. Based on the ‘code age’ of residues in PL h1, therefore, membrane bound cells could have formed as early as stage 7. Consistent with this estimate, other indicators of cell formation converge at a time of origin between stage 8 and 9 of code evolution (cf. Section 5.7) and this fits with Fdxn being a relic from a pre-cell system.

5.4. Mutability of acidic attachment segment and appearance of cells

Pre-cell proteins with a highly negative attachment segment were likely to be highly unstable, after the loss of function in this segment, on replacement of the primordial surface system by cells. Present findings support this assessment.

Following the putative emergence of cells at stage 8–9 of code evolution, one of three pre-divergence substitutions in Fdxn antecedents occurred within the attachment segment. It evidently involved a Glu7Lys substitution (Fig. 6), which replaced an acidic residue adjacent to the first Cys (site 8) with a basic residue, consistent with selection against the highly charged attachment segment. The mutation rate (1/3) may be noted to exceed comparative segment length, which covered one quarter (18/69 bases) of the pro-Fdxn-[10] gene sequence. Phylogenetic evidence indicates that mutations within the attachment segment accounted for half (11/22) of all base substitutions during prokaryote divergence from the LCA. As the attachment segment then covered less than one-fifth ((2\(\times\)15)/165) of the Fdxn gene base sequence, this segment appears highly mutable.

Comparing pro-Fdxn-[5] with each ancestral prokaryote Fdxn (Fig. 6) reveals the attachment segment (residue site 2–7 and 31–36) in ancestral Fdxn for Bacteria differed by a total of 8 residues (Asp3Tyr, Glu5Thr, Glu6Asp, Glu7Lys, Asp32Val, Glu34Pro, Glu35Asp, Glu36Lys). While the attachment segment in ancestral Fdxn for Archaea differed by 4 residues (Val3Lys, Glu7Lys, Asp32Val, Glu36Lys) from pro-Fdxn-[5]. This compares with a difference outside the attachment segment (sites 1, 8–23, and 30, 37–52) of 7 residues (Val9Ile, Asn16Asp, Gly21Glu, Val23Ile,
Val38Ile, Asn45Ser, Val52Ile) between pro-Fdxn-[5] and ancestral bacteria Fdxn, and 5 residues (Val9Ile, Val 20Arg, Val23Ile, Gly50Tyr, Val52Leu) with ancestral archaea Fdxn.

Combining Fdxn mutations among prokaryote domains reveals that half (12/24) attachment segment residues were replaced overall. In contrast, less than one-fifth (12/68) of residues outside this segment were changed. This difference is highly significant ($\chi^2 = 8.03$, $P = 4.61 \times 10^{-3}$), suggesting strong selection forces were exerted against retention of the attachment segment, following the apparent replacement of the primordial surface system by cells at stage 8–9.

5.5. 5′-Initiation of translation before code expansion

The N-terminal attachment segment of pro-Fdxn-[5] consists effectively of stage-2 amino acids (Asp$^1$, Glu$^1$). By contrast, its C-terminal cofactor-binding region almost exclusively consists of stage-4 and -5 amino acids (Ala$^4$, Pro$^4$, Val$^4$, Cys$^5$, Gly$^5$). Polypeptides with an attachment segment, accordingly, could then have formed before expansion from the N-fixers code. This would require initiation at a codon within the NAN set. None of the standard initiation codons (GUG, AUG, UUG) would be available, however, as they were assigned later during expansion phase. With synthesis proceeding from N→C terminus, the pro-Fdxn-[5] sequence places an Asp$^1$ residue proximal to the initiation codon, so a codon from the GAN set may have been recruited to initiate translation in advanced stages of the N-fixers code. The resulting protein conceivably coupled an N donor function, residing in an amide-rich—Asn$^2$, Gln$^3$—C-terminal segment, with a negatively charged N-terminal segment that anchored it to a positively charged surface.

A Cys14Asp substitution has occurred in low potential Fdxn of the hyperthermophilic archaeon, *P. furiosus* (Zhou and Adams, 1997). The Asp$^1$ carboxylate coordinately bonds to an Fe atom of the cofactor, [4Fe-4S], replacing an S atom of Cys$^5$. Since Asp$^1$ was in the first generation of amino acids to enter the code (Davis, 1999a), pre-stage 5 antecedents of Fdxn and possibly other metal chelating proteins, plausibly preceded pro-Fdxn-[5].

Since local phasing by codons with a mid-A base was no longer possible on expansion from the N-fixers code, initiation of synthesis at the 5′-codon of a translated base sequence constitutes a precondition for expansion (Davis, 1999a). Inclusion of 5′-initiation into protein synthesis during advanced stages of the N-fixers code, consequently, would open the way to expansion from a code with only 16 codons (NAN set) to a code with all 64 codons.

Before the general nature of the genetic code was known, Crick et al. (1957) constructed a commaless code. Present considerations indicate that early protein synthesis utilized a simple commaless code. Local phasing of translation, with the N-fixers code, appears well suited to synthesis on a template containing short sequence repeats. Both Fdxn and PL h1 exhibit early low-order periodicity (Sections 2 and 3), possibly from tandem duplication of an oligomeric antecedent (Eck and Dayhoff, 1966). A template with multiple repeats, and local phasing of the reading frame, would allow translation to recommence after an interruption at any downstream codon and still produce a functional protein.

One-third (any mid-base substitution) of mutations occurring in the N-fixers code result in nonsense transitions to an unassigned codon. While one-sixth (1/4 × 2/3, at 5′- and 3′-codon base) produce a Ter codon. The former is a severe error that blocks upstream translation (Bretsch et al., 1965), whereas the latter simply results in premature chain termination of a seemingly repetitious polypeptide.
Mutation rates at any stage of the pre-divergence era were unlikely to exceed $10^{-2}$ per nucleotide copied, the rate in replicase-free polynucleotide replication (Inoue and Orgel, 1982). With small periodic polypeptides and A-rich templates, early synthesis even at this error rate would still, for example, complete more than 90 per cent of octapeptides. When improved fidelity and 5' initiation replaced local phasing of translation, exposure to unassigned codons receded with expansion of the code to all 64 base triplets.

5.6. Sequence extension by tandem duplication

Ferredoxin and PL residue sequences from 82 and 107 phylogenetically diverse sources, respectively, contained internal periodicities (Figs. 5 and 15). Most conspicuous of these were sequence repeats extending 23-residues in Fdxn and 11-residues in PL transmembrane helix-1. Phylogenetic evidence reveals they were present in LCA Fdxn and PL h1. Hence, tandem duplication of pre-divergence genes (Schwartz and Dayhoff, 1978a, b), as by unequal cross-over during chromosome recombination, provided an early means of sequence extension.

Lysine$^{10}$ occurred at site-7 and -36 in LCA Fdxn. The 29 residue displacement between Lys$^{10}$ sites indicates this residue occurred in the pre-duplication protein, pro-Fdxn-[10]. It is attributed with entering the code at stage 10, three stages or more beyond the ‘code age’ of other residues in pro-Fdxn-[10]. A Glu:Lys substitution seems to have occurred in an antecedent, whose sequence subsequently doubled and fused (Fig. 6). Tandem duplication of the pro-Fdxn-[10] gene (length, 69 bases), therefore, occurred no earlier than stage 10 and, from phylogenetic evidence, no later than LCA emergence.

An earlier lower limit applied to tandem duplication of the ancestral PL h1 gene. Leucine$^{7}$ at site-5 and -16 was the most advanced internally conserved residue in LCA PL h1 (Fig. 16). Duplication + fusion of the pro-PL h1 gene (length, 33 bases) thus occurred no earlier than stage 7. Phylogenetic evidence indicates this event preceded species divergence. Broadly, tandem duplication of the ancestral gene for Fdxn and PL h1 took place after cells appeared at stage 8–9 of code evolution (Section 5.7) and, most likely, after DNA synthesis and gene consolidation into chromosomes had begun.

Homology between PL h1 and h2 (Section 3) indicated both transmembrane helices evolved from a common ancestor that duplicated in the pre-divergence era. Insertion of a polar linking segment would then produce the archetypal hairpin PL molecule. Post-divergence sequence duplication is a feature of both Fdxn and PL evolution. Formation of a double-hairpin proteolipid subunit in V-ATPase (Gogarten and Taiz, 1992; Fig. 16) and multi-hairpin subunit in A-ATPase (Muller et al., 1999), together with polyferredoxin formation, illustrate sequence extension by tandem duplication following species diversification.

Weaker repeats of small segments occurred in both Fdxn and PL (Figs. 5 and 15). It has been suggested the ancestral Fdxn residue sequence initially contained multiple repeats of the tetramer, Ala-Asp-Ser-Gly (Eck and Dayhoff, 1966). Successive identity peaks within the Fdxn residue sequences examined in this study were separated by an average of 2.8 ± 0.7 residues. Inter-peak spaces of 1 and 3 residues were equally frequent and together they accounted for 10 of 14 identity peaks. In the N-terminal region (attachment segment) of pro-Fdxn-[5], 4 (of 5) acidic residues were nearest neighbours. While 3 (of 4) Cys$^5$ residues in the C-terminal region ([4Fe-4S] cluster binding site) were next-next-nearest neighbours. Early structural constraints on the primary
structure of Fdxn, rather than tandem duplication, could thus underlie the weak low order repeats observed in the Fdxn residue sequence. Likewise, the pitch of the LCA ATPase PL h1 transmembrane helix arguably produced its tetramer repeat (Fig. 17).

5.7. Protein hydrophobic domain and the origin of cells

Development of the hydrophobic protein domain emerges as a fundamental innovation in protein evolution. Its relation to emergence of the fluid-mosaic cell membrane is considered here.

A number of lines of evidence indicate membrane covered cells evolved about midway during code evolution (Table 4). Basic amino acids—Arg⁹, Lys¹⁰—form on pathways extending nine or 10 steps from the citrate cycle. By contrast, both acidic amino acids—Asp¹, Glu¹—form on pathways extending only one reaction step from their source in the citrate cycle. This suggests polypeptides were initially polyanions and served to anchor cofactors and other functional groups to a positively charged mineral surface (Section 5.3). Addition of basic amino acids to the code, seen in this context, signaled a major shift in direction in protein evolution. This change followed incorporation of increasingly hydrophobic residues during code expansion. These findings fit with elimination of a primordial surface system on appearance of cells, encapsulated by a permeable phospholipid membrane with embedded hydrophobic proteins (Davis, 1999a). Inclusion of the first basic amino acid in the genetic code at stage 9 indicates cells evolved before stage 9 of code evolution (Table 4).

Proteolipid h1 in LCA H⁺-ATPase has a residue profile with a ‘code age’ of 7.1 ± 0.7 (Table 3). Ancestral PL consequently formed no earlier than stage 6.4–7.8 of code development. A significant feature of the PL h1 sequence was exclusion of charged residues. In particular, the lack of acidic residues precluded direct PL participation in a pre-cell system, reliant on charge attraction. Molecules at the root of biosynthetic pathways, within the ancient pathways of central metabolism, possess multiple anionic groups consistent with participation in a primordial surface system (Wachterhauser, 1988, 1992). A negatively charged N-terminal segment also characterized pro-Fdxn-[5] primary structure (Sections 2 and 5.3). Addition of basic amino acids at stage 9–10, consequently, represented a significant change in the direction of code evolution. Replacement of the primordial surface system by cells has been implicated in this shift. The small hydrophobic PL h1 antecedent, pro-PL-[7] would then partition with the membrane phospholipid bilayer. The PL

Table 4
Stage of code development attributed to aspects of pre-divergence evolution linked to origin of cells

<table>
<thead>
<tr>
<th>Aspect of pre-divergence evoln.</th>
<th>Stage (Limit)³</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteolipid residue profile</td>
<td>7.1 ± 0.7³²</td>
</tr>
<tr>
<td>Septation protein (FtsZ) residue profile</td>
<td>7.5 ± 0.5³²</td>
</tr>
<tr>
<td>Length of phospholipid synthesis path</td>
<td>10³³ (Lower)</td>
</tr>
<tr>
<td>Earliest basic amino acid in code</td>
<td>9³³ (Upper)</td>
</tr>
</tbody>
</table>

a Stage of code development corresponds to number of reactions synthesis pathway extends from citrate cycle.

b Source, Table 3.

c Source, Rodwell (1969), Michal (1978) and Davis (1999a).
residue profile places its origin no earlier than stage 7. Permeable cell membranes accordingly originated as early as stage 7 (Table 4).

The length of the phospholipid pathway also sets a lower limit on the origin of cells. Their synthesis commences with acetyl-coenzyme A formation from citrate, followed by cyclic addition of an acyl group to a growing hydrocarbon chain. Two hydrocarbon chains ultimately become attached to a phosphoglycerol molecule (Michal, 1978). The phospholipid pathway extends 10 reactions from the citrate cycle, on discounting repetitious steps. Growth of the phospholipid pathway is taken to occur at a comparable rate, for present purposes, as amino acid pathways originating at oxaloacetate and \( \alpha \)-ketoglutarate. The phospholipid bilayer then originated no earlier than stage 10 of code evolution (Table 4). Considering the assumptions involved in this estimate, it conforms with appearance of cells between expansion and capture phases (stage 8) of code evolution.

Prokaryote septation protein, FtsZ, originated in the pre-divergence era (Table 3). This protein, together with that of other early cell division proteins, seems likely to have coevolved with cells. The FtsZ residue profile places its origin at no earlier than stage 7.5–7.0. Considering this as a lower limit also for the time of origin of cells indicates they appeared no earlier than stage 6.9 to 8.1 of code evolution (Table 4).

On combining these estimates, cells appeared no earlier than stage 6.4–10. The mid-range of lower bound estimates is stage 8.2. Addition of the first basic amino acid placed an upper bound of stage 9 on the origin of cells (Table 4). It may be concluded that cells evolved at some time between stage 8 and 9 of code evolution. This estimate fits well with a significant elevation in mutation rate within the highly acidic pro-Fdxn attachment segment, after stage 10, attributed to the transition from a surface system to cells (Sections 2 and 5.4).

5.8. Retarded origin of DNA and the RNA–DNA heteroduplex

Core subunits of DNA dependent RNA polymerase and 5’→3’ flap exonuclease (FEN-1), exhibit significant sequence homology among bacteria, archaea and eukaryote sources, as required of pre-divergence proteins (Gropp et al., 1986; Sweetser et al., 1987; Langer et al., 1995; Edgell and Doolittle, 1997; Leipe et al., 1999; Section 4). Their role in the synthesis and removal of hybrid molecules possessing single DNA and RNA strands conforms with pre-divergence polymerization of a heteroduplex. Transcription protein homology contrasts with the heterogeneity between DNA polymerases from bacteria and archaea/eukaryotes. Leipe and coworkers suggest this indicates the transition from RNA genes to a DNA genome was not completed until after species diversification. In the event that DNA polymerase originated independently in bacteria and archaea/eukaryotes, the lack of evidence for a pre-divergence DNA polymerase becomes understandable. It also implies that the LCA had a hybrid DNA–RNA genome.

An initial evaluation of ‘code age’ among replication proteins, including ancestral RNAP core subunits, FEN-1, and DNA polymerase (Davis, 1999b) independently led to the same conclusion. Results in Table 3 relating to FEN-1 and RNAP core subunits involve residue sequences from 67 to 138 sources, in three species domains. LCA FEN-1 had an estimated ‘code age’ of stage 9.1 ± 1.4. While LCA \( \alpha \), \( \beta \) and \( \beta’ \) RNAP subunits had a ‘code age’ of stage 9.3 ± 1.4, 10.9 ± 0.8 and 10.9 ± 0.9, respectively. They yield an overlapping set of intervals that span stage 7.7–11.8, with a mid-range value of stage 9.8. This sets a lower limit on their time of origin in the early capture
phase of code evolution. From these results, the DNA–RNA heteroduplex formed no earlier than stage 10 and no later than emergence of the LCA.

RNR residue sequences from 67 sources in bacteria, archaea and eukaryotes revealed the reductase contained a conserved set of amino acids that matched codon assignments at stage $11.0 \pm 0.9$ (Table 3). According this, synthesis of deoxyribonucleotides began no earlier than stage 10.1–11.9. The lower end of the range in ‘code age’ for deoxynucleotide synthesis coincides with the mid-range value for emergence of ancestral FEN-1 and RNAP core subunits. Stage 10 of code evolution thus arises as a probable lower limit for appearance of the DNA–RNA heteroduplex.

This transitional LCA DNA–RNA genome requires pre-divergence RNA dependent DNA synthesis. Conserved amino acids in RT residue sequences from 23 sources in eubacteria and eukaryotes revealed, however, that LCA RT formed no earlier than stage 13 or 14 (Table 3). According to this estimate, RT formed 3 stages after the DNA–RNA genome, whose formation was possibly delayed until late in the pre-divergence era. Alternatively, RT formation was delayed. An early reliance on ribozyme catalyzed RNA dependent DNA polymerization could have retarded RT formation. A ribozymal RNA dependent RNA polymerase (Cech, 1986) is credited with RNA replication before the first proteins evolved.

It may seem paradoxical that cells of Bacteria and Archaea/Eukarya share a DNA genome, but have a common ancestor with only a transitional DNA–RNA genome. This finding implies that a fundamental molecular entity, such as the LCA genome, was still malleable in the post-divergence era. It is relevant to recall, in this connection, that the LCA genetic code was also modified by addition of f-Met.

5.9. Gene consolidation

Notwithstanding involvement of a transitional genome, pre-divergence replication appears sufficiently advanced, from an enzymological standpoint, to accommodate chromosome doubling (Edgell and Doolittle, 1997; Leipe et al., 1999; Davis, 1999b). The presence of Topo I orthologues in Bacteria, Archaea and Eukarya confirms that this enzyme was among LCA replication proteins. Topoisomerase I residue sequences from 39 sources in three species domains conserved amino acids with a frequency distribution that matched the codon array at stage-14 (Table 3). Consolidation of genes into chromosome(s) had evidently proceeded to the point that a reduction in double helix winding was required by the late pre-divergence era. Chromosome formation thus followed synthesis of the DNA–RNA heteroduplex. It is inviting to conjecture that linear extension of the genome consequent on gene consolidation became possible after the synthesis of comparatively thermostable DNA strands, under early Earth conditions (Baross, 1998).

5.10. Overview and the final frontier

Amino acid sequences of great antiquity were analyzed here to chronicle events in protein evolution before bacteria, archaea and eukaryotes diverged from the LCA. Initially they were seen to be small polyanionic molecules, akin to an RNA adaptor, implicated in anchoring a cofactor, or functional residue, to a positively charged mineral surface. Thereafter, proteins evolved into
Ordering events in early protein evolution relied on matching the set of conserved amino acids in an ancient residue sequence with the biosynthetic order of amino acid addition to the genetic code (Davis, 1999a). Support for the proposed path of code evolution, on which these studies rest, came mainly from insights it provided into the source of hitherto unexplained features of the genetic code. Table 5 enumerates these features and it is evident that a range of factors shaped the standard code. They include biosynthetic (tRNA cofactors, path length), chemical (\(\text{CO}_2, \text{NH}_4^+\) fixation, autocatalytic reactions), physical chemical (residue charge, polarity/hydrophobicity), kinetic (codon-anticodon affinity, peptidyl transferase reactivity), error avoidance (residue clusters), and historical (small founder amino acid set) influences. The biosynthetic model is unencumbered by a requirement that all precursor/product pairs have nearest-neighbour base triplets, contrary to the belief of Wong (1975). In fact, the model provides a mechanism whereby four precursor/product pairs (Asp/Thr, Asp/Met, Glu/Pro, Asp(Glu)/Arg), excluded from consideration by Wong, were assigned codons that differ by more than one base (Davis, 2002). Homology among tRNA core structure groups (Saks and Sampson, 1995) indicated the precursor of amino acids with non-conforming codons was misacylated onto a variant acceptor for a sibling amino acid, producing sibling/product nearest-neighbour codons.

Additional evidence for the biosynthetic model was obtained from the primary structure of ‘fossil’ proteins. Two predictions of the model were confirmed in this survey. First, results in Section 2 establish that Asn\(^2\) was incorporated into the earliest protein identified (Section 5.2), the stage 5 pre-duplication antecedent of Fdxn. This revealed Asn\(^2\) was an early addition to the genetic code, consistent with N fixers forming the first code. Second, the set of amino acids at conserved sites in the 10 ‘fossil’ proteins examined (Sections 4 and 5.1) exhibited an earlier ‘code age’ (shorter synthesis path length) than non-conserved amino acids. This helped validate the biosynthetic order of amino acid addition to the genetic code (Davis, 1999a).

‘Fossil’ proteins of increasing ‘code age’ exhibit function coherence, giving added credence to findings obtained from the biosynthetic model of code evolution. Pro-ferredoxin [5], as noted, was the earliest protein identified. The primary structure of this small protein (length, 23-residues) included a putative N-terminal attachment segment, consistent with its participation in a primordial surface system (Section 2). Pro-proteolipid-[7] followed with a conserved residue profile having a ‘code age’ of stage 7. Omission of a negatively charged attachment segment, combined with the hydrophathy of its residue sequence, identified pre-duplication PL as a marker for cell membrane formation (Section 3). It was closely followed by prokaryote cell division protein, FtsZ, whose conserved residues had a ‘code age’ of stage 7.5 (Section 4). Subsequently, proteins implicated in DNA synthesis, transcription and replication followed, whose conserved residues had a code age of stage 10–11 (Section 4). In accord with expectation, the transition from RNA genes to a DNA genome did not begin until (i) development of catalysts with a hydrophobic domain, to channel radical intermediates toward deoxyribonucleotide synthesis, and (ii) cells had emerged. Finally, Topo-I and RT had a stage-14 residue profile, characteristic of proteins evolving late in the pre-divergence era (Section 4). Late formation of the DNA topology altering protein, Topo-I, conformed with consolidation of RNA genes into chromosomes, after synthesis of more thermostable DNA strands. On the other hand, late RT formation may appear misplaced. Had RT coevolved with RNAP, it could have helped replicate the DNA–RNA heteroduplex...
Proposed origin of previously identified features of the genetic code

<table>
<thead>
<tr>
<th>Feature of code</th>
<th>Ref.</th>
<th>Origin according to biosynthetic model of code evolutiona</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Polarity/hydrophobicity difference between amino acids with NAN and NUN codons</td>
<td>Woese (1965)</td>
<td>Polar amino acids, assigned NAN codons, formed the first code. Addition of increasingly hydrophobic amino acids accompanied expansion of the code to all 64 base triplets. NUN codons were assigned most slowly and, consequently, acquired the most hydrophobic amino acids. Proteins evolved outside → in.</td>
</tr>
<tr>
<td>2. Biosynthetically related amino acids have neighbouring codons</td>
<td>Nirenberg et al. (1966)</td>
<td>Amino acid addition to the code accompanied extension of an existing synthesis pathway and formation of a variant tRNA cofactor, in accord with related amino acids having tRNA with core homology.</td>
</tr>
<tr>
<td>3. All eight family boxes (sets of four) contain codons with G or C as a 5' and/or mid-base</td>
<td>Dunnill (1966)</td>
<td>During rapid expansion from the first code, codons were assigned four-at-a-time, initially filling the G/C, G/C,N triplet sets. These codon assignments were stabilized by elevated G-C bp binding energy.</td>
</tr>
<tr>
<td>4. Six of eight family boxes have the smallest amino acids in the code</td>
<td>Taylor and Coates (1989)</td>
<td>Initial expansion from the small first code assigned family boxes to Ala, Pro, Ser and Val. In the next two stages Gly and Thr received a box each. Their mean molecular weight (103.3) is three-quarters of the average for all coded amino acids.</td>
</tr>
<tr>
<td>5. Codon mid-base has most coding significance</td>
<td>Perlwitz et al. (1988)</td>
<td>Mid-base codon substitutions added 10 amino acids during code expansion. By comparison, 3 of 4 amino acids were solely specified by the 5'-base of the first code, and 6 by 3'-base changes during post-expansion codon capture.</td>
</tr>
<tr>
<td>6. Codons for biosynthetically related amino acids exhibit 5'-base invariance</td>
<td>Taylor and Coates (1989)</td>
<td>Mid-base substitutions to anticodons in the 3'-NUU set (NAN codons) led to expansion from the first code. tRNA cofactors in primordial pathways coupled these changes to synthesis of new amino acids.</td>
</tr>
<tr>
<td>7. Acidic amino acids are synthesized on short paths while basic amino acids have long paths</td>
<td>Davis (1999a)</td>
<td>Early proteins were polyanionic, consistent with binding to a mineral surface. Basic amino acids were evidently incorporated after membrane-covered cells had evolved.</td>
</tr>
<tr>
<td>8. All N-fixing amino acids have codons in the NAN set, the oldest part of the genetic code</td>
<td>Davis (1999a)</td>
<td>Protein synthesis originated from N-fixation reactions, coupled to the autocatalytic reductive citrate cycle, consistent with a chemo-autotrophic origin of life.</td>
</tr>
<tr>
<td>9. Amino acids with ring structures form on long paths related to nucleotide synthesis</td>
<td>Garrett and Grisham (1999)</td>
<td>Nucleotide-like amino acids, notably His and Trp, were late additions to the code suggesting ribozymal catalytic mechanisms became incorporated into enzymes in the final stages of code formation.</td>
</tr>
</tbody>
</table>

a Davis (1999a).
(Sections 4 and 5.8). A ribozyme RNA dependent DNA ‘polymerase’ plausibly delayed the evolution of RT.

Ribozyme RNA dependent RNA ‘polymerase’ (Cech, 1986) evidently catalyzed replication of tRNA and proto-ribosome molecules in pre-code translation. RNAR from 15 (+)-strand viruses (Xiong and Eickbush, 1990) displayed a mature residue profile and low homology (identity, 16 per cent) at 168 sites (Davis, 1999b), consistent with a ribozyme retarded origin, possibly in the post-divergence era. The biosynthetic model points to a phase of pre-code translation, where acceptor and template RNA were non-informational components of the proto-ribosome, still undifferentiated from early rRNA (Davis, 1999a). Inspection of *E. coli* ribosomal protein sequences (Giri et al., 1984) reveals they contain a mature residue profile with standard frequencies, compatible with formation after completion of the genetic code. Only S8, L11 and L12 lacked His^{13} and Trp^{14}. Consistent with its limited residue range, the L12 phylogenetic distribution suggests it is a pre-divergence protein (Shimmin et al., 1989). Most ribosome mass is contributed by rRNA, which still apparently retains functional significance in amino acid condensation (Cate et al., 1999). Early translation molecules and their ribozymal replicase could even predate RNA, since a glycerol-2-phosphate backbone would link prochiral, acyclic monomers (Joyce et al., 1987).

The deepest branches of the prokaryote ‘tree’ constructed from 16S RNA sequences (Fox et al., 1980) unexpectedly revealed the LCA was a comparatively advanced thermophilic autotroph (Woese, 1987; Kandler, 1998; Woese et al., 2000). Phylogenetic evidence subsequently placed this node between Bacteria and Archaea/Eukarya (Section 3). Beyond this, however, the era of pre-divergence evolution remained obscure. A beginning has now been made in bringing molecular evolution before the LCA into focus. At its deepest level, the present investigation furnishes evidence of a primitive, fundamentally different kind of self-propagating entity. It appears to have been a system of autocatalytic reactions in contact with a positively charged mineral surface, as inferred by Wachterhauser (1992). Mystery still surrounds its origin and subsequent evolution, reminiscent of the LCA two decades ago, before the Woesian revolution in prokaryote phylogenetics. This issue now looms as the final frontier in the long quest to understand how life began on Earth.

Inclusion of a small, highly acidic segment at the N-terminus of the earliest protein residue sequence identified (Fig. 9) provides compelling evidence for molecular attachment to a positively charged mineral surface. Multianionic molecules likewise occur at the core of metabolism, in the ancient pathways forming the citrate cycle, central trunk and pentose cycle. These pathways branch to biosynthetic pathways for amino acids, purines, pyrimidines, fatty acids, phospholipids, isoprenoids, and porphyrins (Hartman, 1975; Wachterhauser, 1992). In addition, mineral surfaces catalyze simple organosynthesis reactions (Gabel and Ponnamperuma, 1967; Reid and Orgel, 1967; Pitsch et al., 1995; Heinen and Lauwers, 1996; Huber and Wachterhauser, 1997). A shift in the thermodynamics of condensation reactions involving surface-bound monomers toward product formation (Wachterhauser, 1992; Arrhenius et al., 1997) reasonably accounts for elevated polymerization among mineral-bound polypeptides and polynucleotides (Ferris et al., 1996; Hill et al., 1998). Furthermore, Wachterhauser (1998) noted that the narrow kinetic energy distribution of surface-bound reactants would enhance thermostability, compared with dissolved reactants, at elevated early Earth temperatures.

Reduced Fdxn [4Fe-4S] is a proton donor at two sites in the reductive citrate cycle of anaerobe, *Chlorobium thiosulfatophilum*, where oxaloacetate is autocatalytically produced from CO₂ (Evans
et al., 1966). The reliance on an FeS cofactor suggests this ancient pathway may be retrodicted to reductive organosynthesis with CO$_2$, as carbon source, and free energy from oxidation of FeS, in the presence of H$_2$S (Wachterhauser, 1992). In an early form of the reductive citrate cycle summarized in Eq. (9a),

\[
\begin{align*}
4\text{CO}_2 + 4\text{H}_2\text{O} &\rightarrow 4\text{H}_2\text{CO}_3 \\
4\text{H}_2\text{CO}_3 + \text{H}_2\text{C}.\text{CO} (\text{COOH})_2 &\rightarrow 2\text{H}_2\text{C}.\text{CO} (\text{COOH})_2 + 2\text{H}_2\text{O} \\
5\text{FeS} + 5\text{H}_2\text{S} &\rightarrow 5\text{FeS}_2 + 5\text{H}_2\text{O}
\end{align*}
\tag{9a}
\]

oxaloacetatic acid catalyzes conversion of four molecules of hydrated CO$_2$ to one oxaloacetic acid molecule (line 2), coupled with FeS oxidation at a mineral surface. Conversion of some oxaloacetate to aspartate, in a second N-fixation reaction, would form an autocatalytic network able to accelerate CO$_2$ fixation. Bar-Nun et al. (1994) reported that all first generation additions to the code (Davis, 1999a)—Asp$^1$, Glu$^1$, Asn$^2$, Gln$^2$—exhibit carbonic anhydrase-like activity. A citrate cycle with attached Asp branch may be depicted as

\[
\begin{align*}
4\text{CO}_2 + 4\text{H}_2\text{O} + \text{H}_2\text{NCH} (\text{CH}_2\text{COOH}) \text{COOH} &\rightarrow 4\text{H}_2\text{CO}_3 + \text{H}_2\text{NCH} (\text{CH}_2\text{COOH}) \text{COOH} \\
4\text{H}_2\text{CO}_3 + \text{H}_2\text{C}.\text{CO} (\text{COOH})_2 &\rightarrow 2\text{H}_2\text{C}.\text{CO} (\text{COOH})_2 + 2\text{H}_2\text{O} \\
5\text{FeS} + 5\text{H}_2\text{S} &\rightarrow 5\text{FeS}_2 + 5\text{H}_2\text{O} \\
\text{NH}_3 + \text{H}_2\text{C}.\text{CO} (\text{COOH})_2 &\rightarrow \text{H}_2\text{NCH} (\text{CH}_2\text{COOH}) \text{COOH} \\
\text{FeS} + \text{H}_2\text{S} &\rightarrow \text{FeS}_2 + \text{H}_2\text{O}
\end{align*}
\tag{9b}
\]

Asp$^1$ occurs as a catalyst on line 1 and a product on line 4, so the reaction network is autocatalytic, like its predecessor (Eq. (9a)). Since Asp$^1$ is a surface-bonder, all reactions in the network could take place at the surface of a FeS containing mineral exposed to CO$_2$, as near a submarine vent.

Extension of the Asp$^1$ path with a second N-fixation reaction yields Asn$^2$, a non-surface bonder bearing an uncharged side chain. Condensation of Asn$^2$ with Asp$^1$ or Glu$^1$ forms a polyanionic polypeptide chain that could anchor the amide residue, or its analogue Gln$^2$, to a mineral surface. An ordered residue sequence is required to optimize acidic and amide residue mole ratios in an early N donor. This necessitates template-directed synthesis and thereby provides a molecular basis for development of the N-fixers code (Davis, 1999a).

The origin of life is linked in Eqs. (9a) and (9b) to self-organizing autocatalytic reactions that extend to simple molecules, comparable in complexity to those from abiotic sources (Lazcano and Miller, 1999). They convert CO$_2$ to 4C molecules (oxaloacetate, Asp$^1$). The spontaneous autocatalytic formose cycle (Boutelrow, 1861) converts H$_2$CO into a 3C molecule (glyeraldehyde). Phosphorylated components of this cycle, including phospho-glyceraldehyde, dihydroxyacetone-phosphate and erythrose-phosphate, occur in the central trunk, or react with components of this pathway. Empirical evidence bearing on the primordial surface system (Kandler, 1998; Russell
et al., 1998; Wachterhauser, 1998; Lahav et al., 2001) would clearly advance our understanding of the initial steps in the origin of life on Earth.

References


