Philippine Journal of Science
140 (1): 105-113, June 2011
ISSN 0031 - 7683

Short Communication

Proto-ubiquitin: A Bayesian Prediction of an Ancient Protein During the Prokaryotic-Eukaryotic Transition

Custer C. Deocaris[1*], **Lovette F. Cunanan**[2], **Richelda A. Galapia**[2], and **Marla A. Endriga**[2]

[1]College of Home Economics, University of the Philippines, Diliman, Quezon City
[2]Department of Biology, College of Arts and Sciences, University of the Philippines Manila, Ermita, 1000 Manila, Philippines

Only recently have biologists been able to apply mathematical and biochemical tools to preview lifestyles of ancient life forms and "travel back in time." In this paper, we describe an ancestral reconstruction of ubiquitin to determine its molecular properties during the rise of the eukaryotes. Although ubiquitin is one of the most conserved proteins in eukaryotes, no ubiquitin homolog has been found in prokaryotic genomes sequenced thus far. In an attempt to derive the ancestral ubiquitin, or proto-ubiquitin (proto-Ub), we applied Bayesian statistical theory to estimate posterior probabilities of protein sequences from a minimum evolution tree of 30 extant species. The inferred ancestral sequence was 100% homologous with the ubiquitin of *Brugia malayi*, a parasitic nematode. Among its 76 amino acids, only nine residues have undergone amino acid modification. As no major structural and functional changes happened during the evolution of ubiquitin, we hypothesize that the stressful conditions that led to the creation of this gene after the "Great Oxidation Event" 2.4 billion years ago may have already been "buffered" to date.

Key Words: bioinformatics, eukaryotes, paleobiochemistry, planetary biology, protein evolution

INTRODUCTION

What was Earth like when eukaryotes were just starting to emerge from the oceans? Planetary biologists widely believe that proteotoxic stress from increasing oxygen levels and ocean temperatures around 2 billion (Gyr) years ago prompted the origin of the first eukaryote via endosymbiosis (reviewed recently by Deocaris et al. 2010), and that the steady build-up of oxygen by cyanobacterial photosynthesis helped create the Earth's early atmosphere that shielded organisms from deadly cosmic rays. This eventually ushered in the Great Oxidation Event (GOE, also called the oxygen catastrophe or oxygen crisis), when levels of free oxygen in the atmosphere began to shoot up. This major environmental change happened around 2.4 Gyr ago, coinciding with the largest extinction event in the Earth's history.

Among the oldest pieces of evidence establishing the occurrence of photosynthesis are hydrocarbons extracted from shale in the Pilbara Craton, Australia, indicative of oxygen inundation by the cyanobacteria; and the eventual rise of eukaryotic life. These occurred approximately 2.15 Gyr and 1.78-1.68 Gyr ago, respectively (Han & Runnegar 1992; Rasmussen et al. 2008), consistent with the origin of mitochondria which is estimated to have occurred 1.8 Gyr ago. The latter is based on a molecular clock, and is based on the assumption that the last common mitochondrial bacterium lived at a time close to the endosymbiotic event between α-proteobacteria and eukaryotic cells (Yang et al.

*Corresponding author: cdeocaris@gmail.com

1985). Geological theory and evidence further suggest that the ancient ocean had temperatures similar to those of hot springs. As the ocean started cooling down from 3.5 to 0.5 Gyr ago (Knauth & Lowe 2003), life may have responded by adapting its range of growth temperatures to correspond that of its environment. More sophisticated eukaryote phytoplankton then came about 500 Myrs later in the cooler Proterozoic oceans (Falkowski et al. 2004), and throughout the time that eukaryotic cells have been present, atmospheric oxygen levels have varied from about 15 to 35%. When eukaryotic cells began to actively convert free atmospheric oxygen into water in the electron transport chain, the amount of oxygen gradually began to stabilize to present-day levels of 21%. This occurred only towards the end of the Proterozoic era, 0.6-0.7 Gyr years ago. Interestingly, this also coincides with the appearance of multicellular and larger animals such as the *Ediacara* (Berner 1999; Acquisti et al. 2007).

It therefore appears that protein quality control was a major issue to advance and sustain life during the prokaryotic-eukaryotic transition. There arose a need for perpetual, rapid, and selective removal of damaged proteins coupled to a re-synthesis of molecules, since many of the nascent proteins were also misfolded.

Ubiquitin, a small protein consisting of 76 amino acids, occurs universally among the eukaryotes. It has not been detected so far in the other two branches of life, the archae and the bacteria (see Figure 1C). Its amino acid sequence has been highly conserved throughout evolution, which may be a necessary consequence of its astonishing multifunctionality. It was first documented by Poole et al. (1978), while investigating the mode of action of anti-malarial drugs. They found that lysosome, for a long time believed to destroy intracellular proteins, is actually not the key player in this process. Instead, proteins within a cell are broken down by a non-lysosomal mechanism now known as the ubiquitin system, so named by Hershko et al. in the late 1970s (Ciechanover 2006).

Ubiquitin, in conjunction with 26S proteasome, a multicatalytic protease complex, forms the ubiquitin-proteasome system (UPS). Its best characterized biological role involves covalent conjugation of multi-ubiquitin chains or monomers of ubiquitin to target proteins, signaling their proteolysis or regulating other nondegradative cellular functions, respectively. Degradation of polyubiquitinated substrates is carried out by a large protease complex, the 26S proteasome, that does not generally recognize non-modified substrates. Such process is central to many cellular functions, primarily those that govern protein homoestasis, ribosome biogenesis, DNA repair, cell-cycle progression, stress response, and cell survival (Zwickl et al. 1999).

The search for the origins of ubiquitin has been undertaken by previous investigators using the traditional structure and sequence similarity data mining strategy. However, there is a greater likelihood of chance detection of sequence similarity if proteins no more than 100 amino acids long are analyzed (Bienkowska et al. 2003). In our study, we used Bayesian tools to determine the origin of ubiquitin by deriving its ancestral amino acid sequence, we termed as 'proto-ubiquitin' (proto-Ub), and comparing it with the contemporary ubiquitin. We believe that this will contribute to filling in gaps of information on the molecular evolution of this vital protein.

METHODS

Ubiquitin sequences of 30 extant species were used in the study. They were obtained from Uniprot database (http://www.expasy.ch/cgi-bin/sprot-search-ful; Boeckmann et al. 2003) and were chosen to represent all the taxa that have available ubiquitin sequence. The organisms were: *Acanthamoeba castellanii* (amoeba), *Acetabularia cliftonii* (a green alga), *Aglaothamnion neglectum* (a red alga), *Bos taurus* (bovine), *Caenorhabditis elegans* (nematode), *Canis familiaris* (dog), *Chlamydomonas reinhardtii* (a green alga), *Coprinus congregatus* (inky cap fungus), *Drosophila melanogaster* (fruit fly), *Equus caballus* (horse), *Euplotes eurystomus* (a ciliate), *Felis silvestris catus* (cat), *Gallus gallus* (chicken), *Geodia cydonium* (sponge), *Homo sapiens* (human), *Ictalurus punctatus* (channel catfish), *Leishmania major* (protozoan), *Mus musculus* (mouse), *Nicotiana sylvestris* (wood tobacco), *Ophiophagus hannah* (king cobra), *Oryza sativa* subsp. japonica (rice), *Pan troglodytes* (chimpanzee), *Plutella xylostella* (diamondback moth), *Saccharomyces cerevisiae* (Baker's yeast), *Solanum lycopersicum* (tomato), *Spodoptera frugiperda* (fall armyworm), *Strongylocentrotus purpuratus* (purple sea urchin), *Trypanosoma brucei brucei* (a protozoan), *Xenopus laevis* (African clawed frog) and *Zea mays* (maize).

The computational approach used in this study was modified from Gaucher et al (2003). The amino acid sequences of the species mentioned were retrieved from Uniprot database (Boeckmann et al. 2003) and aligned with CLUSTALW (Chenna et al. 2003) via MEGA4 (Tamura et al. 2007). Using the gamma distribution and JTT model, the distance matrix was obtained and was utilized to create a phylogenetic tree using the minimum evolution criterion. For each site of the inferred sequence at a phylogenetic node, posterior values for 20 amino acids represent the probability that a particular amino acid occupied a specific site in the protein during its evolutionary history. This posterior probability distribution is then calculated using PAML4 (Yang 2007) based on a

matrix of amino-acid replacement probabilities, amino acid equilibrium (stationary) frequencies, phylogenetic branch lengths, and site-specific replacement rates. The ancestral protein comprised the sequence of amino acids that garnered the highest posterior probability at each site (Gaucher et al. 2003; Gaucher et al. 2008).

A protein-protein BLAST (Altschul et al. 1990) search against the NCBI database (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was done to find other proteins homologous with that of the derived ancestral sequence. ProtParam (http://www.expasy.ch/tools/protparam.html; Gasteiger et al. 2005) and Swiss-Model (http://swissmodel.expasy.org//SWISS-MODEL.html; Guex & Peitsch 1997; Schwede et al. 2003; Arnold et al. 2006) were used in characterizing the present-day ubiquitins and the derived ancestral sequence. Characterization involved the analysis of the primary, secondary, and tertiary structure of the proteins of interest. ProtParam (Gasteiger et al. 2005) was used for the computation of the physico-chemical properties of the proteins such as molecular weight and theoretical pI. The three-dimensional structure of the ancestral sequence was generated using Swiss-Model (Guex & Peitsch 1997; Schwede et al. 2003; Arnold et al. 2006) and Rasmol (Sayle & Milner-White 1995) was employed to visualize it.

RESULTS

Derivation of proto-ubiquitin

The multiple protein sequence alignment of 30 ubiquitins is shown in Fig. 1A. The columns in red indicate the sites that have undergone changes. This protein is highly conserved as there are 67 out 76 sites that have invariable amino acids. Following Gaucher et al (2003), we then constructed a minimum evolution (ME) tree to illustrate the relationship of the ubiquitin sequences (Fig. 1B). The tree has a total branch length of 0.202 substitutions/site. It can be noted that two clades split from the root. The lower clade includes plants, fungi, and some protozoans, while the upper clade includes all chordates and three protozoans – *Leishmania major*, *Trypanosoma brucei brucei*, and *Aglaothamnion neglectum*. When compared with a phylogenetic tree from literature based on rRNA sequences (Fig. 1C), our results reveal that the groupings in both trees agree, albeit, with some exceptions. Such discrepancies are not unexpected as individual genes may not follow identical historical paths (Vogel 1997; Rambaut et al. 2001; Mace & Holden 2005).

PAML4 (Yang 2007) was used to generate the ancestral sequences. The amino acid sequences of the 30 species and the topology of the ME tree were used as input data. Four ancestral sequences were derived and are arbitrarily designated as nodes 31-34 in Fig. 2A. These correspond to the predecessor ubiquitin sequences of the species under consideration, the most primitive of which is node 31. This is the sequence we generated for the most ancient ubiquitin molecule, or the proto-Ub.

Proto-Ub characterization

BLAST (Altschul 1990) produced numerous sequences with significant alignments, all of which are eukaryotic. Two sequences were found to be identical to the derived ancestral sequence. They each have a bit score of 155 and an E-value of 8×10^{-37}. The two sequences have the same source organism, *Brugia malayi*: *Brugia malayi* (1) was described by Ghedin et al. in 2007 while *Brugia malayi* (2) is derived from an unpublished paper. We also constructed the three-dimensional model of proto-Ub using Swiss-Model (Guex & Peitsch 1997; Schwede et al. 2003; Arnold et al. 2006). There was no observed change in the sites critical for ubiquitin function. The amino acids important to the function of the protein include the surface lysine residues, especially Lys48, and Gly76, although the latter is not in the 3-D model generated. Lysine residues serve as points of attachment for Gly76 of other ubiquitins, particularly during poly-ubiquitination. Gly76, on the other hand, is for linking to other ubiquitins, regulatory enzymes (E1, E2, and E3), and target proteins (Fig. 2C).

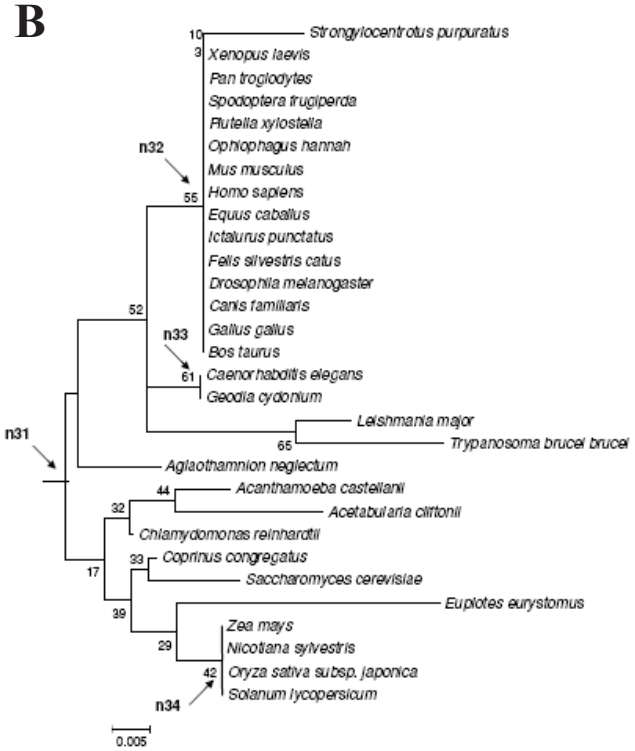
Physico-chemical characteristics of proto-Ub and modern groups of species are shown in Table 1. For each parameter, values for proto-Ub do not deviate much from those for modern sequences. Each species has a unique molecular weight except for the proto-Ub and *Brugia malayi,* which exhibit similar values for all the parameters considered. *T. brucei brucei* has the lowest molecular weight of 8506.8 g/mol while *A. castellanii* has the highest at 8595.8 g/mol. In terms of the theoretical pI, majority of the species have a value of 6.56 while few possess a slightly higher value, particularly *T. brucei brucei* and *L. major* with 6.57. *E. eurystomus* has a slightly lower value at 6.55. These values indicate that ubiquitin is stable at physiological pH.

The instability index (II) of a protein indicates its stability in vitro, while the aliphatic index indicates the relative volume occupied by the aliphatic side chains of alanine, valine, isoleucine, and leucine (Gasteiger et al. 2005). Proteins having an II value less than 40 are classified as stable, and those with value greater than 40 are unstable. The instability indices indicate that all proteins are stable except for that of *S. purpuratus*. Meanwhile, values for the aliphatic index of proto-Ub range from 98.68 to 103.95.

A

```
                        10         20         30         40         50         60         70
human           MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
bovin           MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
canfa           MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
drome           MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
horse           MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
felca           MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
chick           MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
ictpu           MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
mouse           MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
ophha           MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
pantr           MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
pluxy           MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
spofr           MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
xenla           MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
caeel           MQIFVKTLTG KTITLEVEAS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
geocy           MQIFVKTLTG KTITLEVEAS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
strpu           MQIFVKTLTG KTITLEVEPS DSIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
aglne           MQIFVKTLTG KTITLEVESS DTIENVKTKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
chlre           MQIFVKTLTG KTITLEVESS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV LRLRGG
copco           MQIFVKTLTG KTITLEVESS DTIDNVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
leima           MQIFVKTLTG KTIALEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EEGRTLSDYN IQKESTLHLV LRLRGG
acaca           MQIFVKTLTG KTITLEVESS DTIENVKQKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV LRLRGG
nicsy           MQIFVKTLTG KTITLEVESS DTIDNVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV LRLRGG
orysj           MQIFVKTLTG KTITLEVESS DTIDNVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV LRLRGG
sollc           MQIFVKTLTG KTITLEVESS DTIDNVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV LRLRGG
maize           MQIFVKTLTG KTITLEVESS DTIDNVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV LRLRGG
yeast           MQIFVKTLTG KTITLEVESS DTIDNVKSKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
acecl           MQIFVKTLTG KTITLEVESS DTVENVKSKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV LRLRGG
trybb           MQIFVKTLTG KTIALEVEAS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EEGRTLADYN IQKESTLHLV LRLRGG
eupeu           MQIFVKTLTG KTITLDVEQS DTIDNVKTKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV LRLRGG
```

B



C

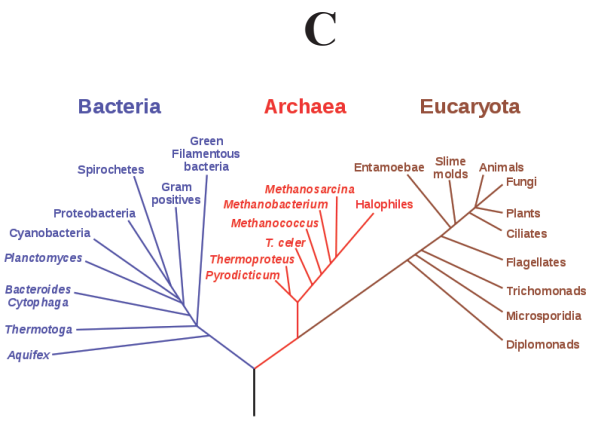


Figure 1. Evolutionary relationship of ubiquitin. (A) Multiple sequence alignment of 30 species using CLUSTALW (Chenna et al. 2003). The ubiquitin of each species considered is 76 amino acids long. The nine variable sites are in red. (B-C) Phylogenetic trees based on ubiquitin (B) and on rRNA sequences (C). (B) The evolutionary history based on 30 sequences considered in the study was inferred using the Minimum Evolution method (Rzhetsky & Nei 1992). The optimal tree with total branch length = 0.202 with bootstrap values is shown. Highlighted nodes (n31-n34) are those which correspond to ancestral sequences in Figure 2A. The evolutionary distances were computed using the JTT matrix-based method and are in the number of amino acid substitutions per site. The rate variation among sites was modeled with a gamma distribution (shape parameter = 1). The ME tree was searched using the Close-Neighbor-Interchange (CNI) algorithm (Nei & Kumar 2000) at a search level of 1. The Neighbor-joining algorithm (Saitou & Nei 1987) was used to generate the initial tree. All positions containing gaps and missing data were eliminated from the dataset. There were 76 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 (Tamura et al. 2007). (C) Evolutionary relationship among selected major groups of organisms inferred from ribosomal RNA sequence comparison used by Carl Woese (Figure retrieved from: http://rst.gsfc.nasa.gov/Sect20/A12.html).

A

```
node31    MQIFVKTLTG KTITLEVESS DTIKNVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
node32    MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
node33    MQIFVKTLTG KTITLEVEAS DTIKNVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
node34    MQIFVKTLTG KTITLEVESS DTIDNVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV LRLRGG
```

B

```
                          10         20         30         40
Brugia malayi (1)      MGCLLHSALF GVWLVPLVGW VRMLWVHNHV KMQIFVKTLT
Brugia malayi (2)                                    MQIFVKTLT
Ancestral Sequence                                   MQIFVKTLT

                          40         50         60         70
Brugia malayi (1)      GKTITLEVES SDTIENVKAK IQDKEGIPPD QQRLIFAGKQ
Brugia malayi (2)      GKTITLEVES SDTIENVKAK IQDKEGIPPD QQRLIFAGKQ
Ancestral Sequence     GKTITLEVES SDTIENVKAK IQDKEGIPPD QQRLIFAGKQ

                          80         90        100        110
Brugia malayi (1)      LEDGRTLSDY NIQKESTLHL VLRLRGGGKK RKKKVYTTPK
Brugia malayi (2)      LEDGRTLSDY NIQKESTLHL VLRLRGGGKK RKKKVYTTPK
Ancestral Sequence     LEDGRTLSDY NIQKESTLHL VLRLRGG

                         120        130        140        150
Brugia malayi (1)      KNKHKRKKVK LAVLKYYKVD ENGKITRLRK ECASPSCGGG
Brugia malayi (2)      KNKHKRKKVK LAVLKYYKVD ENGKITRLRK ECASPSCGGG
Ancestral Sequence

                         160        170        180
Brugia malayi (1)      VFMASHQNRY YCGKCHRTLV MQDPKEKAIG KAK
Brugia malayi (2)      VFMASHQNRY YCGKCHRTLV MQDPKEKAIG KAK
Ancestral Sequence
```

C

Figure 2. Comparisons of reconstructed ancestral sequences. (A) Ancestral ubiquitin sequences derived using an empirical bayesian statistical framework (PAML4; Yang 2007). (B) Alignment of proto-ubiquitin with two *Brugia malayi* ubiquitin sequences derived from BLAST (Altschul et al. 1990). Residues in red indicate the three sequences are identical. (C) Model of proto-Ubiquitin generated using Swiss Model (Guex & Peitsch 1997, Schwede et al. 2003, Arnold et al. 2006) and visualized with RasMol (Sayle & Milner-White 1995). Lys48 is in green while Gly76 is not included in the model.

The Grand Average of Hydropathy (GRAVY) is derived through the summation of the hydropathy values of all the amino acids that constitute a protein, divided by the number of residues in the sequence (Gasteiger et al. 2005), and is indicative of a protein's solubility. Positive GRAVY means that the protein is hydrophobic while negative GRAVY signifies that the protein is hydrophilic. All the species under consideration have negative GRAVY and therefore are all hydrophilic. *A. castellanii* have the highest GRAVY value of -0.514 while *T. brucei brucei* have the lowest at -0.378.

DISCUSSION

The oxygen concentration in our atmosphere reflects the balance between efficient aerobic respiration and utilization of "clean" energy sources, versus damage from oxidation and misfolding of proteins and its co-factors such as iron and molybdenum (Ciechanover 2006). In the absence of compartmentalization, oxidized or misfolded proteins will promiscuously aggregate with undamaged molecules at an extensive rate, leading to the collapse of the entire proteome. Protein quality control, thus, is a major bottleneck in the evolution of energy-efficient metabolism, requiring robust but selective removal of damaged proteins coupled to elaborate synthetic machineries that will replace them. Therefore, in our opinion, the ubiquitin molecule, the ever-present tag of damaged proteins present only in the eukaryotes, represents a suitable molecule which can help address this long-standing question of eukaryotic evolution from the vantage point of our planet's oxygen history, not excluding the early ocean temperatures.

From our data, it is evident that the amino acid sequence of ubiquitin has not changed greatly across the different taxa considered and between the ancestral and modern-day ubiquitin molecules. Among the 76 amino acids, only nine residues have undergone changes, and most of the points of alteration are similar among the groups. The 'choice' of substitute amino acid is almost constant as well, e.g., Ser57 is replaced exclusively by Ala. This, however, does not imply that the nucleotide sequence of the proto-Ub is as conserved as its protein sequence. It has been noted that, even in an individual organism, the bases encoding the molecule of interest vary to some extent (Vrana & Wheeler 1996). Unexpectedly, the ubiquitin protein sequence of *Brugia malayi,* a parasitic nematode that causes elephantiasis, was found identical to the generated ancestral sequence. We know that this species is not the earliest eukaryote to contain the ubiquitin protein as protists and protozoans, which are considered more primitive, also possess ubiquitin-dependent protein degradation machineries. So, what does this tell us?

Without free oxygen, the earliest forms of life may have utilized elements such as iron, processing them to gain

Table 1. Physico-chemical properties of ancestral and extant ubiquitin[a].

Species	MW (g/mol)	pI	Instability Index	Aliphatic Index	GRAVY[b]
Proto-Ub	8554.8	6.56	36.06	100.00	-0.479
Brugia malayi	8554.8	6.56	36.06	100.00	-0.479
H. sapiens[c]	8564.8	6.56	36.06	100.00	-0.489
C. elegans[d]	8538.8	6.56	30.99	101.32	-0.445
S. purpuratus	8550.8	6.56	40.57	100.00	-0.491
A. neglectum	8584.8	6.56	36.06	98.68	-0.512
C. reinhardtii	8538.8	6.56	34.94	101.32	-0.445
C. congregates	8540.7	6.56	30.28	100.00	-0.479
L. major	8548.8	6.57	37.82	101.32	-0.457
A. castellanii	8595.8	6.56	38.06	100.00	-0.514
N. sylvestris[e]	8524.7	6.56	29.16	101.32	-0.445
S. cerevisiae	8556.7	6.56	30.28	98.68	-0.513
A. cliftonii	8540.7	6.56	29.16	98.68	-0.483
T. brucei	8506.8	6.57	31.63	103.95	-0.378
E. eurystomus	8581.8	6.55	32.41	100.00	-0.513

[a] The following properties are constant for all species considered: Total number of negatively charged residues = 11; Total number of positively charged residues = 11; Extinction coefficient (M^{-1} cm^{-1}) = 1490; Estimated half-life (hours): Mammalian reticulocytes in vitro 30, Yeast in vivo >20, *E.coli* in vivo >10.

[b] Grand Average of Hydropathy, the summation of hydropathy values of all the amino acids that constitute a protein divided by the number of residues in the sequence (Gasteiger et al. 2005)

[c] 13 other species with the same sequence as that of human have identical values for parameters considered

[d] *Caenorhabditis elegans* and *Geodia cydonium* have identical values for parameters considered

[e] all plants considered have identical values: *Zea mays*, *Oryza sativa* subsp. japonica, *Nicotiana sylvestris* and *Solanum lycopersicum*

only marginal amounts of energy (Falkowski & Isozaki 2008). The energy race among competing organisms equipped with the ability to exploit solar power, may have, over time, transformed our planet into a virtual toxic dump of oxygen. The impact of oxygen may have been dampened for a time by constant volcanic activity and the movement of land masses, which may have led to large-scale hydrogen sulfide release into surface waters. The surface waters then functioned as an "oxygen sink" while there was a concomitant decrease in methanogenic bacteria [$CO_2 + 2H_2O \rightarrow CH_4 + 2O_2 \rightarrow CO_2 + O_2 +$ 4H (↑space)], which expelled hydrogen to outer space (Catling et al. 2001; Falkowski & Isozaki 2008). These events may have allowed oxygen build-up to fluctuate, preventing the complete annihilation of early life forms and facilitating the evolution of eukaryotic life.

The emergence of ubiquitin being central to the management of toxic damaged proteins in cells is perhaps the molecular "redeemer" that arose when the existence of life was being challenged by protein denaturation from the combined oxidative and thermal stressors. Had the evolution of the ubiquitin-conjugation enzyme come later, mere protein-protein interaction, given extensive surface contact points of ubiquitin, may have been sufficient to change the molecular landscape of a protein, and influence interactions with other proteins, possibly diversifying protein functions. Moreover, free proto-Ub in transitional and early eukaryotes would also have been abundant since present-day cells contain >10 μM ubiquitin (Haas & Bright 1985).

Drawing from the discussion by Wostmann, Tannich & Bakker-Grunwald (1992) on the *Entamoeba* ubiquitin, although our proto-Ub largely conforms to the consensus of present-day organisms, it is deemed that in the ancient proteome milieu, proto-Ub may have offered an additional level of control over protein solubility in the cell by enhancing the solubility of poorly folded or damaged proteins as dictated by its physical properties. Further veering away from the traditional view of ubiquitin as a mere degradation tag, it should be considered that in simple eukaryotes, ubiquitin also functions in manipulating the endocytic pathways and cytoplasmic protein trafficking (Aguilar & Wendland 2003). At the early stage of life on earth, even minor adaptations may have conferred large evolutionary advantages.

Given that the sequence of proto-Ub did not appreciably change after the prokaryote-eukaryote transition, the identity of its prokaryotic predecessor is still

unanswered. Around the same time our manuscript was being reviewed, Hochstrasser (2009) published a provocative analysis on the origins of the ubiquitin system. As suggested by structures of the ubiquitin-related modifier (Urm1) protein in yeast that bears sequence similarity with two prokaryotic sulfur carrier proteins (ThiS and MoaD) (Rudolph et al. 2001, Wang et al. 2001), there is a convincing evolutionary link between ubiquitin and the biosynthesis pathway of sulfur carrier proteins in prokaryotes. Of note, Urm1 does not share significant sequence identity (<14%) with ubiquitin, however, this yeast protein forms a thioester at its C-terminal Gly-Gly with a novel E1-like protein, Uba4, and is conjugated to target proteins similar to ubiquitin (Rudolph et al. 2001, Wang et al. 2001). Given that the evolution of ubiquitin must have involved the process of concerted evolution (Sharp & Li 1987), antecedents of the ubiquitin-protein conjugation needed to evolve as well. The E1 ubiquitin-activating enzyme may have originated from MoeB, an *E. coli* protein required for the biosynthesis of the molybdenum cofactor (Moco) (McGrath, Jentsch & Varshavsky 1991).

Proteasomes, on the other hand, were already in existence prior to the advent of ubiquitination, as these complex proteolytic machines are found in archaea and in some bacteria (Darwin 2009). Notably, a newly-discovered prokaryotic ubiquitin-like protein, Pup, is able to target proteins for proteolysis by the *Mycobacterium tuberculosis* proteasome. Pup is intrinsically disorderd (Chen et al. 2009), sharing no structural nor sequential similarity with Urm1 and ubiquitin. However, it tags proteins for degradation by attaching to substrate lysines via isopeptide bonds in a manner reminiscent of ubiquitin (Festa et al 2009)! In the absence of ubiquitination or pupylation, the proteasome, to a limited extent, is still capable of recognizing non-modified substrates: an established case is that of the polyamine synthesizing mitochondrial enzyme ornithine decarboxylase (ODC) (Takeuchi et al., 2008). Its recent discovery prompted the idea that early organisms had evolved a sophisticated ancient tagging system for protein homeostasis.

It is astonishing to think how prokaryotes during the Great Oxidation Event were able to innovate the structural features of the components of an ancient sulfur transfer pathway and integrate this within the molecular system design of pupylation to increase the efficiency and selectivity of protein degradation, thereby creating the most remarkable protein modification systems of eukaryotes. While eukaryotes have radiated the basic ubiquitin-ligating enzyme into a superfamily (Hochstrasser 2009), why is it that proto-ubiquitin, after its dramatic evolutionary leap, did not anymore change, despite the fact that the Earth's atmosphere reached present-day oxygen and temperature levels only 0.6 and 0.5 Gy ago, respectively? Perhaps the rate of protein misfolding had already started to decrease when proto-Ub evolved, since inferred paleotemperatures based on the melting profile of the recently resurrected ancestral Elongation factor Tu/1A showed cooling of the oceans (Gaucher et al. 2008). The role of the cooperation of mitochondrial chaperones, heat shock proteins that leave the premises of the endosymbiont to wander the cell and assist protein folding and prevent protein denaturation (Deocaris et al. 2006), should also not be discounted. These chaperones are known to reside in the mitochondria and are then summoned to clean up the cellular mess — a feature that might have been useful in conjunction with the ubiquitin system during eukaryotic evolution.

CONCLUSION

Although the evolutionary origin of ubiquitin remains elusive, studying the molecular evolution of the other components of the ubiquitin-mediated protein degradation system, together with their interaction with classic proteins linked to oxidative stress and oxidized protein repair (e.g. thioredoxin, glutathione reductase, methionine sulfide reductase, etc.) and the molecular chaperones (Hsp60, Hsp70, Hsp90, etc.) will provide a systems-wide molecular appreciation of the historic struggles of our unicellular ancestors.

REFERENCES

ACQUISTI C, KLEFFE J, COLLINS S. 2007. Oxygen content of transmembrane proteins over macroevolutionary time scales. Nature 445(7123):47-52.

AGUILAR RC, WENDLAND B. 2003. Ubiquitin: not just for proteasomes anymore. Current Opinion in Cell Biology 15:184–190 .

ALTSCHUL S, GISH W, MILLER W, MYERS E, LIPMAN D. 1990. Basic Local Alignment Search Tool. Journal of Molecular Biology 215:403-410.

ARNOLD K, BORDOLI L, KOPP J, SCHWEDE T. 2006. The SWISS-MODEL Workspace: A web-based environment for protein structure homology modeling. Bioinformatics 22:195-201.

BERNER RA. 1999. Atmospheric oxygen over Phanerozoic time. Proceedings of the National Academy of Sciences USA 96:10955–7.

BIENKOWSKA JR, HARTMAN H, SMITH TF. 2003. A search method for homologs of small proteins. Ubiquitin-like proteins in prokaryotic cells? Protein Engineering 16(12):897-904.

BOECKMANN B, BAIROCH A, APWEILER R, BLATTER M-C, ESTREICHER A, GASTEIGER E, MARTIN MJ, MICHOUD K, O'DONOVAN C, PHAN I, PILBOUT S, SCHNEIDER M. 2003. The Swiss-Prot Protein Knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research 31:365-370.

CATLING DC, ZAHNLE KJ, MCKAY C. 2001. Biogenic methane, hydrogen escape, and the irreversible oxidation of early Earth. Science 293(5531):839-43.

CHEN X, SOLOMON WC, KANG Y, CERDA-MAIRA F, DARWIN KH, WALTERS KJ. 2009. Prokaryotic ubiquitin-like protein pup is intrinsically disordered. Journal of Molecular Biology 392:208–217.

CHENNA R, SUGAWARA H, KOIKE T, LOPEZ R, GIBSON TJ, HIGGINS DG, THOMPSON JD. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Research 31 (13):3497-500.

CIECHANOVER A. 2006. The ubiquitin proteolytic system: from an idea to the patient bed. Proceedings of the American Thoracic Society 3:21-31.

DARWIN KH. 2009. Prokaryotic ubiquitin-like protein (Pup), proteasomes and pathogenesis. Nature Reviews Microbiology 7:485-491.

DEOCARIS CC, KAUL SC, WADHWA R. 2006. On the brotherhood of the mitochondrial chaperones mortalin and heat shock protein 60. Cell Stress and Chaperones 11(2):116-28.

DEOCARIS CC, KAUL SC, WADHWA R. 2010. Mitochondrial heat shock proteins and the 2 billion-year old story of distress. Advances in Mitochondrial Medicine. Roberto Scatena, (ed). UK: Springer Publications [in press].

FALKOWSKI PG, ISOZAKI Y. 2008. Geology: The story of O_2. Science 322(5901):540-2.

FALKOWSKI PG, KATZ ME, KNOLL AH, QUIGG A, RAVEN JA, SCHOFIELD O, TAYLOR FJ. 2004. The evolution of modern eukaryotic phytoplankton. Science 305(5682):354-60.

FESTA RA, MCALLISTER F., PEARCE MJ, MINTSERIS J, BURNS KE, GYGI SP, DARWIN KH. 2009. Prokayrotic Ubiquitin-Like Protein (Pup) Proteome of *Mycobacterium tuberculosis*. PLoS ONE 5(1): e8589.

GASTEIGER E, HOOGLAND C, GATTIKER A, DUVAUD S, WILKINS MR, APPEL RD, BAIROCH A. 2005. Protein Identification and Analysis Tools on the ExPASy Server. In: The Proteomics Protocols Handbook. John M. Walker (ed). New York: Humana Press. p571-607.

GAUCHER EA. 2007. Ancestral sequence reconstruction as a tool to understand natural history and guide synthetic biology: Realizing (and extending) the vision of Zukerkandl and Pauling. In: Ancestral Sequence Reconstruction. David Liberles (ed). United Kingdom: Oxford University Press. p20-33.

GAUCHER EA, GOVINDARAJAN S, GANESH OK. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451(7179):704-7.

GAUCHER EA, THOMSON JM, BURGAN M, BENNER S. 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature 425:285-288.

GHEDIN E, WANG S, SPIRO D, CALER E, ZHAO Q, CRABTREE J, ALLEN JE, DELCHER AL, GUILIANO DB, MIRANDA-SAAVEDRA D, ANGIUOLI SV, CREASY T, AMEDEO P, HAAS B, EL-SAYED NM, WORTMAN JR, FELDBLYUM T, TALLON L, SCHATZ M, SHUMWAY M, KOO H, SALZBERG SL, SCHOBEL S, PERTEA M, POP M, WHITE O, BARTON GJ, CARLOW CKS, CRAWFORD MJ, DAUB J, DIMMIC MW, ESTES CF, FOSTER JM, GANATRA M, GREGORY WF, JOHNSON NM, JIN J, KOMUNIECKI R, KORF I, KUMAR S, LANEY S, LI B, LI W, LINDBLOM TH, LUSTIGMAN S, MA D, MAINA CV, MARTIN DMA, MCCARTER JP, MCREYNOLDS L, MITREVA M, NUTMAN TB, PARKINSON J, PEREGRÍN-ALVAREZ JM, POOLE C, REN Q, SAUNDERS L, SLUDER AE, SMITH K, STANKE M, UNNASCH TR, WARE J, WEI AD, WEIL G, WILLIAMS DJ, ZHANG Y, WILLIAMS SA, FRASER-LIGGETT C, SLATKO B, BLAXTER ML, SCOTT AL. 2007. Draft genome of the filarial nematode parasite *Brugia malayi*. Science 317(5845):1756-1760.

GREGORY R. 2008. Understanding evolutionary trees. Evolution: Education Outreach 1:121–137.

GUEX N, PEITSCH MC. 1997. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling. Electrophoresis 18: 2714-2723.

HAAS AL, BRIGHT PM. 1985. The immunochemical detection and quantitation of intracellular ubiquitin-protein conjugates. Journal of Biological Chemistry 260:12464–73.

HAN TM, RUNNEGAR B. 1992. Macroscopic eukaryotic algae from the 2.1-billion-year-old Negaunee iron-formation, Michigan. Science 257:232–235.

HOCHSTRASSER M. 2009 Origin and Function of Ubiquitin-like Protein Conjugation. Nature 458(7237): 422-429.

KNAUTH LP, LOWE DR. 2003. High Archean climatic temperature inferred from oxygen isotope geochemistry of cherts in the 3.5 Ga Swaziland Supergroup, South Africa. Geological Society of America Bulletin 115:566–580.

MACE R, HOLDEN CJ. 2005. A phylogenetic approach to cultural evolution. Trends in Ecology and Evolution 20:116-21.

MCGRATH JP, JENTSCH S, VARSHAVSKY A. 1991. *UBA1*: an essential yeast gene encoding ubiquitin-activating enzyme. EMBO Journal 10:227-236.

NAUJOKAT C, SARI T. 2007. Concise review: role and function of the ubiquitin-proteasome system in mammalian stem and progenitor cells. Stem Cells 25: 2408 –2418.

POOLE B, OHKUMA S, WARBURTON MJ. 1978. Some aspects of the intracellular breakdown of exogenous and endogenous proteins. In: Protein turnover and lysosome. Segal HL, Doyle DJ (eds). New York: Academic Press. p43-58.

NEI M, KUMAR S. 2000. Molecular Evolution and Phylogenetics. New York: Oxford University Press. 339pp.

RAMBAUT A, ROBERTSON DL, PYBUS OG, PEETERS M, HOLMES EC. 2001. Phylogeny and the origin of HIV-1. Nature 410:1047-8.

RASMUSSEN B, FLETCHER IR, BROCKS JJ, KILBURN MR. 2008. Reassessing the first appearance of eukaryotes and cyanobacteria. Nature 455(7216):1101-4.

RUDOLPH MJ, WUEBBENS MM, RAJAGOPALAN KV, SCHINDELIN H. 2001. Crystal structure of molybdopterin synthase and its evolutionary relationship to ubiquitin activation. Nature Structural Biology 8:42–46.

RZHETSKY A, NEI M. 1992. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. Journal of Molecular Evolution 35(4):367-75.

SAITOU N, NEI M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4(4):406-25.

SAYLE R, MILNER-WHITE EJ. 1995. RasMol: Biomolecular graphics for all. Trends in Biochemical Sciences 20(9):374.

SCHWEDE T, KOPP J, GUEX N, PEITSCH MC. 2003 SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Research 31: 3381-3385.

SHARP PM, LI W-H. 1987. Ubiquitin genes as a paradigm of concerted evolution of tandem repeats. Journal of Molecular Biology 25: 58-64.

TAKEUCHI J, CHEN H, HOYT MA, COFFINO P. 2008. Structural elements of the ubiquitin-independent proteasome degron of ornithine decarboxylase. Biochemical Journal 410(2):401-407.

TAMURA K, DUDLEY J, NEI M, KUMAR S. 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24:1596-1599.

VOGEL G. 1997. Phylogenetic analysis: Getting its day in court. Science 275:1559-60.

VRANA PB, WHEELER WC. 1996. Molecular evolution and phylogenetic utility of the polyubiquitin locus in mammals and higher vertebrates. Molecular Phylogenetic Evolution 6(2):259-69.

WANG C, XI J, BEGLEY TP, NICHOLSON LK. 2001. Solution structure of ThiS and implications for the evolutionary roots of ubiquitin. Nature Structural Biology 8(1):47-51.

WOSTMANN C, TANNICH E, BAKKER-GRUNWALD T. 1992. Ubiquitin of *Entamoeba histolytica* deviates in six amino acid residues from the consensus of all other known ubiquitins. FEBS Letters 308(1): 54-58.

YANG D, OYAIZU Y, OYAIZU H, OLSEN GJ, WOESE CR. 1985. Mitochondrial origins. Proceedings of the National Academy of Sciences USA 82(13):4443-7.

YANG Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24(8): 1586-1591.

ZWICKL P, VOGES D, BAUMEISTER W. 1999. The proteasome: a macromolecular assembly designed for controlled proteolysis. Philosophical Transactions of the Royal Society B: Biological Sciences 345:1501-1511.