Associations of enzyme's evolutionary conservation with structural localizations in Archaea, Bacteria and Eukaryota of the Kreb's cycle and also with volumetric properties of geochronologically categorized Bacteria

J. Dennis Pollack[a*], David Gerard[b], George I. Makhatadze[c], Dennis K. Pearl[d*]

[a]Department of Molecular Virology, Immunology and Medical Genetics, College of Medicine, The Ohio State University, Columbus, Ohio 43210, USA; [b]Department of Mathematics and Statistics, American University, Washington, D.C. 20016, USA; [c]Department of Biological Sciences, Rensselaer Polytechnic Institute, Troy, New York 12180, USA; [d]Department of Statistics, Penn State University, University Park, Pennsylvania 16802, USA

We studied MSA consensus amino acid distributional patterns in 2,844 amino acid sequences of the eight enzymes of the Kreb's oxidative tricarboxylic acid pathway (oTCA) in Archaea, Bacteria and Eukaryota and 5,545 sequences of 33-Bacteria as geochronologically separated enzymes and with multiple-sequence-alignment (MSA) consensus site modal identities. The 33-Bacteria were selected as 20 presumptive examples of early-oldest (Archaean-Hadean) ('Epoch I') or 13 late-newest (contemporary) ('Epoch III') appearing enzymes on Earth. Each MSA consensus' sites were appended with its modal identity, amino acid's % Occupancy, one of nine-graded evolutionary-conservation zones (CZs) and the site's Euclidean distance (Å) to the same atom in the reported functional center of a special ('Scaffold') PDB-sequence in the respective enzyme's FASTA set. MSA consensus sites are 'tetrad'-data points or RAA's (Recovered-amino Acids). Across Domains, the % Occupancies of the eight-dominant RAAs of the Kreb's cycle and the 33-Bacteria were consistently found similarly localized. Compared to Trifonov's '*putative ranked temporal order of the appearance of amino acids on Earth*' (TOAE) the greatest statistical concordance across Domains with tetrad-RAAs was with the most-evolutionary conserved conservation zone (CZ9) typically nearest (Å) their respective enzyme's catalytic/active center. The geochronologically characterized early-oldest Hadean-Archaean Bacteria enzymes compared to late-newest Bacteria enzymes had greater average numbers of amino acid residues/sequence and statistically significantly larger variability in their RAA compositional $Å^3$-volumes. The late-newest Bacteria enzymes of 'Epoch III' had statistically significant lower volumetric values: native $Å^3$-volume, void-volume and volume change on unfolding. Our data has suggested a geochronological trace of '*metabolism's progressive emergence'*.

---

*Corresponding author~~s~~. Email: pollack.1@osu.~~edu~~; dkp13@psu.edu

‘The first step in wisdom is to know the things themselves’

Carl Nilsson Linnæus (1707-1778)

1. Introduction

We studied the 20-biologic amino acids in sequences of eight structurally stable enzymes of the oxidative tricarboxylic acid pathway (Kreb’s pathway or the oTCA) and 33 Bacteria. The 20-biologic amino acids are characterized as Earth’s ‘surviving metabolic fossils’ – their prevalence and encompassing roles are well recognized as a conviction of the continuous evolutionary passage of ‘metabolism’s progressive emergence’ (e.g., Caetano-Anollés, Kim & Caetano-Anollés, 2012; De Duve, 1991; Eigen, 1992; Fry, 2000; Ingles-Prieto et al., 2013; Smith & Morowitz, 2016). Our premise was that there may be a discernable ‘trace’ of their progress in contemporary enzyme sequences.

The oTCA enzyme sequences were used to first determine the distributions of their 20-biologic amino acids across all Domains. The 33-Bacteria enzymes were additionally separated into one of two geochronological periods of Earth’s history, either the Hadean-Archaean era 4,000 to 2,700 Ma (millions-of-years) (‘Epoch I’) or to 2,100 Ma to present (‘Epoch III’). The distinction involved assignment to each enzyme a chronologic “nd-distance or age” value using both: a phylogenetic model, the MANET data base developed by Caetano-Anollés and associates (Kim & Caetano-Anollés, 2012; Kim, Mittenthal & Caetano-Anollés, 2006; Wang et al., 2011) and the expanded scale of Earth’s geochronological record in the ‘International Chronostratigrphic Chart’ (Cohen, Finney, Hibbard & Fan, 2013). Multiple sequence alignments were constructed of all enzyme sets. MSA consensus site of each set were uniquely characterized by their: 1) site’s modal amino acid’s identity, 2) the % Occupancy of the modal amino acid, 3) its position in a one-to-nine zone scale of sequentially graded evolutionary conservation values and 4) our calculation of the Euclidean distance (Å) from the modal amino acids Cα in each evolutionary conservation zones to the *same* atom in the enzyme’s reported catalytic/active center. The MSA modal consensus line site assembly is characterized as a coalescent ‘four data point’ – our ‘Recovered Amino-acid’ (RAA) that we also call a “tetrad”.

The RAAs of the ‘Epoch I and III’ 33-Bacteria were additionally distinguished by comparisons to a sequentially graded historical scale using the ‘putative temporal order of amino acids’ appearances on Earth’ – the TOAE (Trifonov, 2000). We also calculated both the Epoch’s average residue count per enzyme sequence and by the ‘ProteinVolume’ program enzyme to determine their molecular volume ($Å^3$), their void volume and volume change upon unfolding (Chen & Makhatadze, 2015, 2017a, 2017b).

We studied protein sequences using the familiar ‘top-down’ procedure (Granick, 1957; Lipmann, 1965) and characterized the Epochs as either ‘early-oldest’ ’Epoch I’ or ‘late-newest’ ‘Epoch III’. The approach is often criticized with reasonable and cautionary implications that derived protein chemistries

may be unreliable historical accounts by lacking discernable and consequential marks modified or destroyed in their evolutionary passage to modern biochemistry (e.g., Lazcano and Miler, 1996, 1999; Peretó, Fani, Leguina & Lazcano, 1997; Raggi, Bada & Lazcano, 2016; Strasdeit, 2010). Top-down approaches are more specifically discussed elsewhere (e.g., Ikehara, 2016; Morowitz, 1992).

We value and have used in this study large sets of critically assembled enzyme sequences as contemporary based reproducible evidenciary molecular constructs of evolutionary processes by carefully considering their role as 'the only true guide to how life came to be is life as we know it today is "the top-down approach" ' (Lane, 2010). *By encompassing a geochronological time scale we have attempted to more specifically recognize attributable evolutionary changes in contemporary Bacteria protein sequences by nd-distance or -age values that may distinguish their historical placement in Earth's 'Epoch I' (4,000-2,700 Ma) or 'Epoch III' (2,100 Ma-present).*

2. Materials and Methods

2.1. Enzyme Sequences of the oxidative TCA and 33 geochronologcally characterized Bacteria: their FASTA sets and MSA consensus sites each with a conservation and distance value to respective catalytic/active centers (C/ACs)

We studied the distributions of the 20-biologic amino acids of 2,844 protein sequences of the eight oTCA enzymes from Archaea (336), Bacteria (2,122) and Eukaryota (366) (Table 1) and a 5,545 protein sequence set of 33 geochronologically assigned Bacteria enzymes (Table 2). The 41 enzyme examples are represented by their reported presence in one to no more than six arbitrarily chosen unduplicated species per genera. Each homologous enzyme sequence collection includes a most consequential PDB sequence(s) called the 'Scaffold'. The Scaffold's sequence contains one of the 20-biologic amino acid residues that is *reported* to be functionally or mechanistically involved at its enzyme's catalytic/active center (C/AC) that we call an 'Anchor-amino acid' (AAA) (Berman et al., 2000; de Beer et al., 2013; Furnham et al., 2014; Laskowski, 2016). The Anchor-amino acid contains a PDB-identified *atom* that is also reported to be functionally or mechanistically involved and called the 'Anchor-atom' (AA) used for all Euclidean distance measures. The AA represents a relatively 'precise' locus of its enzyme's catalytic/active center's (C/AC). The AA is used by the Yasara program as the *same* terminus in calculating its Euclidean distance (Å) to the Cα of each MSA consensus amino acid at each of its respective consensus MSA's sites of each enzyme (Krieger & Vriend, 2014). When a Scaffold's Anchor-atom is not available we substituted its Anchor-amino acid's Cα.

The 2,844 sequences (Table 1) of the eight enzymes of the *oxidative* tricarboxylic acid pathway (Krebs cycle) (oTCA) were identified following nomenclature of the International Union of Biochemistry and Molecular Biology (IUBMB): citrate synthase (EC 2.3.3.1), aconitate hydratase (EC 4.2.1.3),

isocitrate dehydrogenase (EC 1.1.1.41), 2-oxoglutarate dehydrogenase (EC 1.2.4.2), succinyl-CoA ligase (EC 6.2.1.4), succinate dehydrogenase (EC 1.3.99.1), fumarate hydratase (EC 4.2.1.2) and malate dehydrogenase (EC 1.1.1.37). All enzyme sequences were selected from ExPASy and Brenda data base searches (Artimo et al., 2012; Gasteiger et al., 2003; Schomburg et al., 2004). Oxidative TCA sequence collections were almost all Chain A. Domain (or Super-Kingdom) FASTA sets for the oTCA studies, some of these were additionally distinguished, e.g., Order, Class or Family, Firmicutes, alpha-, beta-, or gamma-proteobacteria, a flavoprotein subunit, 'mitochondrial', or by their cytoplasmic cellular location. We did not study distinguishing enzymes of the 'reverse' CO_2-fixing reductive TCA (rTCA) pathway: ATP-citrate lyase, 2-oxogluarate: ferredoxin oxidoreductase or fumarate reductase. The rTCA has been described as a degenerate or an idiosyncratic pathway (Becerra, Rivas, Garcia-Ferris, Lazcano & Peretó, 2014; Zubarev, Rappoport & Aspuru-Guzik, 2015). We, however, relative to the rTCA, recognize some contrary and persuasive emphases, also including the 'universality" of the TCA and being the most-extant biochemistry candidate for a 'self-maintaining cycle' (Fuchs, 2011; Smith & Morowitz, 2016).

The second set of this study were 5,545, almost entirely Chain A, sequences of the 33 geochrono-logically positioned Bacteria enzymes (Table 2) and as noted above were classified as either twenty Hadean-Archaean early-oldest ('Epoch I', 4,000-2,700 Ma) examples or thirteen late-newest ('Epoch III', 2,100 Ma-present). The few enzymes of the 'Epoch II' period (2,700-2,100 Ma) were not studied in order to minimally separate the 'Epoch I and III' chronology.

2.2. Recovered-amino acid or RAA: the unique data-tetrad

Our analyses involve multiple sequence alignments (MSA) and the determination of a unique multi-discriminatory value – a data-tetrad used throughout the study called a *Recovered-amino acid* or 'RAA' and escribed above.

'Epoch I' or "Epoch III' homologous enzyme FASTA sequence sets were analyzed by the MUSCLE multiple sequence alignment (MSA) program (Edgar, 2004). The MSAs and their consensus strings were conveniently viewed in JalView and edited rarely and only at excessively free-trailing sequence ends (Waterhouse et al., 2009). As reported earlier (Pollack et al., 2013), we assigned a conservation value for each MSA site using the ConSurf program (Ashkenazy et al., 2016) and a Euclidean distance to its respective C/AC with the Yasara program (Krieger & Vriend, 2014). *The critical feature of these analyses is the necessity of using the* same *Scaffold sequence as the query sequence in* both *programs and to first align conservation and distance data and secondarily relate the pair to the same MSA consensus modal site associated with their outputs and then finally align with their respective distance data.* We used Excel© to readily align, form and assemble the compound tetrad-data RAAs.

The ConSurf program determines for each MSA modal consensus site amino acid, a score that is a measure of its evolutionary conservation in a sequential scale of 1 to 9 quantitatively consecutive bins, called Conservation Zones or 'CZs'. In conformity with the ConSurf authors (Ashkenazy et al., 2016), we designated their program's *least -variable* evolutionary level as our 'most-conserved positions' that were typically nearest our C/ACs and identified as Conservation Zone 9 (ConZone 9 or CZ9), while their *most-variable* evolutionary set was rather named our 'least-conserved' or CZ1 set – those generally furthest (Å) from the C/AC.

The Euclidean distance measures extend from each Cα of each MSA consensus amino acid to the same atom (Anchor-atom), it was first necessary to demonstrate that an MSA site's conservation and distance values at respective C/ACs were *separately relatable* to amino acid % Occupancies. Assays with distance values without conservation involvements were compared to our RAA 'standard' analysis (with conservation) in appropriate pairs by their mutual concordances using Kendall's ranked correlation test (Wessa, 2017). *The concordances indicated high statistical notice that our conservation and distance measures respective to the C/AC were separately associated with comparable amino acid % Occupancies*.

2.3. LOcally weighted non-parametric polynomial regrESSions analyses (LOESS) of the RAAs of the oTCA pathway enzymes

The LOESS wire-frame plots add an additional 3D-perspective to 4,220 recovered oTCA RAA data. LOESS is a 3D-Graphic investigation of our oTCA RAAs that presents the contiguous graphic distribution of all averaged individual amino acid data. Further, our 'standard' analyses consider nine averaged conservation zones, whereas the LOESS wire-frames partition and describe the same data in a grid of about thirty conservation zones. The method is a computationally intensive locally weighted least squares non-parametric regression method (Cleveland, 1979). It fits a smooth curve or surface to a cloud of data and is a method often used to visualize trending. In our usage we examine at each site: the % Occupancy, the percentile distance from the C/AC and the percentile conservation score. Each figure is viewed as a contiguous sheet of approximately 780 '3D'-sites.

LOESS wire-frame projections of the individual un-averaged and un-grouped RAA data were constructed for the oTCA RAAs of alanine, aspartic and glutamic acids, glycine, isoleucine, leucine, lysine and valine, ranked by % occupancy as: GALIVDEK (R Development Core Team, 2011), using a statistics package with a span $\alpha = 0.75$. These compose the same dominant eight RAA amino set that we previously reported for glycolysis (*op. cit.*). Our LOESS plots of the estimated % Occupancy concentration are also color-coded by their Standard Errors (s.e.) (Cleveland & Grosse, 1991). The figure's standard errors are primarily determined by their sample sizes in that area of the figures and thus

the color-coding also offers additional insight on the relationship between the distance and our conservation percentiles for each plotted RAA-amino acid.

2.4. Geochronologically positioned 'Epoch' enzymes: studies estimating the relative ages of RAA amino acid protein architectures

2.4.1. Studies of the 20-Bacteria in 'early-oldest-Epoch I' (~4,000-2,700 Ma = million years ago*) and 13' late-newest-Epoch III' (2,100 Ma-present) periods by their 'nd-age' grouping compared either to each other or independently to the ranked putative Temporal Order of appearance of Amino acids on Earth*' (TOAE)

The Bacteria Scaffold enzymes of the 20 'early-oldest-Epoch I', described as Archaean and by their 'nd-age' values of 0.029-0.241 (Table 2A) were categorized using both the MANET data base (ver. 2) and the GTS (Cohen et al., 2013; Kim & Caetano-Anollés, 2010, 2011). The 'Epoch III's' collection of 13 Bacteria late-newest enzymes (Table 2B) had 'nd-age' group values of 0.675 to 0.905. Using Kendall's ranked correlation test (Wessa, 2017), we made various ranked comparisons of the Epoch cohorts RAA contents to each other as well as to Edward Trifonov's alphabet of the '*putative temporal order of appearance of* (the 20-biologic) *amino acids on Earth*' (GADVSPELTRIQNKHFCMYW = the 'TOAE' sequence) (Trifonov et al., 2000, 2004, 2009).

The 'nd-age' value model associates enzymes with independently reported *geological* time of origin and compares them to the enzyme's normalized phylogenetic node-distance from the hypothetical ancestral fold for example at the base of their phylogenetic tree. The nd-values structurally identify Domains by counting the number of their nodes from the base of phylogenetic trees to each Tree's leaf (Caetano-Anollés, Kim & Caetano-Anollés, 2012; Caetano-Anollés, Wang, Caetano-Anollés & Mittenthal, 2009; Kim & Caetano-Anollés, 2012; Kim, Mittenthal & Caetano-Anollés, 2006; Nath, Mtchell & Caetano-Anollés, 2014).

2.4.2. Statistical distinctions of RAA Epoch composition and structure by: 1) 'Epoch' Å^3- molar-volume and 2) ProteinVolume (PV) analyses

We calculated the RAA amino acid *compositional* Å^3-molar-volume of each enzyme in each Epoch set using both their published individual Å^3-volumes and their averaged % Occupancies in each enzyme. % Occupancy values were each multiplied by ofher reported Å^3 homologous values: the amino acid volume '*v*'-properties of Grantham (1974), the 'ACD MolVol' values used by Ilardo and Freeland (2014), or the Voronoi polyhedral procedure by Harpaz, Gerstein & Chothia (1994) or those described by Zamyatnin (1984). The homologous values were appropriately combined to obtain an average volume of each of the 20-RAA amino acids in each of the enzymes in order to estimate what we call the enzymes mean RAA

amino acid '*compositional* Å^3-molar volume' and with a standard deviation. Statistical comparisons of the means of such analyses between the Epoch sets or to the TOAE were *not* statistically significant, however, differences in their standard deviations were, noted later in Section *3.2.3.*

Other and statistically significant distinctions between the 20 'Epoch I' and 13 of 'Epoch III' enzyme sequences (Table 2) were found to involve their native and unfolded state ensembles and were recovered using the 'ProteinVolume' software available at <http://gmlab.bio.rpi.edu/> (Chen and Makhatadze (2015, 2017a, 2017b). For each protein listed in Table 2, we carried-out all-atom explicit solvent molecular dynamic (MD) simulations using identical settings as in our previous study (Chen & Makhatadze, 2017a). Briefly, molecular dynamics simulations of the protein in the native state were done using GROMACS v 4.5.7 with CHARMM27 parameter set (Brooks et al., 2009; Pronk et al., 2013). The electroneutrality of the system was achieved by adding Na^+ and Cl^- ions and included 0.1 molar excess NaCl. Simulations were performed at 1 bar of pressure, 300K. All proteins underwent 1,000 steps of energy minimization, 200 psec of constant volume equilibration, 200 psec of constant pressure equilibration, and 50 nsec of production simulation. A native ensemble of 50 structures was extracted from the production trajectory (1 structure/ns). All proteins remained folded throughout the simulation using stable all-atom RMSD with respect to the crystal structure as a criterion. Starting structures were obtains from PDB and missing side chains were 're-build' using MODELLER ver. 9.11 (Eswar et al., 2003). The unfolded state ensemble for each protein consisting of 1,000 structures was generated using 'TraDes' (Feldman & Hogue, 2002).

Prior to the volume calculations, the structures were processed as previously described (Chen & Makhatadze, 2015, 2017a). There, the statistically significant volume changes upon unfolding were identified as: $\Delta V_{Tot} = V_{Tot,U} - V_{Tot,N} = V_{Void,U} - V_{Void,N} + V_{Hyd,U} - V_{Hyd,N} = \Delta V_{Void} + \Delta V_{Hyd}$ where $V_{Tot,N}$ and $V_{Tot,U}$ are the total volumes of the native and unfolded states, respectively, $V_{Void,N}$ and $V_{Void,U}$ are the void volumes of the native and unfolded states, $V_{Hyd,N}$ and $V_{Hyd,U}$ are the hydration volumes of the native and unfolded states, ΔV_{Void} is the change in void volume upon unfolding and ΔV_{Hyd} is the change in the hydration volume upon unfolding. All final volume values are discussed in Section *3.2.5*.

Two-sided statistical significance for the difference between median values of various variables calculated for 'Epoch I' versus 'Epoch III' enzymes were examined using the Mann-Whitney test, differences in means were tested using Welch's test, and differences in variability were studied using Levene's test (Levene, 1960).

Analyses associated with Conservation Zone CZ9 and correlations with the TOAE were examined in a linear model with conservation zone and 'Epoch' as nominal explanatory variables and making zone-to-zone pairwise comparisons using Tukey's multiple comparisons correction (Hsu, 1996). We calculated the differences between the CZ9 correlation and the other zones' correlations as the

response variable and the conservation zone of comparison (treated nominally) and 'Epoch I' vs 'Epoch III' as the explanatory variables in the same type of analysis. Each calculation method is additionally detailed in their respective Results and Figure Legends. Residual plots were used to check the validity of model assumptions (data not shown). Also, a check on the inferential statements regarding the effect of 'Epoch' was made by comparing the p-value arising from our normal-theory model to an estimated p-value using 10,000 simulated data sets under the null hypothesis. These checks (data not shown) did not find any problems with the methods used. The reported Results reflect the inferential statements from the analyses using the normal distribution-theory linear model (Nelder & Wedderburn, 1972*)*.

3. Results

3.1. Distributions of Recovered Amino Acids

3.1.1. LOESS: Local regression - A graphically discriminating distributional analyses of the eight major oTCA RAAs

RAAs of the oTCA Bacteria in the eight LOESS examples of Fig. 1A (Panels A-D) and Fig. 1B (Panels E-H) (**Ed. both near here**) indicated with strong statistical support that glycine (Fig. 1A) is dominant in the most-conserved site zones and less so in the 'least'-conserved zones (Cleveland & Grosse, 1991). In glycine, as standard errors (s.e.) are a direct function of the number of sites with a specific distance and conservation level, the areas of greatest reliability is where their s.e. is in the 0.4-0.6 range and tend to fall along a surface-diagonal (the principal diagonal) from a region highly-conserved and close to the C/AC sites to least-conserved and distant from the C/AC sites. Conversely, glutamic acid a lysine are found least present in the most-conserved zones closest to the C/AC and most present in the 'least'-conserved zones furthest from the C/AC (Figs. 1B, 1D). Aspartic acid has low % Occupancy at middle conservation zones and high % Occupancy at the 'least'- and most-conserved zones (Figs. 1C). Alanine is widely distributed in LOESS analyses (Fig. 2E). Leucine, valine and isoleucine are concentrated in mid-regions (Figs. 2F, 2G, 2H). These LOESS surface images are in overall close accord with both the oTCA quantitative data Tables (shown later) and those we have previously reported in the LOESS analyses of glycolysis (*op. cit*.). Our investigation of the glycolysis pathway also found no appreciable difference between domains seen with conservation and distance to the C/AC. Note that the LOESS plots show where a particular amino acid is located, (given the amino acid), while the oTCA quantitative data (described later) provide information on what amino acids are prevalent at particular locations. The composite 3D-LOESS enzyme figures used ConSurf's outputs and the graphic program 'FirstGlance' in Jmol at <http://firstglance.jmol.org> (Martz, 2012).

3.1.2. Comparative % Occupancies of 20-RAA amino acid distributions in evolutionary conservation zones (CZs) of the oTCA enzymes in Archaea, Bacteria and Eukaryota

We compared the ranked oTCA RAA distributional data across three Domains to our previously reported values for the 20 ranked RAAs of glycolysis (*op. cit.*) and found close similarities (K-τ = 0.926 comparing all their sites, K-τ = 0.779 for only CZ9s and K-τ = 0.800 for only CZ1s) (Table 3). Of course, lower K-τ values would be expected when considering the smaller sample sizes and greater homogeneity in individual conservation zones, such as CZ9 or CZ1 examined individually. The cumulative data indicated a very high statistical concordance between glycolysis and oTCA ranked sequences of all RAA amino acids in the nine evolutionary conservation zones, K-τ = 0.926.

. We show in Fig. 2 the increasing averaged distances (Å) of 6,831 oTCA RAAs by their Domains from the most-conserved evolutionary consensus sites (CZ9) nearest their catalytic/active centers (C/AC) to their 'least'-conserved and most-distant sites (CZ1) is common to *all* Domain RAAs. We estimate that the average increasing (RMSE) (Å) distance of these non-overlapping oTCA analyses from their C/AC is approximately 1.05 Å per evolutionary conservation zone. The difference *between* the three Domains is marginally significantly different ($p \approx 0.062$). The nearly parallel lines show that the relationship between average distance and conservation appears to be a uniform all-Domain character-istic.

Table 4A, B, C, D show the ranked %Occupancies of all 20 biologic amino acids in each evolutionary conservation zone (CZ9 through CZ1) in each Domain. There the eight dominant and ranked RAAs in each CZ of each Domain are highlighted and the averaged Euclidean distances (Å) of the Cαs of all 20-RAAs in each CZ to their respective C/AC's Anchor-atom are included. In Table 4D are the pooled oTCA RAA % Occupancies of the study.

We found that glycine (G) (blue) is dominant at the most-conserved evolutionary conservation zone sites (i.e., CZ9 and CZ) and on average closest to Scaffold's functional Anchor-atom. Glycine (G) % Occupancies values (blue) *decrease* in moving to less-conserved sites (CZ1) that are furthest from their C/ACs. Conversely, concentrations of polar lysine (K) and glutamic acid (E) (both green) *increase* in moving to the same less conserved CZ3-1 sites farthest from the C/AC. A mutual K/E relationship has been reported (Manavalan & Ponnuswamy, 1977). Hydrophobic non-polar leucine (L), valine (V) and isoleucine (I) maintain elevated concentrations in mid-regions (CZ8 to CZ3) (all orange), as frequently reported (e.g., Arunachalam & Gautham, 2008). RAA polar aspartic acid (D, green) is not confidently localized. Alanine was less concentrated at mid-CZ6-CZ3 sites than at both the highest and lowest conservation sites; this same RAA alanine trend was not found studying glycolysis enzymes (*op. cit*.). The RAAs of non-polar leucine, valine and isoleucine were dominant in mid-conservation zones.

The Tables 4A, B, C and summated in D show Results of all the oTCA data in each Domain by % RAA Averaged Occupancies each of all nine CZs. The CZ RAA data are accompanied by their averaged

distance (Å) to respective C/ACs. The Tables identify the RAA mixtures in each CZ and highlight the eight dominantly ordered RAAs of the study: ALEVGDIK. We would characterize the structural consistency as a 'pattern'. The aggregations are diffusely shaped; the % Occupancy of individual RAAs are noted in the CZ columns and as 'diffuse-aggregations' (DAG) are discussed later (Section *4.4*.). Graphic presentations of such aggregations are shown for pyruvate kinase in Fig. 3 and in the Supplementary File 2 for three glycolysis enzymes at https://www.asc.ohio-state.edu/pearl.1//pgp/ (Pollack et al., 2013); an updated version is available from the authors.

We found progressive differences between the RAA % Occupancy rankings in succeeding conservation zones and to their rankings in the columns labelled 'Averaged % Occupancies of all RAAs' of Tables 4 and 5. For example, K (lysine) and E (glutamic acid) % Occupancies are found most dominant in Domain CZ1s but are more distributed in the all-CZ-RAA summary column (Section *4.4.)*.

3.1.3. Comparisons of oTCA, glycolysis and the 33-Bacteria set RAA contents in all evolutionary conservation zones to TOAE: The dominance of evolutionary conservation zone (CZ9)

Figure 4 graphically shows the average ± s.e.m. of the Kendall's tau (K-τ) correlations between our RAA-amino acid ranked K-t data of oTCA, glycolysis and the 33 geochronologically assigned (but here undistinguished between 'Epoch I and III') Bacteria enzyme sequences (Table 2). The ranked % Occupancy values of each conservation zone were correlated with *the putative temporal ordering of their appearance on Earth* ('TOAE') (Trifinov, 2009). We again found that the closest correlation to TOAE was at the CZ9 site: the difference after correcting for multiple comparisons between CZ9 and each other zone from CZ7 to CZ1 was highly significant, $p < 0.005$. The plots show a generally decreasing relationship between the RAA amino acid distributions and the TOAE ordering from the most-conserved sites, typically nearest the C/AC, toward less conserved sites. *The K-τ correlation with CZ9 was higher than found using all respective CZ8 to CZ2 collections. The relationships are very similar for oTCA, the 'Epoch I and III' Bacteria enzymes and for those we reported for glycolysis (op. cit.).*

3.2. The geochronologically distinguished 33-Bacteria enzymes of 'Epoch I' and 'Epoch III'

3.2.1. Relationships of RAA distributions of 'Epoch I' and 'Epoch III'

We found between "Epochs I and III' (Tables 5A and 5B) a non-random significant *distributional uniformity* between their ranked % Occupancies (K-τ = 0.849, p = 0.0001). These composite RAA trends are essentially identical to those we found in the oTCA Domains (Table 4) and our previous glycolysis studies (*op. cit*.). The oTCA Domain similarities are in some contradiction to the literature and are discussed in Section *4.1.1*. After combining all data in both Epochs and re-calculation, the ranked average RAA % Occupancy of the combined Tables 5A and B is ALGEVDRITKPSFNY-QHMWC.

3.2.2. Average RAA residue count per sequence

We found that 8,764 unique sequences representing the 20 enzymes identified as 'Epoch I' (~4,000-2,700 Ma) (Table 2A**)** had an average (± std dev) number of amino acid residues per enzyme sequence = 439.2 ± 319.9, with a median number of residues per enzyme sequence of 357 (the median being more representative of the lengths in order to offset the effect on the average of a few outliers). The 3,489 sequences representing 13 'Epoch III' (2,100 Ma-present) enzymes (Table 2B) had a lower average residues/enzyme sequence of 290.7 ± 192 and a median number of residues/enzyme sequence of 281. The comparisons of these Epoch means or their medians were *not* statistically significant though they do trend in support of our characterization of the nd-age derived chronological distinctions of 'Epoch I' and 'Epoch III' discussed below.

3.2.3. Distinguishing 'Epoch I' from 'Epoch III': analyses of 'Epoch I' and 'Epoch III' the critical differences from the CZ9 catalytic/active centers of all "Epoch I' or 'Epoch III' enzymes to all their surrounding RAA Cαs

We made a further analysis of the strong relationship between our ranked amino acid distribution in CZ9 and the TOAE by examining the drop-off in the K-τ correlations as we go from CZ9 to the other zones. Figure 5 shows the results of this analysis of the distinctiveness of the most conserved sites in 'Epoch I' versus 'Epoch III' by recording the arithmetic difference (= difference value or 'DV') between the CZ9 K-τ correlation with the TOAE and the similarly calculated K-τ correlation with the TOAE for each of the other conservation zones (CZ). Figure 5 displays two approximately parallel curves with average DV's for the 20 early-oldest 'Epoch I' enzyme set that are higher than for the 13 late-newest 'Epoch III' enzymes ($p \approx 0.007$). Thus, CZ9 is seen to be more distinguished from the other zones in 'Epoch I' enzymes compared with 'Epoch III' enzymes.

To provide an example of the calculations at the '9vs8' 'Epoch I' site for one of the 20 'Epoch I' enzymes, acetyl-CoA synthase (ACS), we offer that a K-τ value of 0.631 is recovered between TOAE and its CZ9 amino acid distribution and a K-τ value of 0.591 between TOAE and its CZ8 amino acid distribution. The difference (DV) is 0.631-0.591 = 0.040 for ACS. This ACS value was averaged with those similarly obtained for the same CZ category using the other 19 'Epoch I' enzymes to yield the studies smallest averaged (n=20) DV value at '9vs8', i.e., there the black circle = 0.105. The bars in the figure show the enzyme-to-enzyme standard errors. Analogous calculations were made for each of the 20 'Epoch I' enzymes at all CZ sites and then entirely again for the 13 'Epoch III' enzymes of Fig. 5.

The DVs on the y-axis are the plain-arithmetic differences between *two* 'paired' K-t values and the x-axis gives 'Conservation Zone-distance values' that first involve CZ9 and CZ8 and then replace

CZ8 by succeeding CZs to CZ1, thereby more critically indicating *de*creasing evolutionary conservation and distance from the functional C/AC of their CZ9. Importantly, the averaged *distance*-across the conservation zones is 2.92 Å ± 0.77 is higher for 'Epoch I' than 'Epoch III' enzyme sets, an approximate ~9 % increase.

The DV data in Fig. 5 show that increasing Å-distance from respective C/ACs to all CZ9s are concomitant with averaged evolutionary conservation values proceeding away (Å) from CZ9 and that the early-oldest-'Epoch I' enzymes at the most conserved areas (CZ9) have an RAA-amino acid distributions that are most associated with the order of their appearances on Earth.

3.2.4. Distinguishing analyses of 'Epoch I' and "Epoch III': Distinctions found in enzyme RAA amino acid 'compositional' Å^3-molar volumes

In comparisons of enzyme *amino acid-compositional* Å^3-molar volume data between Epochs we found that their means were *not* statistically significant. However, importantly, there was a much greater and significant result when comparing their *standard deviations*: 36,014 ('Epoch I') versus 20,119 ('Epoch III'), p ≈ 0.041 (Levene's test). These results explicitly suggested a structural distinction of the 'Epoch III' enzymes perhaps recognizing a greater residue 'compactness' or 'consistency'.

3.2.5. Distinguishing analyses of 'Epoch I' and "Epoch III': Volumetric features in geochronologically distinguished enzymes recovered by 'ProteinVolume' analyses

The statistically significant values in amino acid compositional Å^3-molar volumes (Section *3.2.4.*) between 'Epoch I and III' enzymes suggested that differential contributions to volume changes were evolutionarily conserved and may be measureable. Using the 'ProteinVolume' program we found statistically significant differences between 'Epoch I and III' enzymes (Chen & Makhatadze, 2015, 2017a, 2017b).

The mean (64,562 Å^3 versus 37,124, p = 0.02) and median (52,590 Å^3 versus 38,360, p = 0.02) volume of the native state (V_{native}) of older 'Epoch I' enzymes were significantly higher than 'Epoch III'. Likewise, the 'Epoch I' enzymes also have significantly higher mean (16,424 Å^3 versus 9,054, p = 0.03) and median (13,500 Å^3 versus 4,093, p = 0.02) void volumes in the native state? (V_{Void}).

We also found that 'Epoch I' enzymes exhibited greater variability overall with respect to the volume change upon unfolding (ΔV_{Tot}) *where the standard deviation was 5.7 fold greater for the 'Epoch I' enzymes that for those of 'Epoch III'. Further, the earlier 'Epoch I' enzymes have significantly higher mean volumes (by Welch's test) and median volumes (by the Mann-Whitney test) for each of the VSE, void and native measures.Of all 33 Bacteria examined, the six or seven largest volumes were from the early-oldest 'Epoch I' group while 4 of the 5 smallest were from the late-newest 'Epoch III' group. The*

significant volumetric differences distinguishing the enzymes of 'Epoch I' from 'Epoch III' are also interpreted as support for their separate geochronological assignments based on nd-distance values (Cohen et al., 2013; Kim, Mittenthal & Caetano-Anollés, 2006).

The amount of void volume change we report and the other change-differences upon unfolding have important implications as they reflect the relative efficiency of packing of the residues in the core of the native state. The authors Gilson, Marshall-Christensen, Choi & Shakhnovich (2017) indicated that lower packing density of the native protein, i.e., larger in magnitude volume changes upon unfolding, as in our older 'Epoch I' enzymes provides for more rapid evolution of three-dimensional structures – our Result trend in that direction. Further, the authors offer that the driving force for such structural evolution may be an increasing compactness (Sections *3.2.4., 3.2.5.*) - representing increasing thermodynamic stability. Packing interactions, particularly between non-polar residues, as ILV, have been shown to have paramount importance where better packing leads to higher protein stability (Dill, 1990; Gerstein, Sonnhammer & Chothia, 1994; Liang & Dill, 2001; Makhatadze & Privalov, 1995).

4. Discussion

We temporally characterized enzyme sequences by special nd-distance phylogenetic analyses and geologic periodicity as either Hadean-Archaean or Archaean or 'contemporary', i.e., 'Epoch I or 'Epoch III'. The assignment of the Hadean-Archaean at ~4,000 Ma (= 4.0 Ga) (Table 2) is taken as an approximate geochronological "boundary" of our temporal study (Cohen, Finney, Gibbard & Fan, 2013, McGuiness, 2010; Sleep, 2010). We accept that the age of Life's 'beginning' is characterized as 'absolutely beyond controversy' within 3,600-2,800 Ma (Russell, 2016). We would consider an older age based on the isotopic record of life from 3,800 Ma deduced by ^{13}C-^{14}C isotope fractionation studies indicative of carbon-fixation (Schildowski, 1988) and reservedly studying 4,100 Ma zircons (Bell, Boehnke, Harrison & Mao, 2017).

We have used tryptophan, as the most-recent and consequential evolutionary sign-post. It is reported as the last-recovered amino acid at about ~3,420 Ma and was considered the most recent addition to a formative-genetic code in a contemporary 'protein world' (Fournier & Alm, 2015). We imagine that before 3,420 Ma there were tolerable photosynthetic microbial environments perhaps with ribozymes, an expanding early genetic code and functional catalytic antecedents of tryptophan-poor or tryptophan-less pre-protein(s), representing, e.g., a Trp-less 'pre-acetyl-CoA synthase' sequence of 'Epoch I' that would have an nd-age value less than 0.029 (Table 2).

4.1. The common distributional homology of RAAs in the oTCA of Archaea, Bacteria and Eukaryota and in the geochronologically separated 33-Bacteria

4.1.1. The oTCA

Our determinations of the general similar Domain oTCA RAA-amino acid distributions (% Occupancies) are in apparent *dis*agreement with other reports that distinguish the *un*characterized amino acid contents of Archaea, Bacteria and Eukaryota (e.g., Bogatyreva, Finkelstein & Galzitskaya, 2006; Karlin et al., 2002; Pe'er et al., 2004; Zaia, Zaia & de Santana, 2008). Such reports, using current data bases, identify and estimate the distribution of amino acid residue contents in unstructured Archaea, Bacteria or Eukaryota and have agreeably found distinctive Domain differences.

In response, our data is not methodologically comparable. We study oTCA Domain content after initial separations using amino acid residues characterized by their conservation values and Euclidean proximities (Å) to their respective functional centers (RAAs). The Results are considered revealing, additive and complimentary (Table 4).

In an our glycolysis report, as here, but there also using logistic regression methods and 4,645 curated sequences of 10 glycolytic enzymes and interconnected data of RAA amino acid occupancy and distances to respective C/AC's, *no* statistically significant Domain differences were found between their amino acids RAA % Occupancy's after controlling for differences in conservation (Pollack et al., 2013). Possible exceptions were more cysteine and lysine in Eukarya and more aspartic acid in Archaea.

4.1.2. The 33-Bacteria

The differences between the 33-Bacteria of 'Epochs I and III' in their RAA *distributional* content and ranked % Occupancy (Section *3.2.1.*), as those in Domains of oTCA (Section *4.1.1.*), were *not* statistically different – rather, highly concordant, however, in other comparisons that we studied and have presented they are, in Sections: *3.2.2*) residue number per sequence, in *3.2.3.)* the variable distances from their functional centers in CZ9s, and the degree to which CZ9 is distinguished by its association with the TAOE, in *3.2.4.*) RAA compositional $Å^3$-molar-volumes and in *3.2.5.)* analyses between their native volumes, void volumes and unfolded volume status'.

A particularly relevant report concerning RAA distributional conformity involves the 'HP-model' of Guseva, Zuckermann & Dill (2017). The HP-model describes relatively stable random short-chain hydrophobic-polar sequences capable of folding and compact collapse in water that can elongate, effect the elongation of similar molecules and act as primitive catalytic protein-like polymers. *We suppose that our 33-Bacteria may start as similar primitive amino acid sequences in an early generative period of 'Epoch I', close to, within or even prior to 4,000-3,800Ma, and may be functional antecedents to the catalytic/active centers we have described.*

4.2. CZ9 – Enzyme Conservation Zone 9: the most-conserved and functional centers as earliest enzymatic progenitors?

Of all nine evolutionary conservation zones in all enzymes, our CZ9 sets are most concordant to the 'temporal appearance of amino acids on Earth' (TOAE) (Trifonov, 2009). Compositionally, functionally and structurally the CZ9 loci may represent a very early prebiotic evolutionary conservation zone. The CZ9 oTCA examples generally have both fewest residues of any CZs and are most poor in tryptophan (Section 3.2.2.). Further, the ranked RAA contents of the CZ9s per se of geochronologically positioned early-oldest 'Epoch I' collections have higher ranked K-τ-correlations to the 'putative temporal order of appearance of (the 20-biologic) amino acids on Earth' (TOAE) than those of the late-newest ranked 'Epoch III' cohort.

We also found that of the nine oTCA conservation zones, the CZ9 locus nearest the C/AC contains functional catalytic sites that is most-conserved and and dominated by glycine. Glycine has flexible conformational properties and the 'first' peptides were glycine rich (Guimarães, 2011; Yan & Sun, 1997). Glycine is a more effective H-bond acceptor with more configurational entropy than other side-chain amino acids and contributes to the maintenance of stability and mutational 'robustness' (Bloom et al., 2006; Duax et al., 2009; Matthews et al., 1987). Further, a glycine rich motif is the most conserved element among all protein sequences and fixed at the earliest stages of protein evolution (Gorbalenya & Koonin 1989; Koonin 2012). ~~G~~lycine's significant and relevant characteristics were reported in a similar study of glycolysis enzymes (Pollack et al., 2013).

4.3. 'Alphabets': the ranked % Occupancy distributions of 20-biologic amino acids

Two published amino acid alphabets are of particular relevance to our study. One study emphasizes amino acid size, charge and hydrophobicity by separation into two unranked 10-alphabet collections of the 20-biologic amino acids (Ilardo & Freeland, 2014). Their ten early-oldest set is ADEGILPSTV and was described by the authors as available to earliest life while the 10 remaining late-newest residues as later derived. Our earliest-set of 10 is ALEVGIKDTP (Table 5A).

The second report described a specifically ranked set of each of the 20-biologic amino acid alphabet based on the reported contents of experimental observables in non-biological contexts, e.g., meteorites, 'icy grains', atmospheric, hydrothermal and some chemical syntheses (Higgs & Pudritz, 2007, 2009). The authors indicated that the top-ten prebiotic amino acids ranked in decreasing abundance were GADEVSILPT. We note their similar presence and relative positions to those of the TOAE, i.e., GADRPTLSEV. Further, in analogy, their order was considered thermodynamically predictable by their relative increasing standard free energies of formation (ΔG_f^0), i.e., their 'earlies', may be considered thermodynamically 'less-costly', while the late-newest amino acids are 'more-costly'. The authors also

hypothesized that the10 amino 'earlies' were frequent at the time of the origin of the genetic code and were the first coded amino acids by having the smallest ΔG_f^o.

4.4. Common 'Diffuse-Conservation-Aggregations' ('DAG'): the localized and conservationally ranked abundances of dominant RAAs

In all Domains, we consistently found (Section *3.1.1*) structurally and evolutionary conserved collections with dominance of particular RAA amino acids as 'diffuse-aggregations' (DAGs) (Fig. 3 and Supplementary File 1). DAGs were found in analyses of 51 enzymes (Tables 4 and 5 and reported in Pollack et al., 2010, 2013). *These results bear on the characterization of specific amino acid dominances in protein structure that we find conditional on their specific evolutionary conservation value and proximity to respective functional centers.*

The apparently central locations of isoleucine, leucine, valine DAGs are analogized to the ILV side chain cores in the presence of intrinsically disordered regions described by Kathuria, Chan, Nobrega, Özen & Matthews (2015) (Tables 4 and 5). These authors reported that large ILV hydrophobic clusters would impede solvent exchange and serve as cores of stability impeding water penetration and effecting hydrogen bond-networks. In other studies of intrinsically unstructured proteins: isoleucine, valine and leucine have been described as structural order-promoting amino acids (Campen, Williams, Brown, Meng, Uversky, et al., 2008; Mohan, Oldfield, Radivojac,Vacic, Cortese, et al., 2006).

4.5. The direct relationships of distance and conservation measures to crystal structures

We suggest that our MSA consensus-conservation calibrated RAA-distribution studies have a strong supportive relationship to those of A. Mittal and co-workers. They determined the amino acid incidences and coordinates of all the Cαs backbone crystal structures in various folded proteins including three of the 26 oTCA cycle PDB-enzyme structures that we also studied: PDBs: 1l5j, 8acn and 1nek (Table 1) (Mittal & Jayaram, 2011). They reported that folded proteins were characterized by spatially well-defined, distance dependent and universal 'neighborhoods*'* of common amino acid members. This opinion is referable to our own findings of 'aggregations' of ranked RAA-amino acids in all the enzyme sets that we studied and our inference of internal organized collections of amino acids. Also notable in their studies were the implications that protein folding is primarily influenced by the frequency of occurrence of the amino acids in the primary sequence, a point we also consider is an association to our distributional studies (Mittal, Jayaram, Shenoy & Bawa, 2010). We ranked their reported distributions of 20-biologic amino acids in 3,718 folded various proteins of Archaea, Bacteria and Eukaryota as LAGVEKSIDT-RPNFQYHMCW. In order to compare this sequence to that of our oTCA analyses we re-calculated *all* our 'averaged % Occupancies' oTCA data (Table 5) but *without* association to either conservation or

distance. The unmodified 'raw' 20-amino acid % Occupancy data of all 2,844 oTCA enzyme sequences is: AGLEVIDKPTRSNFYQHMCW. The K-τ derived concordance of the two 20-ranked amino acid enzyme alphabets was very high at K-τ = 0.863 of 1.0.

4.6. 'Indirect' Epoch assignments of the oTCA enzymes

Indirect clues of the relevance of nd-relationships and oTCA 'Epoch' distinctions of Tables 2 were found after examination of the eight oTCA enzymes of Table 1 using MANET data base's Enzyme Commission and PDB identifications and metabolic notations. Eight oTCA enzyme sequences supported our geochronological positioning by having nd-values with an average of 0.171 ± 0.116, that is, within the ~4,000-2,700 Ma Archaean nd-range of Table 2's 'Epoch I' 0.029-0.241 values, they were: three aconitate hydratases, two succinate-CoA ligases, two succinate dehydrogenases, two fumarate hydratases and a malate dehydrogenase. Furthermore and agreeably compatible, acetyl-CoA synthase (nd = 0.029) and ATPase (nd = 0.044) are listed first in Table 2, both are considered as among Earth's very earliest enzymes (Adam, Borrel & Gribaldo, (A2018); Fuchs, 2011; Nitschke & Russell, 2013; Weiss et al., 2016).

Although, eight sequences of five oTCA enzymes are insufficient they putatively support our geochronological positioning of oTCA and the usefulness of nd-value scalar analyses. At the other more 'contemporary' 'nd-extreme' of Table 2 in Part B, we consider that both enzymes RuBisCo (nd = 0.645) and SOD (nd = 0.905) as 'Epoch III' (~2,100 Ma – present) are also temporally compatible examples with their reported major presence and rise of atmospheric oxygen during Earth's Great Oxygen Event at ~2,322 Ma (Bekker et al., 2004; Slesak, Slesak & Kruk, 2017).

*5. C*onclusion: in retrospect -molecular structural changes and our hypothetical 'trace' of 'metabolism's progressive emergence'

Although, the lysine (K) and glutamic acid (E) rankings in 'Epoch I' and 'Epoch III' at, e.g., in CZ1 (Table 5) when compared to the same RAAs in the 'all-CZ-RAA summary column are strikingly different, they alone do not satisfactorily support our research hypothesis of the likely existence of chemical and physical signatures that temporally characterize or 'trace' any continuous passage of 'metabolism's progressive emergence'. Distributional data, as well in Table 4, likewise emphasize the contribution of evolutionary conservation values and proximities (Å) to respective C/ACs in distinguishing RAA structural positions and the non-random distribution of dominant enzyme amino acids. The 'trace', we now believe, may not be clearly resolved in comparisons we have made as for example, using distributional data that may insufficiently include the earliest examples of amino acid aggregations. Obviously, as one option, a concentration of more early-oldest 'Epoch I' nd-age assigned

sequences, either Hadean-Archaean or Hadean, are needed to study for an earliest evolutionary role by our or similar distributional analyses – obtaining a sufficient number of such Hadean or Archaean examples does not appear to us likely.

Analyses specifically involving the evolutionary zone (CZ9) revealed unanticipated evolutionary 'traces'. The average distances (Å) of all individual CZ9's consensus RAAs Cα to their respective Scaffold's Anchor-atom were closer than those from the RAA collections of any other CZ of any other enzyme. Also, the ranked RAA amino acid % Occupancies of the CZ9 early-oldest 'Epoch I' amino acids were significantly more concordant to the TOAE than to those ranked CZ9 of the late-newest 'Epoch III' enzymes that suggests or identifies a contiguous evolutionary 'trace' between "Epoch I and III'. In additional comparisons and with more conclusiveness, we found that the greater differences of 'Epoch I' in our 33-Bacteria studies of $Å^3$-volumes and the variability of their standard deviations were critical parameters. We calculated the averaged volumetric features of individual amino acids in all Epoch enzyme sequences and found by three consequential and statistically significant volumetric distinctions between 'Epoch I and III' enzymes (Section *3.2.5.*). 'Epoch I' Bacteria enzyme sequences occupied greater $Å^3$-volumes and greater void-volumes in the native state and greater variability in volume changes upon unfolding. *We interpret the temporal linked stabilizing physical-volumetric distinctions between the Bacteria enzymes of 'Epoch I' and 'Epoch III' as concomitant geochronological and evolutionary evidences in of 'traces' of 'metabolism's progressive emergence'.*

Acknowledgements

The study is dedicated to the memories of Professors Robert Cawrse Cleverdon, Ph.D., University of Connecticut, Storrs, CT and Harold J. Morowitz, Ph.D., George Mason University, Fairfax, VA for their friendship, encouragement and advice. We most appreciatively thank MEMSP and BP (JDP) and the librarians of 'Iliad' at The Ohio State University Interlibrary Services for absolutely essential help and Professor Gustavo Caetano-Anollés, University of Illinois, Urbana, IL for very generously sharing unpublished nd-values of some enzymes identified(†) in Table 2. We have been and are continuously encouraged by many highly motivating reports (e.g., Romero, Rabin & Tawfik, 2016) and are indebted to numerous, thoughtful and challenging personal opinions and evidences of others that have been crucial to our study of the evolution of prebiotic-enzymes.

Disclosure statement

The authors declare they have no competing interests and have read and approve the final manuscript.

Funding

The work was supported by the US National Science Foundation grants numbered CHEM-1506468 and CHEM-1803045 (to G.I.M).

References

Adam, P. S., Borrel, G. & Gribaldo, S. (2018) Evolutionary history of carbon monoxide dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes. *Proceedings of the National Academy of Sciences, USA.* doi:10.1073/pnas.1716667115.

Ævarsson, A., Seger, K., Turley, S., Sokatch, J. R. & Hol, W. G. J. (1999). Crystal structure of 2-oxoisovalerate and the architecture of 2-oxo acid dehydrogenase multienzyme complexes. *Nature Structural Biology, 6*, 785-792.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov, D., Liechti, R., Moretti, S., Mostaguir, K., Redaschi, N., Rossier, G., Xenarios, I. & Stockinger, H. (2012). ExPASy: SIB bioinformatics resource panel. *Nucleic Acids Research, 40,* issue W1, July, W597-W603. doi:10.1093/nar/gks400

Arunachalam, J. & Gautham, N. (2008). Hydrophobic clusters in protein structures. *Proteins, 71*, 2012-2025.

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T. & Ben-Tal, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research, 44*. doi:10.1093/nar/gkw408

Becerra, A., Rivas, M., Garcia-Ferris, C., Lazcano, A. & Peretó, J. (2014). A phylogenetic approach to the early evolution of autotrophy: the case of the reverse TCA and the reductive acetyl-CoA pathways. *International Microbiology, 17*, 91-97.

de Beer, T. A. P., Laskowski, R. A., Duban, M-E., Chan, A. W. E., Anderson, W. F. & Thornton, J. M. (2013). *LigSearch*: a knowledge-based web server to identify likely ligands for a protein target. *Acta Crystallographica, D69*, 2395-2402.

Bekker, A., Holland, H. D., Wang, P.L., Rumble III, D., Stein, W. J., Hannah, J. L., Coetzee, L. L. & Beukes, N. J. (2004). Dating the rise of atmospheric oxygen. *Nature, 425*, 117-120.

Bell, E. A., Boehnke, P., Harrison, M. & Mao, W. L. (2015). Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. *Proceedings of the National Academy of Sciences, USA.* doi:10.1073/pnas.1517557112

Berman, H. M, Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research, 28*, 235-242. doi:10.1093/nar/28.1.235

Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences, USA, 103*, 5869–5874.

Bogatyreva, N. S., Finkelstein, A. V., Galzitskaya, O. (2006). Trend of amino acid composition of proteins of different taxa. *Journal of Bioinformatics and Computational Biology*, *4*(2), 597-608.

Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., Karplus, M. (2009).CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry, 30*, 1545-1614.

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

Caetano-Anollés, G., Wang, M., Caetano-Anollés, D. & Mittenthal, J. E. (2009). The origin, evolution and structure of the protein world. *Biochemical Journal, 417*, 621-637.

Campen, A., Williams, R. M., Brown, C. J., Weng, J., Uversky, V. N. and Dunker, A. K. (2008). TOP-IDP-scale: A new amino acid scale measuring propensity for intrinsic disorder. *Protein and Peptide Letters, 15*(9), 956-963.

Chen, C. R. & Makhatadze, G. I. (2015). Protein Volume: calculating molecular van der Waals and void volumes of proteins. *BMC Informatics, 16,* 101. doi:10.1186/s12859-015-0531-2

Chen, C. R. & Makhatadze, G. I. (2017a). Molecular determinant of the effects of hydrostatic pressure on protein folding stability. *Nature Communications, 8,* 14561. doi:10.1038/ncommuns14561

Chen, C. R. & Makhatadze, G. I. (2017b). Molecular determinants of temperature dependence of protein volume change upon unfolding. *Journal of Physical Chemistry B, 121*, 8300-8310.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*, 829-836.

Cleveland, W. S. & Grosse. E. (1991). Computational methods for local regression. *Statistics and Computing, 1*, 47-62.

Cohen, K. M., Finney, S. C., Gibbard, P. L. & Fan, J.-X. (2013). The ICS International Chronostratigraphic Chart. *Episodes 36*(3), 199-204. (http://www.stratigraphy.org (2017 revisions), there see, 'Chart Time Scale', also @: <http://www.stratigraphy.org/ICSchart/ChronostratChart2013-01.pdf>).

De Duve, C. (1991). *Blueprint for a Cell: The Nature and Origin of Life.* Neil Patterson Publishers, Burlington, North Carolina.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, *29,* 7133-7155.

Duax, W. L., Huether, R., Pletnev, V., Umland T. C. & Weeks, C. M. (2009). Divergent evolution of a Rossmann fold and identification of the oldest surviving ancestor. *International Journal of Bioinformatics Research and Applications, 5*, 280–294.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high through-put. *Nucleic Acids Research, 32*(5), 1792-1797.

Eigen, M. (1992). *Steps Towards Life*. Oxford University Press, Oxford.

Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V. A., Pieper, U., Stuart, A. C., Martı-Renom, M. A., Madhusudhan, M. S., Yerkovich, B. & Šali, A. (2003). Tools for comparative protein structure modeling and analysis. *Nucleic Acids Research*, *31,* 3375-3380.

Feldman, H. J. & Hogue, C. W. (2002). Probabilistic sampling of protein conformations: new hope for brute force? *Proteins, 46*, 8-23.

Fournier, G. P. & Alm, E. J. (2015). Ancestral reconstruction of a pre-LUCA aminoacyl-tRNA syntetase ancestor supports the late addition of trp to the genetic code. *Journal of Molecular Evolution, 80*, 171-185.

Frank, R. A. W., Price, A. J., Northrup, F. D., Perham, R. N. & Luisi, B. F. (2007). Crystal structure of the E1 component of the Escherichia coli 2-oxoglutarate dehydrogenase multienzyme complex. *Journal of Molecular Biology, 368*, 639-651.

Fraser, M. E., James, M. N. G., Bridger, W. A. & Wolodko, W. T. (2000). Phosphorylated and dephosphorylated structures of pig heart, GTP-specific succinyl-CoA synthetase. *Journal of Molecular Biology, 299*, 1325-1339

Fry, I. (2000). *The Emergence of Life on Earth.* Rutgers University Press, New Brunswick, New Jersey.

Fuchs, G. (2011). Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annual Reviews of Microbiology, 65*, 631-658.

Furnham, N., Holliday, G. L., de Beer, T. A., Jacobsen, J. O. B., Pearson, W. R. & Thornton, J. M. (2014). The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Research, 42*, D485-D489.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. & Bairoch, A. (2003). ExPASy: the proteomic server for in-depth protein knowledge and analyses. *Nucleic Acids Research, 31*, 3784-3788.

Gerstein, M. 1, Sonnhammer, E. L. & Chothia, C. (1994).Volume changes in protein evolution. *Journal of Molecular Biology, 236*, 1067-1078.

Gilson, A. I., Marshall-Christensen, A., Choi, J.-M. & Shakhnovich, E. I. (2017). The role of evolutionary selection in the dynamics of protein structure evolution. *Biophysical Journal, 112, 7*, 1350-1365. doi: 10.1016/j.bpj.2017.02.029

Gorbalenya, A. E. & Koonin, E. V. (1989). Viral proteins containing the purine NTP-binding sequence pattern. *Nucleic Acids Research, 17*, 8413–8439

Granick, S. (1957). Speculations on the origin and evolution of photosynthesis. *Annals of the New York Academy of Sciences 69,*292-308.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science 185*, 862-854.

Guimarães, R. C. (2011). Metabolic basis for the self-referential genetic code. *Origins of Life and Evolution of the Biosphere, 41,* 357–371.

van der Gulik, P., Massar, S., Gilis, D., Buhrman, H. & Rooman, M. (2009). The first peptides: the evolutionary transition between prebiotic amino acids and early proteins. *Journal of Theoretical Biology, 261*, 531-539.

Guseva, E., Zuckermann, R. N. & Dill, K. A. (2017). Foldamer hypothesis for the growth and sequence differentiation of prebiotic polymers. *Proceedings of the National Academy of Sciences, USA*. doi:10.1073/pnas.1620179114

Harpaz, Y., Gerstein, M. & Chothia, C. (1994). Volume changes on protein folding. *Structure, 2*, 641-649.

Higgs, P. G. & Pudritz, R. E. (2007). From planetary disks to prebiotic amino acids and the origin of the genetic code. In Ralph E. Pudritz, Paul G. Higgs and Jonathan R. Stone *(*eds.*), Planetary Systems and the Origin of Life* (pp. 62-88). Cambridge, Cambridge University Press.

Higgs, P. G. & Pudritz, R. E. (2009). A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology, 9*, 483-490.

Hsu, J. C. (1996). *Multiple Comparison Theory and Methods*. Chapman & Hall, New York.

Ikehara, K. (2016) Evolutionary steps in the emergence of Life deduced from the bottom-up approach and GADV hypothesis (top-down approach). *Life, 6*, 6. doi:10.3390/life6010006

Ilardo, M. A. & Freeland, S. J. (2014). Testing for adaptive signatures of amino acid alphabet evolution using chemistry space. *Journal of Systems Chemistry, 5*(1), 1-9.

Ingles-Prieto, A., Ibarra-Molero, B., Delgado-Delgado, A., Perez-Jimenez, R., Fernandez, J. M., Gaucher, E. A., Sanchez-Ruiz, J. M. & Gavira, J. A. (2013). Conservation of protein structure over four billion years. *Structure 21*, 1690-1697.

Jha, A. N., Vishveshwara, S. & Banavar, J. R. (2010). Amino acid interaction preferences in proteins. *Protein Science, 19*, 603-616.

Karlin, S., Brocchieri, L., Trent, J., Blaisdell, B. E. & Mrázek, J. (2002). Heterogeneity of genome and proteome content in Bacteria, Archaea, and Eukarya. *Theoretical Population Biology, 61*, 367-390.

Kathuria, S. V., Chan, Y. H., Nobrega, R. P., Özen, A. & Matthews, C. R. (2015). Clusters of isoleucine, leucine, and valine side chains define cores of stability in high-energy states of globular proteins: Sequence determinants of structure and stability. *Protein Science, 25(3).* doi:10.1002/pro.2860

Kim, K. M. & Caetano-Anollés, G. (2010). Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data. *Molecular Biology and Evolution, 27*(7), 1710-1733.

Kim, K. M. & Caetano-Anollés, G. (2011). The proteomic complexity and rise of the primordial ancestor of diversified life. *BMC Evolutionary Biology, 11*, 140-164.

Kim, K. M. & Caetano-Anollés, G. (2012). The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms. *BMC Evolutionary Biology, 12*, 13.

Kim, H. S., Mittenthal, J. & Caetano-Anollés, G. (2006). MANET: using evolution of protein architecture in metabolic networks. *BMC Bioinformatics, 7*, 351. doi:10.1186/1471-2105-7-351

Koonin, E. V. (2012). *The logic of chance: the nature and origin of biological evolution*. Pearson Education, Inc, F.T. Press Science, Upper Saddle River, New Jersey.

Krieger E., & Vriend, G. (2014). YASARA view – molecular graphics for all devices – from smartphones to workstations. *Bioinformatics, 30*(20), 2981-2982.

Lane, N. (2010). *Life Ascending: The Ten Great Inventions of Evolution*. W. W. Norton & Co., New York.

Laskowski, R. A. (2016). Protein structure data bases. *Methods in Molecular Biology, 1415*, 31-53. doi:10.1007/978-1-4939-3572-7-2

Lazcano, A. & Miller, S. L. (1996). The origin and early evolution of life: prebiotic chemistry, the pre-RNA world and time. *Cell, 85*, 793-798.

Lazcano, A. & Miller, S. L. (1999). On the origin of metabolic pathways. *Journal of Molecular Evolution, 49*, 424-431.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, (ed.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, CA, pp. 278-292.

Liang, J. & Dill, K. A. (2001). Are proteins well-packed? *Biophysical Journal, 81*, 751-766.

Lipmann, F. (1965). Projecting backward from the present stage of evolution of biosynthesis. In *The Origins of Prebiological Systems and of their Molecular Matrices.* K. Harada and S. W. Fox, (eds.), Academic Press, New York, pp. 259-280.

Ma, B-G., Chen, L., Ji, H-F., Chen, Z-H., Yang, F-R., Wang, L., Qu, G., Jiang, Y-Y., Ji, C. & Zhang, H-Y. (2008). Characters of very ancient proteins. *Biochemical and Biophysical Research Communications, 366*, 607-611.

Makhatadze, G. I, Privalov, P. L. (1995). Energetics of protein structure. *Advances in Protein Chemistry, 47*, 307-425.

Manavalan, P. & Ponnuswamy P. K. (1977). A study of the preferred environment of amino acid residues in globular proteins. *Archives of Biochemistry and Biophysics, 184*, 476-487.

Martz, E. (2012). 'Introduction to Evolutionary Conservation'. *Proteopedia*. doi:10.14576/514849.1541287

McGuiness, E. T. (2010). Some molecular moments of the Hadean and Archaean aeons: a retrospective overview from the interfacing years of the second to third millennia. *Chemical Reviews, 110*, 5191-5215.

Mittal, A. & Jayaram, B. (2011). Backbones of folded proteins reveal novel invariant amino acid neighborhoods. *Journal of Biomolecular Structure and Dynamics, 28*, 443-454.

Mittal, A., Jayaram, B., Shenoy, S. & Bawa, T.S. (2010). A stoichiometry driven universal spatial organization of backbones of folded proteins: are their Chargaff's rules for protein folding? *Journal of Biomolecular Structure and Dynamics, 28*, 133-142.

Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K. & Uversky, V. N. (2006). Analysis of molecular recognition features (MoFRs). *Journal of Molecular Biology, 362*, 1043-1059.

Morowitz, H. J. (1992). *Beginnings of Cellular Life*. *Metabolism Recapitulates Biogenesis*. Yale University Press, New Haven & London.

Nath, N., Mitchel, J. B. O., Caetano-Anollés, G. (2014). The natural history of biochemical mechanisms. *PLOS Computational Biology, 10*(5), e1003642.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A (General), 135*(3), 370-384.

Nitschke, W. & Russell, M. J. (2013). Beating the acetyl coenzyme A-pathway to the origin of life. *Philosophical Transactions of the Royal Society B, 368*, 20120258.

Pe'er, I., Felder, C. E., Man, O., Silman, I., Sussman, J. L. & Beckmann J. S. (2004). Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla. *Proteins: Structure, Function, and Bioinformatics, 54*, 20-40.

Peretó, J., Fani, R., Leguina, J. I. & Lazcano, A. (1997). Enzyme evolution and the development of metabolic pathways. In Cornish-Bowden, A. (ed.), *New Beer In An Old Bottle: Edward Buchner and the Growth of Biochemical Knowledge.* University of València, València, pp. 173-198. ISBN: 84-370-3328-4 (GUADA Litografía, S. L., Camí Nou de Picanya 3, 46014, València; also, via Google Scholar and Amazon)

Pollack, J. D., Pan, X. & Pearl, D. K. (2010). Concentration of specific amino acids at the catalytic/active centers of highly-conserved 'housekeeping' enzymes of central metabolism in Archaea, Bacteria and Eukarya: is there a widely conserved chemical signal of prebiotic assembly? *Origins of Life and Evolution of Biospheres, 40*, 273-302.

Pollack, J. D., Gerard, D. & Pearl, D. K. (2013). Uniquely localized intra-molecular amino acid concentrations at the glycolytic enzyme catalytic/active centers of Archaea, Bacteria and Eukarya are associated with their proposed temporal appearances on Earth. *Origins of Life and Evolution of Biospheres, 43*, 161-187.

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., Hess, B. & Lindahl, E. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics (Oxford, England), 29*, 845-854.

R Development Core Team. (2011). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.

Raggi, L., Bada, J. L. & Lazcano, A. (2016). On the lack of evolutionary continuity between prebiotic peptides and extant enzymes. *Physical Chemistry Chemical Physics.* doi:10.1039/c6cp00793g

Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *Journal of Molecular Biology, 82,* 1-14.

Risso, V. A., Manssour-Triedo, F., Delgado-Delgado, A., Arco, R., Barroso-delJesus, A., Ingles-Prieto, A., Godoy-Ruiz, R., Gavira, J. A., Gaucher, E. A., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. (2015). Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. *Molecular Biology and Evolution, 32*(2), 440-455.

Risso, V. A., Martinez-Rodriguez, S., Candel, A. M., Krüger, D. M., Pantoja-Uceda, D., Ortega-Muñoz, M., Santoyo-Gonzalez, F., Gaucher, E. A., Kamerlin, S. C. L., Bruix, M., Gavira, J. A. & Sanchez-

Ruiz, J. M. (2017). *De novo* active sites for resurrected Precambrian enzymes. *Nature Communications* 8:16113. doi:10.1038/ncomms16113

Romero, M. L. R., Rabin, A. & Tawfik, D. S. (2016). Functional proteins from short peptides: Dayhoff's hypothesis turns 50. *Angewandte Chemie International Edition, 35*, 15966-15971.

Russell, M. (2016). Emergence of life and its early history. Chapter. 3, pp. 19-54. In H. L. Ehrlich, D. K. Newman & A. Kappler (eds.), *Ehrlich's Geomicrobiology*, 6th Edition, CRC Press, Boca Raton.

Schidlowski, M. (1988). A 3,800-million-year isotopic record of life from carbon in sedimentary rocks. *Nature, 333*(6171), 313-318.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G. and Schomburg, D. (2004). BRENDA, the enzyme data base: updates and major new developments. *Nucleic Acids Research, 32*, D431-D433. doi:10.1093/nar/gkh081

Slesak, I., Slesak, H. & Kruk, J. (2017). RubisCO early oxygenase activity: a kinetic and evolutionary perspective. *BioEssays 39*, 1700071. doi:10.1002/bies.201700071

Sleep, N. H. (2010). The Hadean-Archaean environment. *Cold Spring Harbor Perspectives in Biology 2*:a002527.

Smith, E. & Morowitz, H. (2016). *The Origin and Nature of Life on Earth: The Emergence of the Fourth Geosphere*. Cambridge University Press, New York.

Strasdeit, H. (2010). Chemical evolution and early Earth's and Mars' environmental conditions. *Paleodiversity, 3*, Supplement: 107-116.

Trifonov, E. N. (2000). Consensus temporal order of amino acids and evolution of the triplet code. *Gene, 261*, 139-151.

Trifonov, E. N. (2004). The triplet code from first principles. *Journal of Biomolecular Structure and Dynamics 22*, 1-11.

Trifonov E. N. (2009). The origin of the genetic code and the earliest oligopeptides. *Research in Microbiology, 160*, 481-486.

Wang, M., Jiang, Y-Y., Kim, K. M., Qu, G., Ji, H-F., Mittenthal, J. E., Zhang, H-U. & Caetano-Anollés, G. (2011). A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. *Molecular Biology and Evolution, 28*, 567-582.

Waterhouse, A. M., Proctor, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. (2009). Jalview Version, 2 – a multiple sequence alignment editor and analysis workbench. *Bioinformatics, 25*(9), 1189-1191. doi:10.1093/bioinformatics/btp033

Wessa, P. (2017). Kendall tau Rank Correlation (v1.0.13). In *Free Statistics Software* (v1.2.1), Office for Research Development and Education, https://www.wessa.net/rwasp_kendall.wasp/

Weiss, M. C., Sousa, F. L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S. & Martin, W. F. (2016). The physiology and habitat of the last universal common ancestor. *Nature Microbiology*. doi:10.1038/nmicrobiol.2016.116

Yan, B. X. & Sun, Y. Q. (1997) Glycine residues provide flexibility for enzyme active sites. *Journal of Biological Chemistry, 272*, 3190–3194.

Zaia, D. A. M., Zaia, C. T. B. V. & de Santana, H. (2008). Which amino acids should be used in prebiotic chemistry studies? *Origins of Life and Evolution of Biospheres, 38*, 469-488.

Zamyatnin, A. A. (1984). Amino acid, peptide, and protein volume in solution. *Annual Review of Biophysics and Bioengineering, 13,* 145-165. doi:10.1146/annurev.bb.13.060184.001045

Zubarev D. Y., Rappoport, D. & Aspuru-Guzik, A. (2015). Uncertainty of prebiotic scenarios: the case of the non-enzymatic reverse tricarboxylic acid cycle, *Scientific Reports, 5*, 8009. doi:10.1038/srep08009

Table 1. oTCA's enzymes: sequences and their Domain, 'Scaffold', 'Anchor-Amino acid' and its 'Anchor-Atom'

Enzyme(s)	EC	Sequences in Each Domain Set	Each 'Domain Set's 'Scaffold(s)'	PDB	ExPASy	Anchor-Amino Acid' and its 'Anchor-Atom'	Notes
Citrate synthase	2.3.3.16	Archaea (95)	*Pyrococcus furiosus*	1aj8	Q53554	D312OG2	
		Bacteria (352)	*Escherichia coli*	4g6b	P0ABH7	D362OD2	
			Thermus thermophilus	1ixe	Q5SIM6	D312OD2	
		Eukaryota (95)	*Sus scrofa*	3enj	P00889	D375OD2	a
			Gallus gallus	5csc	P23007	D375OD2	
Aconitate hydratase	4.2.1.3	Bacteria (189)	*Escherichia coli*	1l5j	P36683	S244OG	b
		Eukaryota (24)	*Sus scrofa*	1b0j	P16276	D100OD2	c
			Bos taurus	8acn	P20004	D100OD2	
Isocitrate dehydrogenase	1.1.1.41	Eukaryota (41)	*Saccharomyces cereviseae*	3blv	P28834	D217OD2	d
2-Oxoglutarate dehydrogenase	1.2.4.2	Bacteria (260)	*Escherichia coli*	2jgd	P0AFG3	H298NE2	e
Succinyl-CoA ligase	6.2.1.4-5	Archaea (149)	*Methanocaldococcus jannaschii*	2yv1	Q58643	D214OE2	f
		Bacteria (429)	*Escherichia coli*	2scu	P0AGE9	NEP246ND1	f,g
		Eukaryota (69)	*Sus scrofa*	1euc	O19069	H29NE2	h
Succinate dehydrogenase	1.3.5.1	Bacteria (123)	*Escherichia coli*	1nek	P0AC41	H354NE2	i
Fumarate hydratase ("fumarase") malate forming	4.2.1.2	Bacteria (379)	*Escherichia coli*[j]	1fuq	P05042	H188NE2	j
			Rickettsia prowazekii	3gtd	Q9ZCQ4	K324NZ	
		Eukaryota (59)	*Homo sapiens*	3e04	P07954	E378OE2	k
			Saccharomyces cereviseae	1yfm	P08417	H213NE2	
Malate dehydrogenase	1.1.1.37	Archaea (92)	*Haloarcula marismortui*	1hlp	Q07841	D166OD2	l
			Archaeoglobus fulgidus	2x0i	O08349	D168OD2	
		Bacteria (390)	*Brucella abortus*	3gvh	Q2YLR9	H176NE2	
			Escherichia coli	1emd	P61889	H177NE2	
		Eukaryota (78)	*Sus scrofa*	1mld	P00346	H176NE2	
			Citrulus lanatus	1sev	P19446	H220NE2	m
			Homo sapiens	2dfd	P40926	H182NE2	
Table seqs totals (2,824)		3A/7B/6E: 336/2,122/366					

Notes: [a]Sequences (= seqs) are mitochondrial; [b]'Scaffold' assignment; [c]aconitate hydratase seqs are predominantly PDBs aco2 and acnB; [d]seqs are predominantly mitochondrial and [NAD] α-subunit (Site 1); [e]seqs are predominantly E1 component (Frank et al., 2007, Ævarsson et al., 1999); [f]seqs are all [ADP-forming] subunits α and β; [g]'NEP', i.e., the Anchor-atom for PDB-2scu is the ND1 atom of the Anchor-amino acid N1-phosphonohistidine designated PDB NEP(A)246ND1 or PDB-2scu HETATM 1787; [h]six 'fragment' seqs are included in the 69, they are variously ADP or GDP (or both)-forming and are α-subunit and mitochondrial (Fraser, James, Bridger & Wolodko, 2000); [i]seqs are Gammaproteobacteria flavoprotein sub-units; [j]seqs are predominantly Gammaproteobacteria Class II; [k]seqs are predominantly Alphaproteobacteria Class II ; [l]seqs are Archaea Phylum Euryarchaeota; [m]four 'fragment' seqs are included in the predominantly mitochondrial set

Table 2. Bacteria enzyme sequences: geochronologically related 'nd-Age' distances, taxonomy,'Scaffolds', and their 'Anchor-Amino acid' and its 'Anchor Atom'*

		ENZYME (recommended by Nomenclature Committee of the International Union of Biochemistry and Molecular Biology and Brenda data base)	nd-age*	EC	Taxonomic Identities of Bacteria Sequences in each FASTA-MSA Homologous Enzyme Set	Selected Scaffold(s) in FASTA-MSA Enzyme Set		
						Bacteria	PDB**	ExPASy	Anchor-Amino Acid' and its 'Anchor-Atom'
A. 'Epoch I' 4,000-2,700 Million (Ma) Years*	1	Acetate-CoA ligase (acetyl-CoA synthetase)	0.029	6.2.1.1	Bacteria (30)	*Salmonella enterica*	2p2f	Q8ZKF6	K609CD
	2	H^+-transporting two-sector ATPase	0.044	3.6.3.14	Firmicutes (231)	*Bacillus* PS3	1sky	P07677	R191(E)CA
	3	Adenine phosphoribosyltransferase	0.051†	2.4.2.7	Gammaproteobacteria (393)	*Escherichia coli*	2dy0	P69503	R66NH2
	4	Uracil phosphoribosyltransferase	0.051	2.4.2.9	Proteobacteria (533)	*Escherichia coli*	2ehj	P0A8F0	D139OD2
	5	Transaldolase (pentose/EMP)	0.055	2.2.1.2	Gammaproteobacteria (60)	*Escherichia coli*	1onr	P0A870	K132NZ
	6	Aspartate transaminase	0.069	2.6.1.1	Proteobacteria (152)	*Escherichia coli*	1aam	P00509	K258NZ
	7	Glutamate synthase, NADPH	0.073	1.4.1.13	Proteobacteria (25)	*Azospirillum brasilense*	1ea0	Q05755	C1SG
	8	Adenylate kinase	0.077	2.7.4.3	Bacteria (147)§	*Bacillus subtilis*	1p3j	P16304	R160NH2
						Escherichia coli	4ake	P69441	R156NH2
	9	3-Isopropylmalate dehydrogenase	0.095†	1.1.1.85	Gammaproteobacteria (302)	*Escherichia coli*	1cm7	P30125	K195NZ
	10	IMP (inosine 5'PO_4) dehydrogenase	0.128	1.1.1.205	Firmicutes (125)	*Streptococcus pyogenes*	1zfj	P0C0H6	C310SG
	11	Fumarate reductase, quinol [rTCA]	0.131†	1.3.5.4	Bacteria (143)	*Escherichia coli*	2b76	P00363	R390NH2
	12	Dihydroorotase	0.153	3.5.2.3	Firmicutes (112)	*Bacillus anthracis*	3mpg	Q81WF0	D304OD2
	13	Triose-phosphate isomerase [EMP]	0.157	5.3.1.1	Firmicutes (269)	*Geobacillus stearothermophilus*	1btm	P00943	E166OE1
	14	Nucleoside-diphosphate kinase	0.164†	2.7.4.6	Bacteria (268)§	*Mycobacterium tuberculosis*	1k44	P9WJH7	H117NE2
						Virgibacillus halodenitrificans	1nb2	Q7SIA9	H116NE2
	15	Histidinol dehydrogenase	0.186†	1.1.1.23	Bacteria (132)	*Escherichia coli*	1kae	P06988	H327NE2
	16	Ribose-phosphate diphosphokinase	0.193†	2.7.6.1	Firmicutes (217)	*Bacillus subtilis*	1dkr	P14193	H135NE2
	17	Fructose-*bis* phosphate aldolase	0.193	4.1.2.13	Gammaproteobacteria (140)	*Escherichia coli*	1b57	P0AB71	D109OD2
	18	Inorganic diphosphatase	0.212†	3.6.1.1	Gammaproteobacteria (294)§	*Escherichia coli*	1igp	P0A7A9	D67OD2
						Oleispira antarctica	3i4q	D0VWZ3	D68OD2
	19	Formate dehydrogenase	0.225	1.2.1.2	Proteobacteria (37)	*Escherichia coli* ‡	1kqf	P24183	H197NE2
	20	Thioredoxin-disulfide reductase	0.241	1.8.1.9	Bacteria (464)§	*Escherichia coli*	1c0l	P0A9P4	C138SG
						Helicobacter pylori	3ish	P56431	C136SG
		Total sequences (5,545)			4,074 = 73.3%				
B. 'Epoch III' 2,100 Million (Ma) Years - Present*	21	Ribulose-*bis* PO_4 carboxylase (RuBisCo)†	0.675†	4.1.1.39	Bacteria (143)§	*Alkaligenes eutrophus*	1bxn	P0C2C2	K178NZ
						Rhodospirillum rubrum	1rba	P04718	K166NZ
						Thermosynechococcus elongatus ◊	2ybv	Q8DIS5	K175NZ
	22	Tryptophan 2,3-dioxygenase	0.690	1.13.11.11	Gammaproteobacteria (63)	*Xanthomonas campestris*	2aw8	Q8PDA8	H55NE2
	23	Aspartate racemase	0.730	5.1.1.13	Proteobacteria (269)§	*Yersinia pestis*	3ojc	Q74SZ5	C198SG
						Salmonella enterica, enterica	3s81	Q8ZJZ9	C193SG
	24	Ribosylpyrimidine nucleosidase	0.758	3.2.2.8	Gammaproteobacteria (23)	*Escherichia coli*	1q8f	P33022	D75OD2
	25	Acylphosphatase	0.810	3.6.1.7	Firmicutes (58)	*Bacillus subtilis*	2fhm	O35031	N36OD1
	26	Ferredoxin-$NADP^+$ reductase	0.810	1.18.1.2	Bacteria (232)§	*Azotobacter vinelandii*	1a8p	Q44532	Y53O
						Escherichia coli	1fdr	P28861	Y52O
	27	Pyridoxal kinase	0.839	2.7.1.35	Proteobacteria (168)	*Escherichia coli*	1td2	P77150	Q46NE2
	28	Cytidine deaminase	0.854	3.5.4.5	Firmicutes (134)	*Bacillus subtilis*	1jtk	P19079	E55OE2
	29	Alkaline phosphatase	0.858	3.1.3.1	Proteobacteria (56)	*Escherichia coli*	4km4	P00634	H370NE2
	30	Carbon-monoxide dehydrogenase, acceptor	0.865	1.2.99.2	Proteobacteria (30)	*Oligotropha carboxidovorans*	1n5w	P19921	E763(B)OE2
	31	Carbonic anhydratase	0.887†	4.2.1.1	Gammaproteobacteria (159)	*Haemophilus influenzae*	2a8c	P45148	R46NH2
	32	Arginase	0.898	3.5.3.1	Firmicutes (98)	*Bacillus caldovelox*	1cev	P53608	D126OD2
						Mycobacterium tuberculosis	1pzs	P9WEG9	H75NE2
	33	Superoxide dismutase (SOD)	0.905†	1.15.1.1	Bacteria (38)§	*Escherichia coli*	1es0	P0AGD1	H61NE2
						Haemophilus ducreyi	1e9p	Q59452	H95NE2
		Total sequences (5,545)			1,471 = 26.5%				

Notes: All enzyme nd-age values were characterized as 'Epoch I' or 'Epoch III' after 'Epoch I Architectural Diversification' or 'Epoch III Organismal Diversification' described by Caetano-Anollés et al., 2009, Wang et al., 2011 and Kim & Caetano-Anollés, 2012. Compatible numerical geologic ages were characterized and assigned by us using the GTS schedule (Cohen et al., 2015); **PDB structural nomenclature are Chain A unless noted; †from the unpublished studies of Professor Gustavo Caetano-Anollés; §these identify the set of 2-3 Scaffolds whose separate analyses of the replicate Bacteria sequences are averaged; ‡α-subunit (residues 34-850); ◊large chain (see, Methods)

Table 3. Comparisons of RAA % Occupancies between oTCA and glycolysis enzymes across three Domains and their distances to their catalytic/active centers (C/AC*

CZ9 to CZ1				CZ9				CZ1			
oTCA		Glycolysis**		oTCA		Glycolysis**		oTCA		Glycolysis**	
RAA	%OCC	RAA	%OCC	RAA	%OCC	RAA	%OCC	RAA	%OCC	RAA	%OCC
A	10.27	A	11.38	G	16.09	G	16.04	E	18.14	E	15.52
L	9.23	L	9.48	A	10.09	A	8.35	K	14.59	K	15.46
G	9.14	G	9.21	P	6.76	D	7.69	D	10.09	A	10.87
E	7.61	E	8.31	D	6.56	E	7.33	A	9.09	D	8.30
V	7.24	V	8.07	R	6.30	S	6.61	L	6.22	L	7.67
I	6.06	K	7.52	T	6.30	N	6.55	R	5.18	G	7.17
D	6.01	D	6.11	N	5.83	R	6.07	P	5.02	P	6.29
K	5.97	I	5.83	S	5.63	T	6.07	G	3.63	V	4.78
P	5.37	T	4.93	L	5.60	L	4.68	N	3.59	R	3.96
T	5.13	P	4.49	E	5.19	V	4.68	T	3.51	S	3.14
R	4.77	S	4.15	V	5.13	P	4.62	S	3.31	T	3.14
S	4.49	R	4.11	I	4.37	K	3.90	V	3.31	I	2.77
N	3.61	N	3.97	H	3.79	I	3.72	Q	2.87	N	2.77
F	3.31	F	3.11	Q	2.74	H	3.12	Y	2.71	F	2.07
Y	2.72	Y	2.59	K	2.39	Q	3.06	I	2.55	Y	1.89
Q	2.42	H	1.83	M	2.36	M	2.34	H	2.23	Q	1.38
H	2.40	Q	1.81	F	2.22	F	2.10	F	2.03	H	1.07
M	2.23	M	1.80	Y	1.52	Y	2.04	M	0.92	M	0.88
C	1.13	W	0.76	W	0.70	W	0.66	W	0.64	C	0.50
W	0.87	C	0.56	C	0.41	C	0.36	C	0.32	W	0.38
n_{RAA} = 7,284		8,738		1,704		1,665		1,243		1,591	
K-τ = 0.926				K-τ = 0.779				K-τ = 0.800			
Average distance (Å) from Recovered-amino Acid's (nRAAs) Cαs to assigned Anchor-Atom ± SEM in CZs above across three Domains (oTCA malate dehydrogenase Eukaryota distance data were omitted)**											
—		—		19.42 ± 1.33		15.21 ± 0.012		30.82 ± 1.13		27.26 ± 0.056	

**All glycolysis data taken from: Pollack, J. D., Gerard, D. & Pearl, D. K. (2013). *Orig Life Evol Biosph 43*:161-187 (Table 1)

Notes: Comparisons between pathways by Kendall's ranked correlation vtest (K-τ): the ranked contents of their pooled nine conservation zones: all CZ9 through CZ1, or only CZ9s or CZ1s. Also shown are the data's 'Average distance (Å) from Recovered-Amino Acids (RAAs)…..' to a reported functional Anchor-Atom ± SEM in their respective C/AC locus within their Anchor-Amino acid (AAA) and its parent Scaffold (see, Methods)

Table 4. Individual analyses and trends of the % Occupancies of all Recovered-Amino Acid (RAAs) sites in each of the eight oTCA cycle enzymes of Archaea, Bacteria and Eukaryota

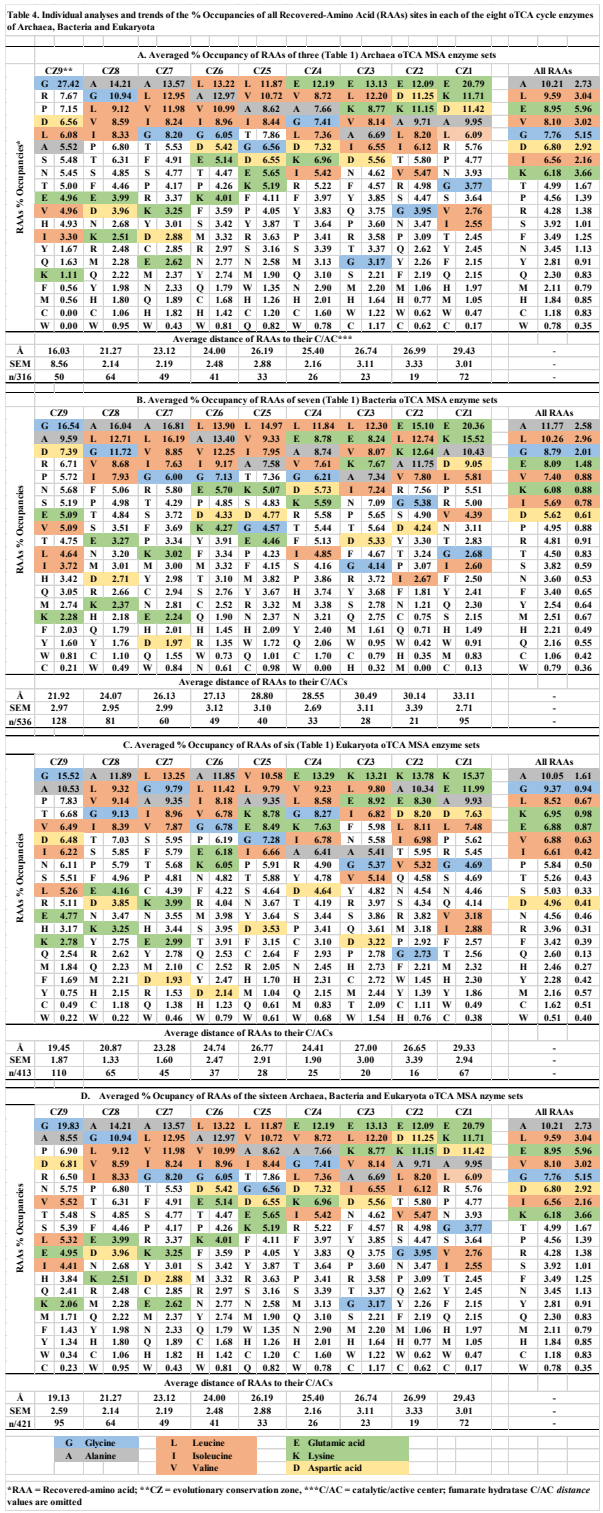

Table 4. Individual analyses and trends of the % Occupancies of all Recovered-Amino Acid (RAAs) sites in each of the eight oTCA cycle enzymes of Archaea, Bacteria and Eukaryota

A. Averaged % Occupancy of RAAs of three (Table 1) Archaea oTCA MSA enzyme sets

RAAs % Occupancies*

CZ9**		CZ8		CZ7		CZ6		CZ5		CZ4		CZ3		CZ2		CZ1		All RAAs		
G	27.42	A	14.21	A	13.57	L	13.22	L	11.87	E	12.19	E	13.13	E	12.09	E	20.79	A	10.21	2.73
R	7.67	G	10.94	L	12.95	A	12.97	V	10.72	V	8.72	L	12.20	D	11.25	K	11.71	L	9.59	3.04
P	7.15	L	9.12	V	11.98	V	10.99	A	8.62	A	7.66	K	8.77	K	11.15	D	11.42	E	8.95	5.96
D	6.56	V	8.59	I	8.24	I	8.96	I	8.44	G	7.41	V	8.14	A	9.71	A	9.95	V	8.10	3.02
L	6.08	I	8.33	G	8.20	G	6.05	T	7.86	L	7.36	A	6.69	L	8.20	L	6.09	G	7.76	5.15
A	5.52	P	6.80	T	5.53	D	5.42	G	6.56	D	7.32	I	6.55	I	6.12	R	5.76	D	6.80	2.92
S	5.48	T	6.31	F	4.91	E	5.14	D	6.55	K	6.96	D	5.56	T	5.80	P	4.77	I	6.56	2.16
N	5.45	S	4.85	S	4.77	T	4.47	E	5.65	I	5.42	N	4.62	V	5.47	N	3.93	K	6.18	3.66
T	5.00	F	4.46	P	4.17	P	4.26	K	5.19	R	5.22	F	4.57	R	4.98	G	3.77	T	4.99	1.67
E	4.96	E	3.99	R	3.37	K	4.01	F	4.11	F	3.97	Y	3.85	S	4.47	S	3.64	P	4.56	1.39
V	4.96	D	3.96	K	3.25	F	3.59	P	4.05	Y	3.83	Q	3.75	G	3.95	V	2.76	R	4.28	1.38
H	4.93	N	2.68	Y	3.01	S	3.42	Y	3.87	T	3.64	P	3.60	N	3.47	I	2.55	S	3.92	1.01
I	3.30	K	2.51	D	2.88	M	3.32	R	3.63	P	3.41	R	3.58	P	3.09	T	2.45	F	3.49	1.25
Y	1.67	R	2.48	C	2.85	R	2.97	S	3.16	S	3.39	T	3.37	Q	2.62	Y	2.45	N	3.45	1.13
Q	1.63	M	2.28	E	2.62	N	2.77	N	2.58	M	3.13	G	3.17	Y	2.26	F	2.15	Y	2.81	0.91
K	1.11	Q	2.22	M	2.37	Y	2.74	M	1.90	Q	3.10	S	2.21	F	2.19	Q	2.15	Q	2.30	0.83
F	0.56	Y	1.98	N	2.33	Q	1.79	W	1.35	N	2.90	M	2.20	M	1.06	H	1.97	M	2.11	0.79
M	0.56	H	1.80	Q	1.89	C	1.68	H	1.26	H	2.01	H	1.64	H	0.77	M	1.05	H	1.84	0.85
C	0.00	C	1.06	H	1.82	H	1.42	C	1.20	C	1.60	W	1.22	W	0.62	W	0.47	C	1.18	0.83
W	0.00	W	0.95	W	0.43	W	0.81	Q	0.82	W	0.78	C	1.17	C	0.62	C	0.17	W	0.78	0.35

Average distance of RAAs to their C/AC***

Å	16.03	21.27	23.12	24.00	26.19	25.40	26.74	26.99	29.43	-
SEM	8.56	2.14	2.19	2.48	2.88	2.16	3.11	3.33	3.01	-
n/316	50	64	49	41	33	26	23	19	72	-

B. Averaged % Occupancy of RAAs of seven (Table 1) Bacteria oTCA MSA enzyme sets

RAAs % Occupancies

CZ9		CZ8		CZ7		CZ6		CZ5		CZ4		CZ3		CZ2		CZ1		All RAAs		
G	16.54	A	16.04	A	16.81	L	13.90	L	14.97	L	11.84	L	12.30	E	15.10	E	20.36	A	11.77	2.58
A	9.59	L	12.71	L	16.19	A	13.40	V	9.33	E	8.78	E	8.24	L	12.74	K	15.52	L	10.26	2.96
D	7.39	G	11.72	V	8.85	V	12.25	I	7.95	A	8.74	V	8.07	K	12.64	A	10.43	G	8.79	2.01
R	6.71	V	8.68	I	7.63	I	9.17	A	7.58	V	7.61	K	7.67	A	11.75	D	9.05	E	8.09	1.48
P	5.72	I	7.93	G	6.00	G	7.13	T	7.36	G	6.21	A	7.34	V	7.80	L	5.81	V	7.40	0.88
N	5.68	F	5.06	R	5.80	E	5.70	K	5.07	D	5.73	I	7.24	R	7.56	P	5.51	K	6.08	0.88
S	5.19	P	4.98	T	4.29	P	4.85	S	4.83	K	5.59	N	7.09	G	5.38	R	5.00	I	5.69	0.78
E	5.09	T	4.84	S	3.72	D	4.33	D	4.77	R	5.58	P	5.65	S	4.90	V	4.39	D	5.62	0.61
V	5.09	S	3.51	F	3.69	K	4.27	G	4.57	T	5.44	T	5.64	D	4.24	N	3.11	P	4.95	0.88
T	4.75	E	3.27	P	3.34	Y	3.91	E	4.46	F	5.13	D	5.33	Y	3.30	T	2.83	R	4.81	0.91
L	4.64	N	3.20	K	3.02	F	3.34	P	4.23	I	4.85	F	4.67	T	3.24	G	2.68	T	4.50	0.83
I	3.72	M	3.01	M	3.00	M	3.32	F	4.15	S	4.16	G	4.14	P	3.07	I	2.60	S	3.82	0.59
H	3.42	D	2.71	Y	2.98	T	3.10	M	3.82	P	3.86	R	3.72	I	2.67	F	2.50	N	3.60	0.53
Q	3.05	R	2.66	C	2.94	S	2.76	Y	3.67	H	3.74	Y	3.68	F	1.81	Y	2.41	F	3.40	0.65
M	2.74	K	2.37	N	2.81	C	2.52	R	3.32	M	3.38	S	2.78	N	1.21	Q	2.30	Y	2.54	0.64
K	2.28	H	2.18	E	2.24	Q	1.90	N	2.37	N	3.21	Q	2.75	C	0.75	S	2.15	M	2.51	0.67
F	2.03	Q	1.79	H	2.01	H	1.45	H	2.09	Y	2.40	M	1.61	Q	0.71	H	1.49	H	2.21	0.49
Y	1.60	Y	1.76	D	1.97	R	1.35	W	1.72	Q	2.06	W	0.95	W	0.42	W	0.91	Q	2.16	0.55
W	0.81	C	1.10	Q	1.55	W	0.73	Q	1.01	C	1.70	C	0.79	H	0.35	M	0.83	C	1.06	0.42
C	0.21	W	0.49	W	0.84	N	0.61	C	0.98	W	0.00	H	0.32	M	0.00	C	0.13	W	0.79	0.36

Average distance of RAAs to their C/ACs

Å	21.92	24.07	26.13	27.13	28.80	28.55	30.49	30.14	33.11	-
SEM	2.97	2.95	2.99	3.12	3.10	2.69	3.11	3.39	2.71	-
n/536	128	81	60	49	40	33	28	21	95	-

C. Averaged % Occupancy of RAAs of six (Table 1) Eukaryota oTCA MSA enzyme sets

RAAs % Occupancies

CZ9		CZ8		CZ7		CZ6		CZ5		CZ4		CZ3		CZ2		CZ1		All RAAs		
G	15.52	A	11.89	L	13.25	A	11.85	V	10.58	E	13.29	K	13.21	K	13.78	K	15.37	A	10.05	1.61
A	10.53	L	9.32	G	9.79	L	11.42	L	9.79	V	9.23	L	9.80	A	10.34	E	11.99	G	9.37	0.94
P	7.83	V	9.14	A	9.35	I	8.18	A	9.35	L	8.58	E	8.92	E	8.30	A	9.93	L	8.52	0.67
T	6.68	G	9.13	I	8.96	V	6.78	K	8.78	G	8.27	I	6.82	D	8.20	D	7.63	K	6.95	0.98
V	6.49	I	8.39	V	7.87	G	6.78	E	8.49	K	7.63	F	5.98	L	8.11	L	7.48	E	6.88	0.87
D	6.48	T	7.03	S	5.95	P	6.19	G	7.28	I	6.78	N	5.58	I	6.98	P	5.62	V	6.88	0.63
I	6.22	S	5.85	F	5.79	E	6.18	I	6.66	A	6.41	A	5.41	T	5.95	R	5.45	I	6.61	0.42
N	6.11	P	5.79	T	5.68	K	6.05	P	5.91	R	4.90	G	5.37	V	5.32	G	4.69	P	5.84	0.50
S	5.51	F	4.96	P	4.81	N	4.82	T	5.88	Y	4.78	V	5.14	Q	4.58	S	4.69	T	5.26	0.43
L	5.26	E	4.16	C	4.39	F	4.22	S	4.64	D	4.64	Y	4.82	N	4.54	N	4.46	S	5.03	0.33
R	5.11	D	3.85	K	3.99	R	4.04	N	3.67	T	4.19	R	3.97	S	4.34	Q	4.14	D	4.96	0.41
E	4.77	N	3.47	N	3.55	M	3.98	Y	3.64	S	3.44	S	3.86	R	3.82	V	3.18	N	4.56	0.46
H	3.17	K	3.25	H	3.44	S	3.95	D	3.53	P	3.41	Q	3.61	M	3.18	I	2.88	R	3.96	0.31
K	2.78	Y	2.75	E	2.99	T	3.91	F	3.15	C	3.10	D	3.22	P	2.92	F	2.57	F	3.42	0.39
Q	2.54	R	2.62	Y	2.78	Q	2.53	C	2.64	F	2.93	P	2.78	G	2.73	T	2.56	Q	2.60	0.13
M	1.84	Q	2.23	M	2.10	C	2.52	R	2.05	N	2.45	H	2.73	F	2.21	M	2.32	H	2.46	0.27
F	1.69	M	2.21	D	1.93	Y	2.47	H	1.70	H	2.31	C	2.72	W	1.45	H	2.30	Y	2.28	0.42
Y	0.75	H	2.15	R	1.53	D	2.14	M	1.04	Q	2.15	M	2.44	Y	1.39	Y	1.86	M	2.16	0.57
C	0.49	C	1.18	Q	1.38	H	1.23	Q	0.61	M	0.83	T	2.09	C	1.11	W	0.49	C	1.62	0.51
W	0.22	W	0.22	W	0.46	W	0.79	W	0.61	W	0.68	W	1.54	H	0.76	C	0.38	W	0.51	0.40

Average distance of RAAs to their C/ACs

Å	19.45	20.87	23.28	24.74	26.77	24.41	27.00	26.65	29.33	-
SEM	1.87	1.33	1.60	2.47	2.91	1.90	3.00	3.39	2.94	-
n/413	110	65	45	37	28	25	20	16	67	-

D. Averaged % Occupancy of RAAs of the sixteen Archaea, Bacteria and Eukaryota oTCA MSA nzyme sets

RAAs % Occupancies

CZ9		CZ8		CZ7		CZ6		CZ5		CZ4		CZ3		CZ2		CZ1		All RAAs		
G	19.83	A	14.21	A	13.57	L	13.22	L	11.87	E	12.19	E	13.13	E	12.09	E	20.79	A	10.21	2.73
A	8.55	G	10.94	L	12.95	A	12.97	V	10.72	V	8.72	L	12.20	D	11.25	K	11.71	L	9.59	3.04
P	6.90	L	9.12	V	11.98	V	10.99	A	8.62	A	7.66	K	8.77	K	11.15	D	11.42	E	8.95	5.96
D	6.81	V	8.59	I	8.24	I	8.96	I	8.44	G	7.41	V	8.14	A	9.71	A	9.95	V	8.10	3.02
R	6.50	I	8.33	G	8.20	G	6.05	T	7.86	L	7.36	A	6.69	L	8.20	L	6.09	G	7.76	5.15
N	5.75	P	6.80	T	5.53	D	5.42	G	6.56	D	7.32	I	6.55	I	6.12	R	5.76	D	6.80	2.92
V	5.52	T	6.31	F	4.91	E	5.14	D	6.55	K	6.96	D	5.56	T	5.80	P	4.77	I	6.56	2.16
T	5.48	S	4.85	S	4.77	T	4.47	E	5.65	I	5.42	N	4.62	V	5.47	N	3.93	K	6.18	3.66
S	5.39	F	4.46	P	4.17	P	4.26	K	5.19	R	5.22	F	4.57	R	4.98	G	3.77	T	4.99	1.67
L	5.32	E	3.99	R	3.37	K	4.01	F	4.11	F	3.97	Y	3.85	S	4.47	S	3.64	P	4.56	1.39
E	4.95	D	3.96	K	3.25	F	3.59	P	4.05	Y	3.83	Q	3.75	G	3.95	V	2.76	R	4.28	1.38
I	4.41	N	2.68	Y	3.01	S	3.42	Y	3.87	T	3.64	P	3.60	N	3.47	I	2.55	S	3.92	1.01
H	3.84	K	2.51	D	2.88	M	3.32	R	3.63	P	3.41	R	3.58	P	3.09	T	2.45	F	3.49	1.25
Q	2.41	R	2.48	C	2.85	R	2.97	S	3.16	S	3.39	T	3.37	Q	2.62	Y	2.45	N	3.45	1.13
K	2.06	M	2.28	E	2.62	N	2.77	N	2.58	M	3.13	G	3.17	Y	2.26	F	2.15	Y	2.81	0.91
M	1.71	Q	2.22	M	2.37	Y	2.74	M	1.90	Q	3.10	S	2.21	F	2.19	Q	2.15	Q	2.30	0.83
F	1.43	Y	1.98	N	2.33	Q	1.79	W	1.35	N	2.90	M	2.20	M	1.06	H	1.97	M	2.11	0.79
Y	1.34	H	1.80	Q	1.89	C	1.68	H	1.26	H	2.01	H	1.64	H	0.77	M	1.05	H	1.84	0.85
W	0.34	C	1.06	H	1.82	H	1.42	C	1.20	C	1.60	W	1.22	W	0.62	W	0.47	C	1.18	0.83
C	0.23	W	0.95	W	0.43	W	0.81	Q	0.82	W	0.78	C	1.17	C	0.62	C	0.17	W	0.78	0.35

Average distance of RAAs to their C/ACs

Å	19.13	21.27	23.12	24.00	26.19	25.40	26.74	26.99	29.43	-
SEM	2.59	2.14	2.19	2.48	2.88	2.16	3.11	3.33	3.01	-
n/421	95	64	49	41	33	26	23	19	72	-

G	Glycine	L	Leucine	E	Glutamic acid	
A	Alanine	I	Isoleucine	K	Lysine	
		V	Valine	D	Aspartic acid	

*RAA = Recovered-amino acid; **CZ = evolutionary conservation zone, ***C/AC = catalytic/active center; fumarate hydratase C/AC *distance* values are omitted

Notes: *Each Domain includes: 1) the '% RAA Occupancy' of the Domains contents and 2), the 'Average Euclidean Distance (Å) …..' of the RAA's Cαs to their assigned C/AC Anchor-atom ± s.e.m. in each of nine evolutionary conservation zones (CZ) and 3) in the three right-hand columns, the averaged % Occupancies of the respective Domain. In Part D, we report the massed analysis of all the % Occupancies of all Recovered-Amino Acid (RAAs) sites of the eight oTCA cycle enzymes across Domains. The sections emphasize the differences of individual RAA % Occupancy trends in pooled analyses to those in individual consecutive ranked evolutionary conservation zones (CZ) with their concomitant distances (Å) to respective C/ACs.

Table 5. RAAs: their % Occupancy and distance to their catalytic/active centers (Å) (12,583) in sequences (5,485) of 33-Bacteria enzymes separated by their geochronological and evolutionary conservation assignments

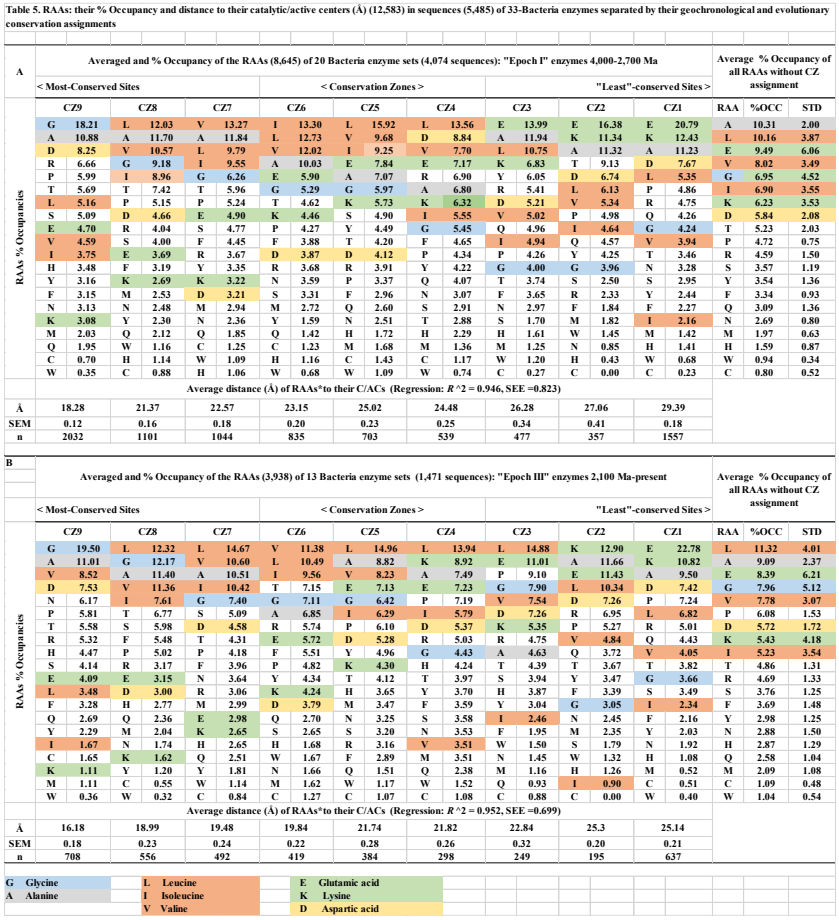

Table 5. RAAs: their % Occupancy and distance to their catalytic/active centers (Å) (12,583) in sequences (5,485) of 33-Bacteria enzymes separated by their geochronological and evolutionary conservation assignments

A

Averaged and % Occupancy of the RAAs (8,645) of 20 Bacteria enzyme sets (4,074 sequences): "Epoch I" enzymes 4,000-2,700 Ma

< Most-Conserved Sites						< Conservation Zones >						"Least"-conserved Sites >						Average % Occupancy of all RAAs without CZ assignment		
CZ9		CZ8		CZ7		CZ6		CZ5		CZ4		CZ3		CZ2		CZ1		RAA	%OCC	STD
G	18.21	L	12.03	V	13.27	I	13.30	L	15.92	L	13.56	E	13.99	E	16.38	E	20.79	A	10.31	2.00
A	10.88	A	11.70	A	11.84	L	12.73	V	9.68	D	8.84	A	11.94	K	11.34	K	12.43	L	10.16	3.87
D	8.25	V	10.57	L	9.79	V	12.02	I	9.25	V	7.70	L	10.75	A	11.32	A	11.23	E	9.49	6.06
R	6.66	G	9.18	I	9.55	A	10.03	E	7.84	E	7.17	K	6.83	T	9.13	D	7.67	V	8.02	3.49
P	5.99	I	8.96	G	6.26	E	5.90	A	7.07	R	6.90	Y	6.05	D	6.74	L	5.35	G	6.95	4.52
T	5.69	T	7.42	T	5.96	G	5.29	G	5.97	A	6.80	R	5.41	L	6.13	P	4.86	I	6.90	3.55
L	5.16	P	5.15	P	5.24	T	4.62	K	5.73	K	6.32	D	5.21	V	5.34	R	4.75	K	6.23	3.53
S	5.09	D	4.66	E	4.90	K	4.46	S	4.90	I	5.55	V	5.02	P	4.98	Q	4.26	D	5.84	2.08
E	4.70	R	4.04	S	4.77	P	4.27	Y	4.49	G	5.45	Q	4.96	I	4.64	G	4.24	T	5.23	2.03
V	4.59	S	4.00	F	4.45	F	3.88	T	4.20	F	4.65	I	4.94	Q	4.57	V	3.94	P	4.72	0.75
I	3.75	E	3.69	R	3.67	D	3.87	D	4.12	P	4.34	P	4.26	Y	4.25	T	3.46	R	4.59	1.50
H	3.48	F	3.19	Y	3.35	R	3.68	R	3.91	Y	4.22	G	4.00	G	3.96	N	3.28	S	3.57	1.19
Y	3.16	K	2.69	K	3.22	N	3.59	P	3.37	Q	4.07	T	3.74	S	2.50	S	2.95	Y	3.54	1.36
F	3.15	M	2.53	D	3.21	S	3.31	F	2.96	N	3.07	F	3.65	R	2.33	Y	2.44	F	3.34	0.93
N	3.13	N	2.48	M	2.94	M	2.72	Q	2.60	S	2.91	N	2.97	F	1.84	F	2.27	Q	3.09	1.36
K	3.08	Y	2.30	N	2.36	Y	1.59	N	2.51	T	2.88	S	1.70	M	1.82	I	2.16	N	2.69	0.80
M	2.03	Q	2.12	Q	1.85	Q	1.42	H	1.72	H	2.29	H	1.61	W	1.45	M	1.42	M	1.97	0.63
Q	1.95	W	1.16	C	1.25	C	1.23	M	1.68	M	1.36	M	1.25	N	0.85	H	1.41	H	1.59	0.87
C	0.70	H	1.14	W	1.09	H	1.16	C	1.43	C	1.17	W	1.20	H	0.43	W	0.68	W	0.94	0.34
W	0.35	C	0.88	H	1.06	W	0.68	W	1.09	W	0.74	C	0.27	C	0.00	C	0.23	C	0.80	0.52

RAAs % Occupancies

Average distance (Å) of RAAs*to their C/ACs (Regression: R^2 = 0.946, SEE =0.823)

Å	18.28	21.37	22.57	23.15	25.02	24.48	26.28	27.06	29.39
SEM	0.12	0.16	0.18	0.20	0.23	0.25	0.34	0.41	0.18
n	2032	1101	1044	835	703	539	477	357	1557

B

Averaged and % Occupancy of the RAAs (3,938) of 13 Bacteria enzyme sets (1,471 sequences): "Epoch III" enzymes 2,100 Ma-present

< Most-Conserved Sites						< Conservation Zones >						"Least"-conserved Sites >						Average % Occupancy of all RAAs without CZ assignment		
CZ9		CZ8		CZ7		CZ6		CZ5		CZ4		CZ3		CZ2		CZ1		RAA	%OCC	STD
G	19.50	L	12.32	L	14.67	V	11.38	L	14.96	L	13.94	L	14.88	K	12.90	E	22.78	L	11.32	4.01
A	11.01	G	12.17	V	10.60	L	10.49	A	8.82	K	8.92	E	11.01	A	11.66	K	10.82	A	9.09	2.37
V	8.52	A	11.40	A	10.51	I	9.56	V	8.23	A	7.49	P	9.10	E	11.43	A	9.50	E	8.39	6.21
D	7.53	V	11.36	I	10.42	T	7.15	E	7.13	E	7.23	G	7.90	L	10.34	D	7.42	G	7.96	5.12
N	6.17	I	7.61	G	7.40	G	7.11	G	6.42	P	7.19	V	7.54	D	7.26	P	7.24	V	7.78	3.07
P	5.81	T	6.77	S	5.09	A	6.85	I	6.29	I	5.79	D	7.26	R	6.95	L	6.82	P	6.08	1.53
T	5.58	S	5.98	D	4.58	R	5.74	P	6.10	D	5.37	K	5.35	P	5.27	R	5.01	D	5.72	1.72
R	5.32	F	5.48	T	4.31	E	5.72	D	5.28	R	5.03	R	4.75	V	4.84	Q	4.43	K	5.43	4.18
H	4.47	P	5.02	P	4.18	F	5.51	Y	4.96	G	4.43	A	4.63	Q	3.72	V	4.05	I	5.23	3.54
S	4.14	R	3.17	F	3.96	P	4.82	K	4.30	H	4.24	T	4.39	T	3.67	T	3.82	T	4.86	1.31
E	4.09	E	3.15	N	3.64	Y	4.34	T	4.12	T	3.97	S	3.94	Y	3.47	G	3.66	R	4.69	1.33
L	3.48	D	3.00	R	3.06	K	4.24	H	3.65	Y	3.70	H	3.87	F	3.39	S	3.49	S	3.76	1.25
F	3.28	H	2.77	M	2.99	D	3.79	M	3.47	F	3.59	Y	3.04	G	3.05	I	2.34	F	3.69	1.48
Q	2.69	Q	2.36	E	2.98	Q	2.70	N	3.25	S	3.58	I	2.46	N	2.45	F	2.16	Y	2.98	1.25
Y	2.29	M	2.04	K	2.65	S	2.65	S	3.20	N	3.53	F	1.95	M	2.35	Y	2.03	N	2.88	1.50
I	1.67	N	1.74	H	2.65	H	1.68	R	3.16	V	3.51	W	1.50	S	1.79	N	1.92	H	2.87	1.29
C	1.65	K	1.62	Q	2.51	W	1.67	F	2.89	M	3.51	N	1.45	W	1.32	H	1.08	Q	2.58	1.04
K	1.11	Y	1.20	Y	1.81	N	1.66	Q	1.51	Q	2.38	M	1.16	H	1.26	M	0.52	M	2.09	1.08
M	1.11	C	0.55	W	1.14	M	1.62	W	1.17	W	1.52	Q	0.93	I	0.90	C	0.51	C	1.09	0.48
W	0.36	W	0.32	C	0.84	C	1.27	C	1.07	C	1.08	C	0.88	C	0.00	W	0.40	W	1.04	0.54

RAAs % Occupancies

Average distance (Å) of RAAs*to their C/ACs (Regression: R^2 = 0.952, SEE =0.699)

Å	16.18	18.99	19.48	19.84	21.74	21.82	22.84	25.3	25.14
SEM	0.18	0.23	0.24	0.22	0.28	0.26	0.32	0.20	0.21
n	708	556	492	419	384	298	249	195	637

G	Glycine	L	Leucine	E	Glutamic acid	
A	Alanine	I	Isoleucine	K	Lysine	
		V	Valine	D	Aspartic acid	

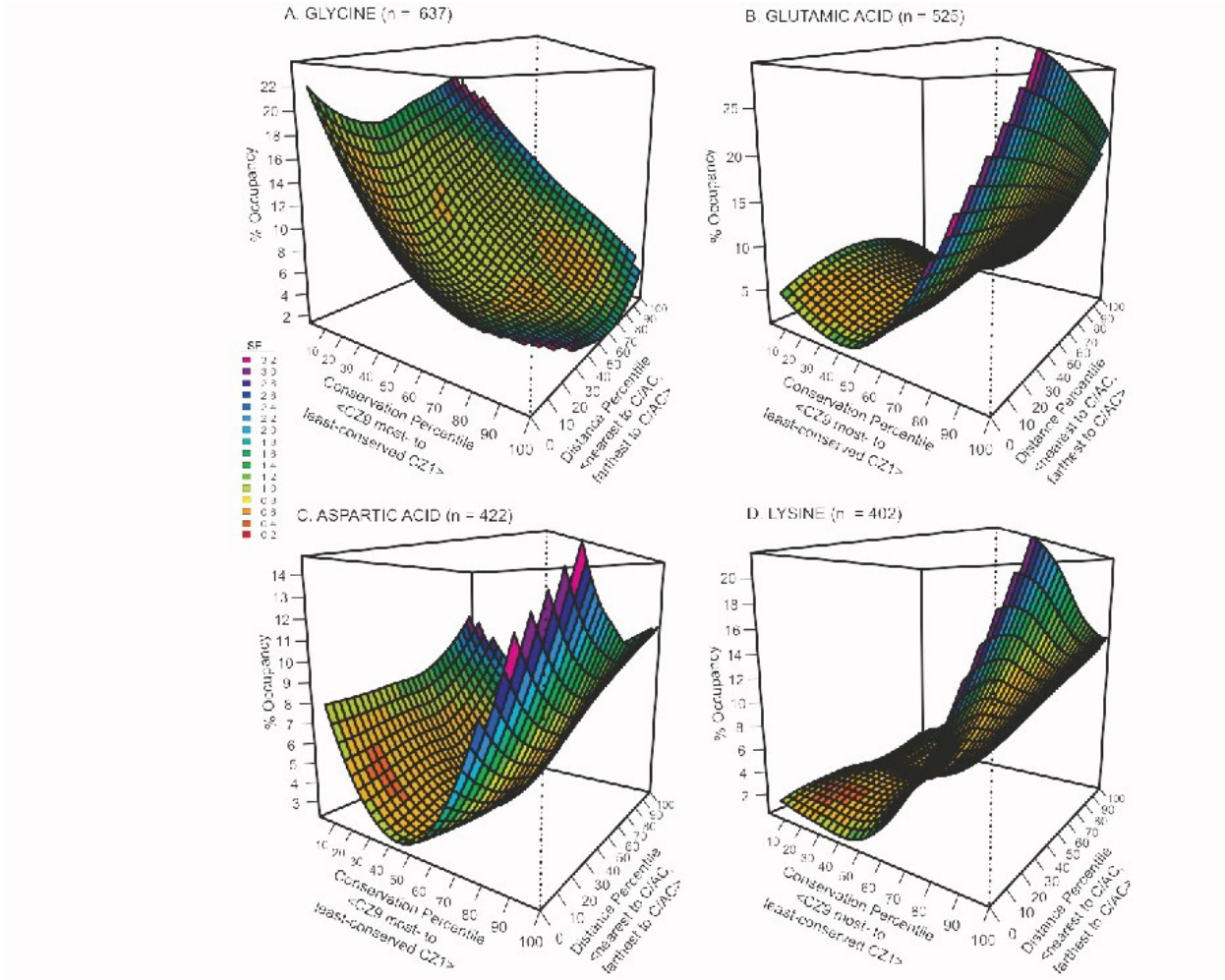

Figure 1A. LOESS surface plots of Recovered-Amino Acid %Occupancy sites as a function of their percentiles of conservation zone and distance to respective C/ACs Percentiles of distance and conservation were calculated within 15 data sets. LOESS surfaces (span = 0.75) and standard errors (s.e.) for the linked data are shown in the attached color standards insert (R Development Core Team, 2011). All of these oTCA surface plots mimic quite closely those also reported for all glycolysis enzymes (Pollack et al., 2013): A: glycine is most abundant in sites that are both closest- and most-distant to their C/ACs and least abundant in low evolutionary conservation zone (CZ) sites. B and D: Glutamic acid and lysine are most abundant in low conservation sites CZ2 and CZ1 furthest from the C/AC. C: Aspartic acid shows high abundance in high and low conservation sites with apparently highest abundance at lowest distances to their C/ACs, however, these values have the highest s.e.

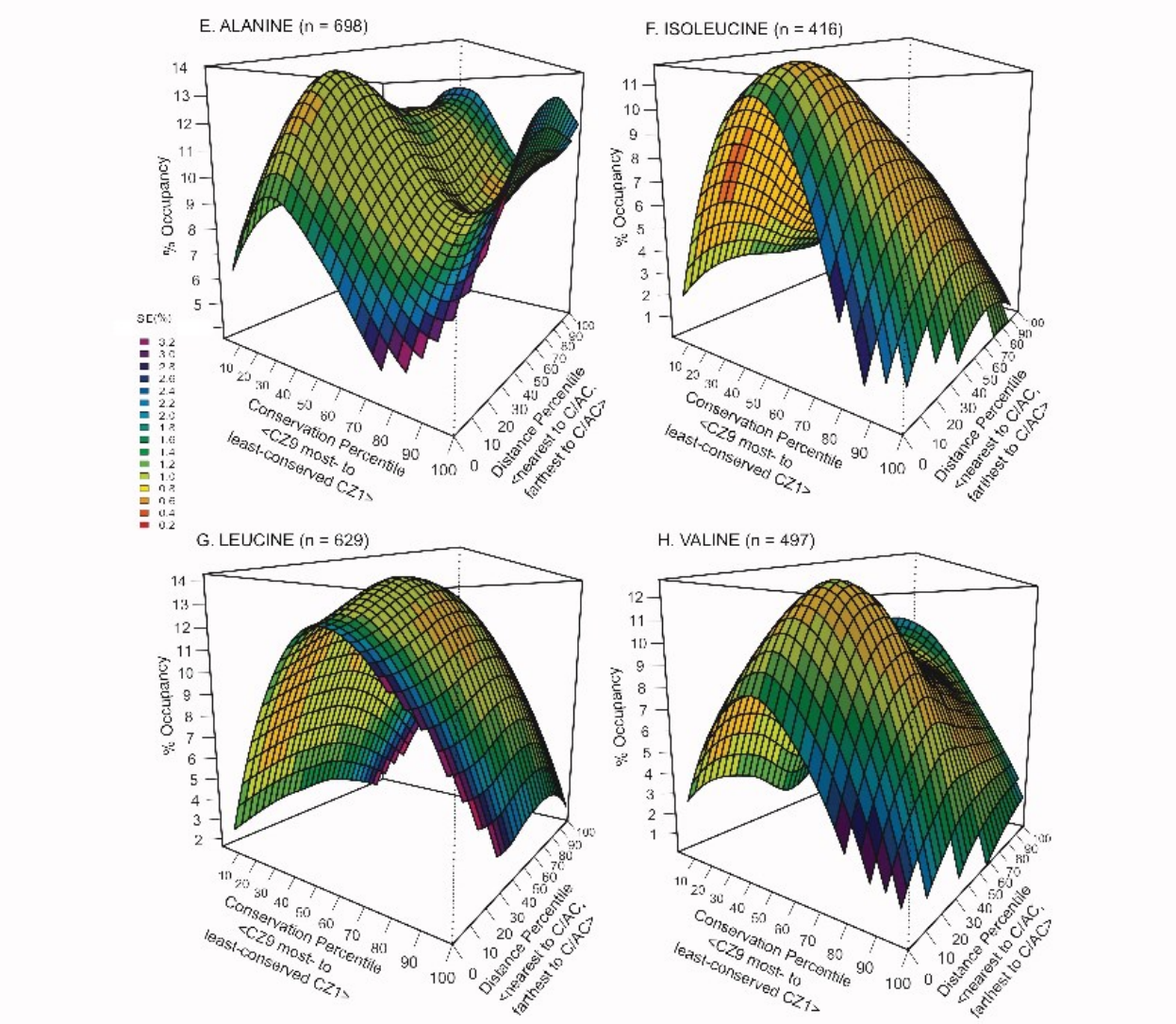

Figure 1B (Fig. 1A continued). E: Alanine sites may be more abundant at their C/AC and periphery. F:, G:, H: Isoleucine, leucine and valine sites are similar as they all demonstrate highest abundances at mid-conservation regions, however, the highest abundance of isoleucine falls off rapidly with increasing distance from the C/AC, while that for valine diminishes with distance from C/ACs. The more reddish colors down through the dark orange areas with highest statistical reliability have SEs in the 0.2 % to 0.4 % range, while the dark green and blue up to the deep purple colors are in areas of lowest reliabilities with SEs of no more than, 1.4 %.

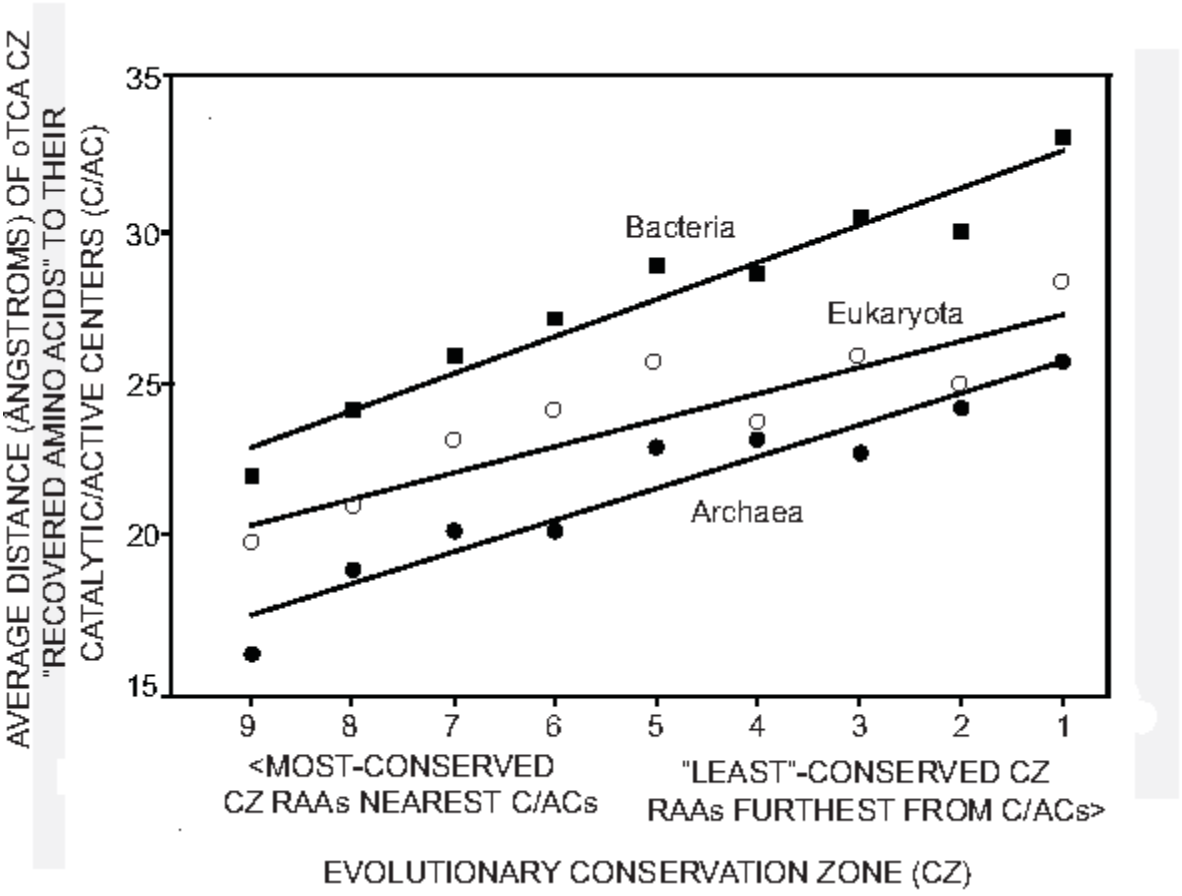

Figure 2. Recovered-Amino Acid (RAAs) sites (6,831) of the eight oTCA cycle enzymes in Archaea, Bacteria or Eukaryota vs their site's distances (Å) to their respective catalytic/active centers (C/AC) For taxonomic identifications, See, Table 1: three Archaea enzymes, 948 RAAs, $R^2 = 0.921$, RMSE (Å) = 0.904, slope =1.055, 95%CI = 0.78-1.33; seven Bacteria enzymes, 3,747 RAAs, $R^2 = 0.945$, RMSE (Å) = 0.865, slope = 1.230, 95% CI = 0.99-1.49; five Eukarya enzymes, 2,136 RAAs, $R^2 = 0.780$, RMSE (Å) = 1.368; slope = 0.879, 95% CI = 0.46-1.30. The parameters of the plot of all 6,831 RAAs of all Domains are: $R^2 = 0.85$, RMSE (Å) = 1.43 ± 0.57, slope = 1.26 ± 0.35 its p-value ≈ 0.0001, 95% CI = 0.25 to 2.52. In all Domains a decrease in averaged evolutionary conservation is directly related to the increase in Euclidean distance (Å) from respective functional loci (C/AC).

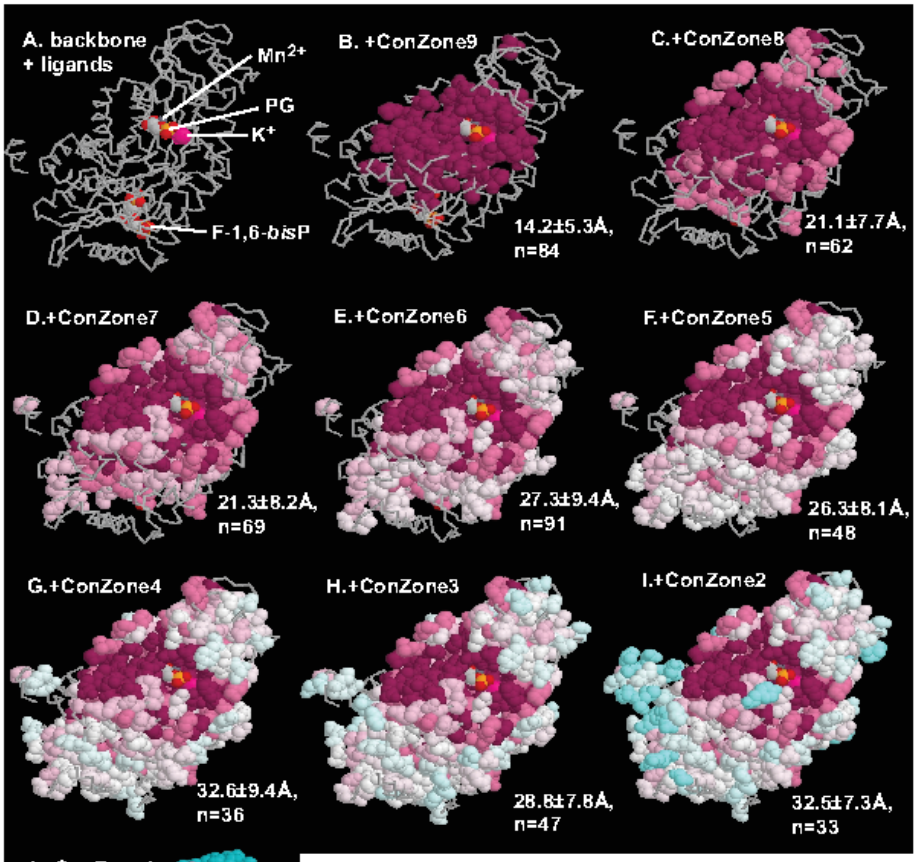

Figure 3. Assembly of diffuse evolutionary conserved aggregations (DAGs) of the averaged MSA consensus site RAAs of 132 pyruvate kinase (EC 2.7.1.40) (PYK) sequences representing Archaea, Bacteria, and Eukaryota revealed by ConSurf and Protein Explorer using the PDB PYK "Scaffold" PDB:1a3w, KPY1_YEAST, *Saccharomyces cerevisiae* (see, Methods). Panel A shows the backbone, catalyically involved ligands (Mn^{2+}, PG, K^+) and the distant native allosteric effector F-1,6-*bis*P. The Panels B-J show the locations of 558 RAA amino acids in sequentially decreasing conserved aggregations that constitute DAGs. Panel B adds the MSA consensus sites that are most-conserved (ConZone9) and colored deepest-magenta according to the scale beneath Panel J. The succeeding Panels C-J show accretion of ConZone8 through ConZone1. The "least"-conserved postions of (ConZone1) are colored deepest-blue or turquoise in Panel J that depicts the complete PK model. The data in the lower right of each Panel, calculated by the Yasara program, is the average distance in Å (± SD, n) from all MSA consensus amino acid Cα positions of that particular ConZone to the same Anchor-atom (Mn^{2+}) in the Scaffold's reported mechanistic center shown in Panel A (see, Methods).

PG = 2-phosphoglycolic acid; F-1,6-*bis*P = fructose-1,6-*bis*phosphate

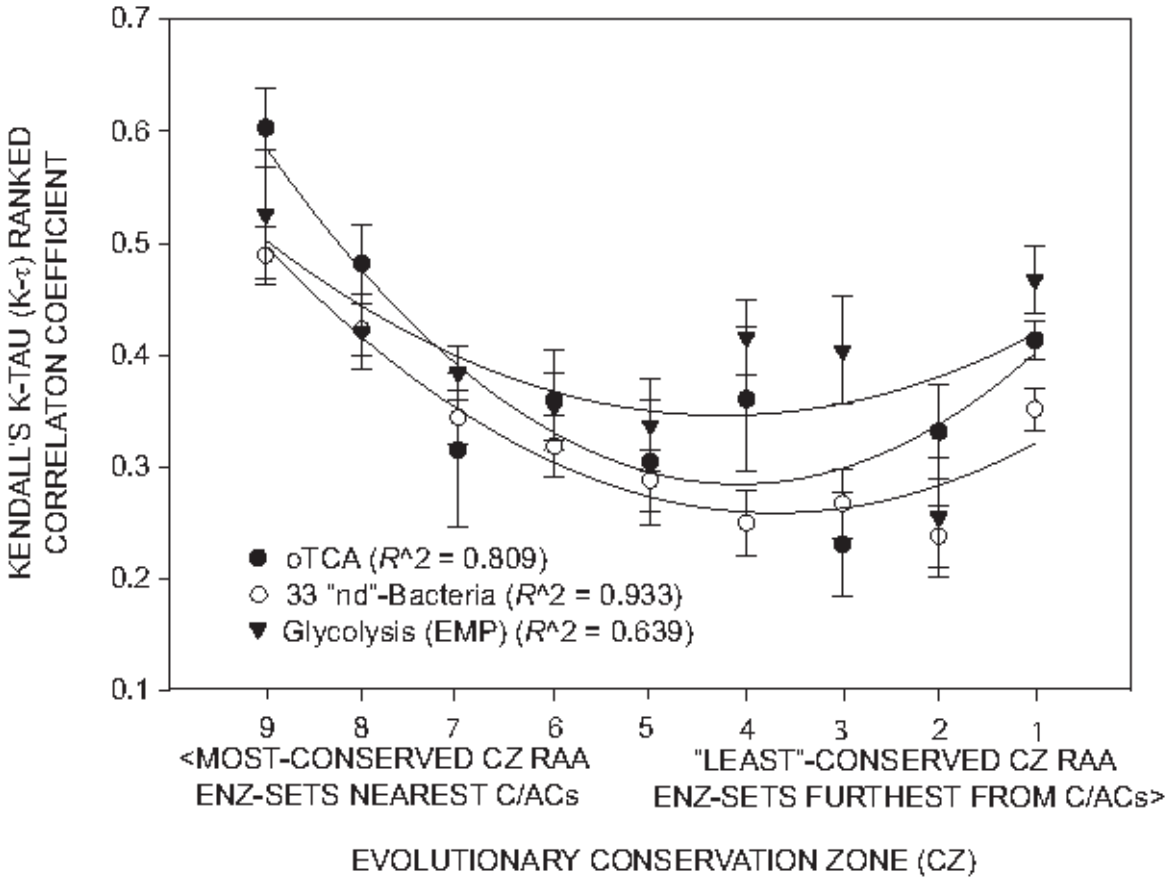

Figure 4. The ranked dominance of the most-conserved evolutionary conservation zone (CZ9) in comparisons to Trifonov's putative temporal sequence (TOAE) Differences between Epochs of the RAAs of averaged analyses of: 1) the eight oTCA enzymes of Table 1, and 2) the 33-Bacteria examples of Table 2 and 3) the 10 glycolysis enzymes (op. cit.). See, Sections *3.1.3*. and *4*. for more details.

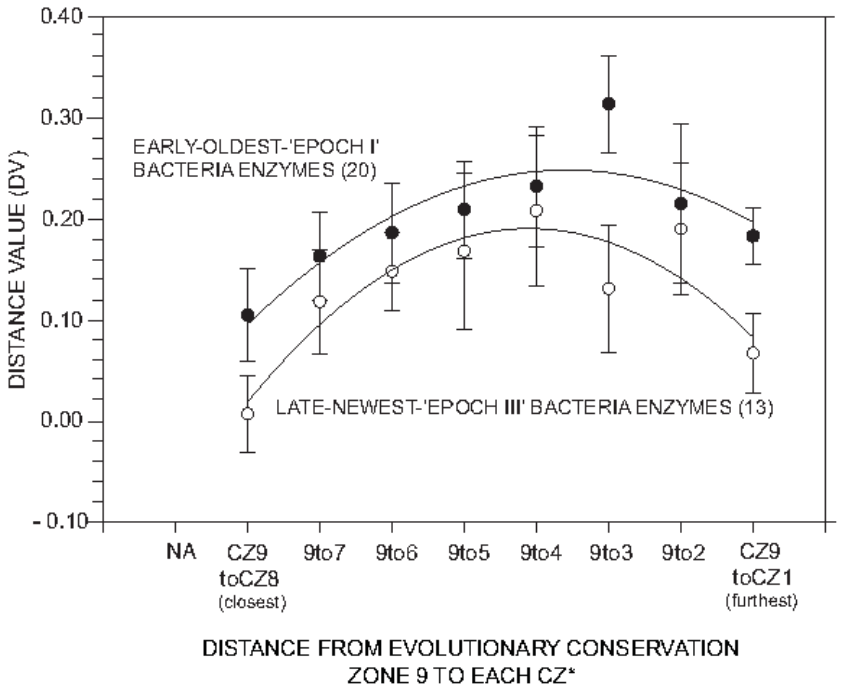

Figure 5. Distinguishing 'Epoch I' from 'Epoch III'. Each point represents the average of the differences between two K-t correlations, one correlation between CZ9's average distance to its functional center vs TOAE's ranked sequence minus the other correlation between the succeeding CZ's average distance to its functional center vs TOAE's ranked data. See, Section *3.2.3*. for more details.