letters to nature

Acknowledgements This research was supported by a NERC grant. We would like to thank Petroleum Development Oman and Oman Ministry of Oil and Gas for making the material available for study and for permission to publish.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for material should be addressed to C.H.W. (c.wellman@sheffield.ac.uk).

Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins

Eric A. Gaucher¹, J. Michael Thomson^{2*}, Michelle F. Burgan³ & Steven A. Benner^{1,2,3}

¹NASA Astrobiology Institute, ²Department of Anatomy and Cell Biology, College of Medicine, and ³Department of Chemistry, University of Florida, Gainesville, Florida 32611-7200, USA

* Present address: Department of Cell and Developmental Biology, University of North Carolina, Chapel Hill, North Carolina 27599, USA

Features of the physical environment surrounding an ancestral organism can be inferred by reconstructing sequences¹⁻⁹ of ancient proteins made by those organisms, resurrecting these proteins in the laboratory, and measuring their properties. Here, we resurrect candidate sequences for elongation factors of the Tu family (EF-Tu) found at ancient nodes in the bacterial evolutionary tree, and measure their activities as a function of temperature. The ancient EF-Tu proteins have temperature optima of 55–65 °C. This value seems to be robust with respect

to uncertainties in the ancestral reconstruction. This suggests that the ancient bacteria that hosted these particular genes were thermophiles, and neither hyperthermophiles nor mesophiles. This conclusion can be compared and contrasted with inferences drawn from an analysis of the lengths of branches in trees joining proteins from contemporary bacteria¹⁰, the distribution of thermophily in derived bacterial lineages¹¹, the inferred G+C content of ancient ribosomal RNA¹², and the geological record combined with assumptions concerning molecular clocks¹³. The study illustrates the use of experimental palaeobiochemistry and assumptions about deep phylogenetic relationships between bacteria to explore the character of ancient life.

This year marks the 40th anniversary of the observation by Pauling and Zuckerkandl that it should be possible to infer the sequences of ancient proteins by comparing the sequences of their descendants¹⁴. Some 25 yr were required, however, before their vision of resurrecting ancient proteins for study was first realized^{1,2,4}. The properties of resurrected ancestral proteins have been used to correlate molecular behaviour with changing geology, ecology and physiology in mammals¹⁵, analyse the evolution of substrate specificity in biomedically important proteases⁵, and identify *in vitro* behaviours of proteins involved in inflammation and vision that are important to changing physiological function^{8,9}.

So far, however, experimental palaeobiochemistry has carried experimental scientists back in time only approximately 240 million years⁸. This has left untouched many of the most intriguing questions about ancient life. One of these relates to the role of thermophily in the history of life on Earth. Various models for environments in the Precambrian have suggested that the Earth was cold and covered with snow. Other models, inspired by the discovery of modern microorganisms that live at high temperatures, suggest that early bacteria may have been thermophiles, or possibly extreme thermophiles. Arguments based on indirect evidence, such as the lengths of branches of various trees, the G+C content of reconstructed ancestral rRNA, the possible cold temperature of early Earth, and the distribution of thermophily in contemporary taxa, have generated contradictory inferences.



Figure 1 The two unrooted universal trees used to reconstruct ancestral bacterial sequences. Archaea and Eukarya serve to provide a node within the bacterial subtree from which ancient sequences can be inferred. Thermophilic lineages are highlighted in bold. Aquificaceae subfamily not shown. **a**, Maximum likelihood topology used to reconstruct

the stem elongation factors from bacteria (ML-stem), or most recent common ancestor of bacteria, and the ancestral sequence for mesophilic lineages only (ML-meso). **b**, Alternative topology used to reconstruct the stem elongation factors from bacteria (Alt-stem).

letters to nature

We reasoned that if we were able to reconstruct the sequences of ancestral proteins from bacteria that lived in the Precambrian, resurrect these proteins in the laboratory, and measure their thermal stabilities, we might be able to obtain direct evidence addressing the temperature(s) at which the particular ancestral bacteria lived. To be suited for the study, the protein would need to have temperaturedependent behaviour, with an optimum at physiological temperature. Furthermore, the rate of divergence of the protein must be slow enough, and the number of available derived and sibling protein sequences large enough, that the ancestral sequences can be reconstructed with minimal ambiguity. In addition, when the inevitable ambiguity is encountered, various sequences capturing that ambiguity must be sampled to see whether the interpretation, however made, is robust with respect to that ambiguity.

Elongation factor Tu (from Bacteria) and elongation factor 1A (from Archaea and Eukarya) are suitable proteins for such a study on all counts. EFs are G proteins that present charged aminoacyltransfer RNAs to the ribosome during translation. Because of their relatively slow rates of sequence divergence, most character states of ancient EF sequences can be robustly reconstructed for proteins that lived on the order of a billion years ago. This has made them frequently useful for phylogenetic analysis. Furthermore, the optimal thermal stabilities of EFs correlate with the optimal growth temperature of the host organism. Thus, EFs from mesophiles, thermophiles and hyperthermophiles-defined as organisms that grow at 20–40, 40–80 and >80 °C, respectively, and represented by species of Escherichia, Thermus and Thermotoga-have temperature optima in their respective ranges¹⁶⁻¹⁸. This is consistent with a previous study based on a large set of proteins in which a correlation coefficient of 0.91 was calculated between environmental temperatures of the host organisms and protein melting temperatures¹⁹.

To infer the sequences of EFs deep within the bacterial lineage, amino acid sequences of 50 EF-Tu proteins from various bacterial



Figure 2 Comparisons of reconstructed ancestral sequences. **a**, Distribution of the marginal posterior probabilities at sites for the three reconstructed ancestral nodes. The overall accuracies for sequence reconstructions of ML-stem, Alt-stem and ML-meso are predicted to be 88%, 88% and 92%, respectively²⁷. **b**, Examples of the per cent sequence identity between ancestral and modern proteins. T.m., *T. maritima*; T.a., *T. aquaticus*; E.c., *E. coli*; G.s., *Geobacillus stearothermophilus*.

lineages were collected, aligned and phylogenetically analysed. Because saturation at silent sites in the DNA sequence had occurred, amino acid sequences were used in the analysis. The differences in the rates of amino acid replacement at different sites in the sequence were captured using a gamma distribution²⁰. To support an analysis of the robustness of our interpretations with respect to plausible changes in our evolutionary models, two phylogenetic trees were used (see Methods). The first was constructed from EF-Tu sequences alone using a combination of phylogenetic tools (see Methods). The second was constructed from the literature, which contains various views of bacterial phylogeny (see, for example, ref. 11). These generally agree among themselves and with the EF-Tu tree. Where they differed, however, we extracted a tree that captured those differences (Fig. 1).

Candidate ancestral sequences were then reconstructed at nodes throughout the bacterial subtree. Marginal reconstructions, opposed to joint reconstructions, were calculated owing to our interest in comparing probabilities of multiple character states at a single interior node and selecting the character with the highest posterior probability²¹. The 'most probabilistic ancestral sequence' (MPAS) was then reconstructed by accepting at each site the amino acid with the highest posterior probability.

The MPASs for the two trees were found to be surprisingly similar to sequences from modern day *Aquifex*; only 4–6 amino acid replacements out of about 400 residues were inferred to have occurred from the most recent common ancestor of bacteria to the modern *Aquifex*. The placement of the branch leading to Aquificaceae at the base of the tree appeared, however, to be due



Figure 3 GDP-binding assay to test thermostability of ancestral and modern EF proteins. The amount of tritium-labelled GDP bound at 0 °C was subtracted from all other temperature values for a given protein. Shown is the relative amount of GDP bound compared with the amount bound at the optimal temperature for each protein. to long-branch attraction, an artefact that has been much discussed^{13,22}. To test this, the 23 sites that displayed no variation within the Archaea/Eukarya subfamily, the Aquificaceae subfamily and the subfamily containing all other bacteria, but not conserved between the subfamilies, were removed from the analysis. The resulting analysis no longer places Aquificaceae at the base of the bacterial lineage. This is consistent with the working of long-branch attraction. To eliminate bias due to this artefact, ancestral sequences were recalculated from a data set that excluded *Aquifex* sequences.

Figure 1 shows the two topologies used to reconstruct ancestral sequences at the node representing the hypothetical organism laying near the presumed stem of the bacterial tree. The number of archaeal and eukaryal sequences (3-20) did not affect the amino acid reconstructions at these nodes: ML-stem (maximum likelihood stem for elongation factors in bacteria) and Alt-stem (alternative stem for the elongation factors in bacteria). The ancestral sequence at the node representing the most recent common ancestor of only mesophilic bacterial lineages was also reconstructed, and named ML-meso (maximum likelihood mesophiles only). This node captures a feature of models that have concluded that the last common ancestor of certain bacteria was mesophilic²². In all, these reconstructed ancestral sequences did not appear to be influenced by long-branch attraction or non-homogeneous modes of molecular evolution, such as changes in the mutability of individual sites in different branches of the bacterial subtree²⁰. Figure 2 presents the range of posterior probabilities for the reconstructed sequences and shows the sequence identity relating the putative sequences and their descendants. ML-stem and Alt-stem are most similar to the sequences of EFs from Thermoanaerobacter tengcongensis (a thermophile) and Thermotoga *maritima* (a hyperthermophile), respectively, and differ from each other by 28 amino acids. ML-meso is most similar to the sequence of EF from *Neisseria meningitides* (a mesophile).

If we assume that similarity in sequence implies similarity in thermostability, we would predict that the stem bacterium was thermophilic or hyperthermophilic, and that the ancestral node constructed without considering thermophiles was mesophilic. To test these predictions based on this (unsubstantiated) assumption, genes encoding the ancestral sequences were synthesized, expressed in an *Escherichia coli* host and purified (see Methods). The thermostabilities of these ancestral EFs, and three representative EFs from contemporary organisms, were then assessed by measuring the ability of each to bind GDP across a range of temperatures.

Each resurrected protein behaved similarly (Fig. 3a). Both ML-stem and Alt-stem bound GDP with a temperature profile similar to that of the thermophilic EF from modern *Thermus aquaticus*, with optimal binding at about 65 °C. Although the sequence similarity was higher between Alt-stem and the modern hyperthermophilic *T. maritima*, the temperature profile of Alt-stem was not similar to that from *T. maritima*, which is maximally active up to at 85 °C (data not shown). The observation that the amino acid sequences of ML-stem and Alt-stem shared only 93% identity, but display the same thermostability profiles, suggests that inferences of this ancestral property are robust with respect to both varying topologies and ancestral character state predictions. This suggests, on the basis of these given evolutionary models, that the temperature in the palaeoenvironment of the ancient bacterium that hosted these reconstructed proteins was approximately 65 °C.

We then asked what inferences might be drawn from a resurrected EF whose sequence was reconstructed from the last common ancestral sequences of contemporary organisms that, for the most part, grow optimally at mesophilic temperatures. The temperature profile of the ancestral protein, which displayed a maximum at 55 °C, suggests that the ancestor of modern mesophiles lived at a higher temperature than its descendants (Fig. 3b). This result demonstrates that the behaviour of an ancestral protein need not be an average of the behaviours of its descendants, and suggests that

phylogenetic-based ancestral sequence reconstructions (per stirpes) should be preferred over consensus sequence reconstructions (per capita) and applied generally²³. The observation is important because it underscores the fact, well known in protein chemistry, that physical behaviour in a protein cannot be reliably predicted by a model that assumes amino acids at each site contribute to protein behaviour independently of the residues at all other sites. This, in turn, implies that an experiment in palaeobiochemistry can yield information beyond that yielded by analysis of descendant proteins alone.

Clearly, as more bacterial genomes are completed, we will need to exploit the resulting opportunity to extend these inferences to other nodes in the universal tree. Also, although our interpretation is robust with respect to the sampling of sequences from the extremes of the evolutionary models presented here, other plausible intermediate sequences will need to be resurrected and tested to establish this robustness more broadly. This notwithstanding, the results presented here show the value of experimental palaeobiochemical reconstructions to explore features in the lifestyles of ancient life forms, and will be particularly valuable when the relation of these ancient life forms is established relative to the root of the universal tree by other methods.

Methods

Computational analyses

Completely sequenced genomes were used to determine the protein families that might support reconstructions in the bacterial tree most successfully. The MasterCatalog (EraGen Biosciences)¹⁵ was used as a starting point, as it presents an evolutionary model for all protein families in a recent version of GenBank. Gene trees were surveyed to identify the family that had the most sequences from the most diverse taxa with the least overall protein sequence divergence, using the PAM (point accepted mutations per 100 amino acids) distance metric. The family that represented the most potential for ancestral reconstruction was elongation factor Tu (Bacteria, EF-Tu) and 1A (Archaea/Eukarya, EF1A). This family was highly conserved, contained members from all of the complete genomes, and GenBank and Swiss-Prot databases contained diverse EF representatives. Sequences were aligned using ClustalW, and minor corrections were performed by hand. The list of organisms, GI and/or Swiss-Prot numbers, and multiple sequence alignment are available on request.

Two phylogenetic trees were used in the computation of ancestral amino acid character states for bacterial EF-Tu (about 400 residues). The maximum likelihood (ML) tree was generated as follows: a distance matrix was generated from the EF sequences using a gamma distribution to capture site-specific rate heterogeneity using the MEGA package²⁴. This distance matrix was then used to build trees inter-relating the sequences from individual lineages of bacteria using the minimum evolution criterion in the PAUP* package²⁵. The subtrees within the nine individual bacterial lineages were thus constrained, and the relationships between the lineages were then inferred using the maximum likelihood algorithm implemented within the MOLPHY package²⁶. The second tree was constructed by identifying in the literature plausible alternative bacterial phylogenies. Many of these combined analyses of multiple protein, DNA and rRNA sequences. Where these trees differed from the first tree, the differences were captured in an alternative tree (Alt).

The computation of ancestral amino acid states was performed using an empirical bayesian statistical framework that incorporated the gamma distribution (PAML)²⁷. For both the ML and alternative trees, ancestral character states were predicted using the Jones, Taylor and Thornton replacement matrix²⁸ (JTT) (See Supplementary Information). Using the Dayhoff²⁹ and WAG³⁰ (Whelan And Goldman) matrices had little effect on the predicted ancestral states.

Synthetic gene construction and protein expression

The ancestral genes were synthesized piecemeal by polymerase chain reaction using complementary 50-nucleotide oligonucleotides (MWG Biotech) with 15–20 base-pair overlap. Ancestral and extant (*E. coli, T. aquaticus* and *T. maritima*) EF genes were cloned in Topo-TA (Invitrogen). Errors resulting from primer synthesis or PCR were fixed using standard site-directed mutagenesis. All genes were sequenced with two times coverage. The genes were subsequently cloned and expressed in the TYB11 vector (IMPACT System; intein-mediated purification with an affinity chitin-binding tag), and purified according to the manufacturer (New England Biolabs). The proteins were eluted from the chitin affinity column in a buffer comprising 20 mM Tris-HCl (pH 8.5), 500 mM NaCl, 10 mM MgCl₂, 5 μ M GDP and 1 μ M phenylmethylsulphonyl fluoride and stored at -20 °C. The samples were filtered and concentrated using Centricon YM-30 (Amicon). SDS–PAGE verified the isolation of a single band of appropriate size: about 44 kDa.

GDP-binding assay

The thermostability of ancestral and extant proteins was determined based on an ability to bind nucleotide across a range of temperatures^{16–18}. Aliquots (40 μ l) of a solution

letters to nature

containing 1 μM EF protein in 20 mM Tris-HCl (pH 8.5), 500 mM NaCl, 10 mM MgCl₂ and 2.5 μM [³H]GDP (specific activity approximately 11.5 Ci mmol⁻¹) (Amersham Pharmacia Biotech) were assayed at different temperatures, first at 10°C intervals between 0 and 100°C and then \pm 5 °C on either side of the temperature optimum. For the ML-meso putative ancestor, binding at 5-°C intervals between 40 and 70 °C was also determined. The amount of [³H]GDP bound to EF was determined as previously described¹⁶.

Received 7 April; accepted 30 July 2003; doi:10.1038/nature01977.

- Malcolm, B. A., Wilson, K. P., Matthews, B. W., Kirsch, J. F. & Wilson, A. C. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature* 345, 86–89 (1990).
- Stackhouse, J., Presnell, S. R., McGeehan, G. M., Nambiar, K. P. & Benner, S. A. The ribonuclease from an extinct bovid. *FEBS Lett.* 262, 104–106 (1990).
- Adey, N. B., Tollefsbol, T. O., Sparks, A. B., Edgell, M. H. & Hutchison, C. A. Molecular resurrection of an extinct ancestral promoter for mouse L1. *Proc. Natl Acad. Sci. USA* 91, 1569–1573 (1994).
- Jermann, T. M., Opitz, J. G., Stackhouse, J. & Benner, S. A. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* 374, 57–59 (1995).
- Chandrasekharan, U. M., Sanker, S., Glynias, M. J., Karnik, S. S. & Husain, A. Angiotensin II forming activity in a reconstructed ancestral chymase. *Science* 271, 502–505 (1996).
- Golding, G. B. & Dean, A. M. The structural basis of molecular adaptation. *Mol. Biol. Evol.* 15, 355–369 (1998).
- Miyazaki, J. et al. Ancestral residues stabilizing 3-isopropylmalate dehydrogenase of an extreme thermophile: Experimental evidence supporting the thermophilic common ancestor hypothesis. J. Biochem. 129, 777–782 (2001).
- Chang, B. S., Jonsson, K., Kazmi, M. A., Donoghue, M. J. & Sakmar, T. P. Recreating a functional ancestral archosaur visual pigment. *Mol. Biol. Evol.* 19, 1483–1489 (2002).
- Zhang, J. Z. & Rosenberg, H. F. Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. *Proc. Natl Acad. Sci. USA* 99, 5486–5491 (2002).
- 10. Woese, C. R. Bacterial evolution. Microbiol. Rev. 51, 221-271 (1987).
- Hugenholtz, P., Goebel, B. M. & Pace, N. R. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J. Bacteriol. 180, 4765–4774 (1998).
- Galtier, N., Tourasse, N. & Gouy, M. A nonhyperthermophilic common ancestor to extant life forms. Science 283, 220–221 (1999).
- Cavalier-Smith, T. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int. J. Syst. Evol. Microbiol.* 52, 7–76 (2002).
- Pauling, L. & Zuckerkandl, E. Chemical paleogenetics: molecular "restoration studies" of extinct forms of life. Acta Chem. Scand. A 17, S9–S16 (1963).
- Benner, S. A., Caraco, M. D., Thomson, J. M. & Gaucher, E. A. Planetary biology-paleontological, geological, and molecular histories of life. *Science* 296, 864–868 (2002).
- Arai, K.-I., Kawakita, M. & Kaziro, Y. Studies on polypeptide elongation factors from *Escherichia coli*. *I. Biol. Chem.* 37, 7029–7037 (1972).
- Nock, S. et al. Properties of isolated domains of the elongation factor Tu from Thermus thermophilus HB8. Eur. J. Biochem. 234, 132–139 (1995).
- Sanangelantoni, A. M., Cammarano, R. & Tiboni, O. Manipulation of the tuf gene provides clues to the localization of sequence element(s) involved in the thermal stability of *Thermotoga maritima* elongation factor Tu. *Microbiol.* **142**, 2525–2532 (1996).
- Gromiha, M. M., Oobatake, M. & Sarai, A. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophys. Chem.* 82, 51–67 (1999).
- Gaucher, E. A., Miyamoto, M. M. & Benner, S. A. Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. *Proc. Natl Acad. Sci. USA* 98, 548–552 (2001).
- Yang, Z. H., Kumar, S. & Nei, M. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141, 1641–1650 (1995).
- Brochier, C. & Philippe, H. Phylogeny: a non-hyperthermophilic ancestor for bacteria. Nature 417, 244 (2002).
- Gaschen, B. et al. Diversity considerations in HIV-1 vaccine selection. Science 296, 2354–2360 (2002).
 Kumar, S., Tamura, K. & Nei, M. MEGA–Molecular evolutionary genetics analysis software for
- microcomputers. Comput. Appl. Biosci. 10, 189–191 (1994). 25. Swofford, D. L. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4.10
- (Sinauer Associates, Sunderland, Massachusetts, 2002).
- Adachi, J. & Hasegawa, M. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. *Comput. Sci. Monogr.* 28, 1–150 (1996).
- Yang, Z. PAML: a program package for phylogenetic analyses by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556 (1999).
- Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequence. *Comput. Appl. Biosci.* 8, 275–282 (1992).
- Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. Atlas of Protein Sequence and Structure 345–358 (National Biomedical Research Foundation, Washington DC, 1978).
- Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* 18, 691–699 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Miyamoto, J. Aris, A. Falcon, S. Sassi, C. West and Z. Yang for discussions and assistance with our research. We also thank C. Knudson and A. M. Sanangelantoni for providing EF clones. Funding is provided by the National Research Council and NASA's Astrobiology Institute (E.A.G.), and the NIH and NASA (S.A.B.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.A.G. (gaucher@ufl.edu). Inferred ancestral sequences have been deposited in GenBank under accession numbers AY305395–AY305397

Patterns of predation in a diverse predator-prey system

A. R. E. Sinclair¹, Simon Mduma^{1,2} & Justin S. Brashares¹

 ¹Centre for Biodiversity Research, 6270 University Boulevard, University of British Columbia, Vancouver, V6T 1Z4, Canada
 ²Tanzania Wildlife Research Institute, Box 661, Arusha, Tanzania

lanzania wilalije Research Institute, Box 661, Arusha, lanzani

There are many cases where animal populations are affected by predators and resources in terrestrial ecosystems¹⁻³, but the factors that determine when one or the other predominates remain poorly understood⁴⁻⁵. Here we show, using 40 years of data from the highly diverse mammal community of the Serengeti ecosystem, East Africa, that the primary cause of mortality for adults of a particular species is determined by two factorsthe species diversity of both the predators and prey and the body size of that prey species relative to other prey and predators. Small ungulates in Serengeti are exposed to more predators, owing to opportunistic predation, than are larger ungulates; they also suffer greater predation rates, and experience strong predation pressure. A threshold occurs at prey body sizes of \sim 150 kg, above which ungulate species have few natural predators and exhibit food limitation. Thus, biodiversity allows both predation (top-down) and resource limitation (bottom-up) to act simultaneously to affect herbivore populations. This result may apply generally in systems where there is a diversity of predators and prey.

The influence of predation and resource availability on population dynamics has long been a focus of ecological research⁶. Yet, we know little about how these top-down and bottom-up forces work together to structure diverse ecosystems^{4–6,7}. This is particularly true for mammal communities, where effects of predation and resource limitation are usually investigated on single species or in simple systems with low species diversity^{8,9}. We examined patterns of predation in the diverse Serengeti ecosystem using data on ungulate populations before, during and after a period of predator removal.

The Serengeti ecosystem $(34-36^{\circ}\text{E}, 1-4^{\circ}\text{S})$ in Tanzania and Kenya, East Africa, is composed of open grassland and savannah. It supports 28 species of ungulates and 10 species of large carnivores that prey on them¹⁰. In any one habitat, there can be seven co-



Figure 1 The range of weights of mammal prey consumed by carnivores of different sizes in the Serengeti ecosystem. There is a large overlap in diet at small prey sizes. Data are from our unpublished observations and published sources^{17,26}.