LETTERS

A formal test of the theory of universal common ancestry

Douglas L. Theobald[1]

Universal common ancestry (UCA) is a central pillar of modern evolutionary theory[1]. As first suggested by Darwin[2], the theory of UCA posits that all extant terrestrial organisms share a common genetic heritage, each being the genealogical descendant of a single species from the distant past[3–6]. The classic evidence for UCA, although massive, is largely restricted to 'local' common ancestry—for example, of specific phyla rather than the entirety of life—and has yet to fully integrate the recent advances from modern phylogenetics and probability theory. Although UCA is widely assumed, it has rarely been subjected to formal quantitative testing[7–10], and this has led to critical commentary emphasizing the intrinsic technical difficulties in empirically evaluating a theory of such broad scope[1,5,8,9,11–15]. Furthermore, several researchers have proposed that early life was characterized by rampant horizontal gene transfer, leading some to question the monophyly of life[11,14,15]. Here I provide the first, to my knowledge, formal, fundamental test of UCA, without assuming that sequence similarity implies genetic kinship. I test UCA by applying model selection theory[5,16,17] to molecular phylogenies, focusing on a set of ubiquitously conserved proteins that are proposed to be orthologous. Among a wide range of biological models involving the independent ancestry of major taxonomic groups, the model selection tests are found to overwhelmingly support UCA irrespective of the presence of horizontal gene transfer and symbiotic fusion events. These results provide powerful statistical evidence corroborating the monophyly of all known life.

In the conclusion of *On the Origin of Species*, Darwin proposed that ''all the organic beings which have ever lived on this earth have descended from some one primordial form''[2]. This theory of UCA—the proposition that all extant life is genetically related—is perhaps the most fundamental premise of modern evolutionary theory, providing a unifying foundation for all life sciences. UCA is now supported by a wealth of evidence from many independent sources[18], including: (1) the agreement between phylogeny and biogeography; (2) the correspondence between phylogeny and the palaeontological record; (3) the existence of numerous predicted transitional fossils; (4) the hierarchical classification of morphological characteristics; (5) the marked similarities of biological structures with different functions (that is, homologies); and (6) the congruence of morphological and molecular phylogenies[9,10]. Although the consilience of these classic arguments provides strong evidence for the common ancestry of higher taxa such as the chordates or metazoans, none expressly address questions such as whether bacteria, yeast and humans are all genetically related. However, the 'universal' in universal common ancestry is primarily supported by two further lines of evidence: various key commonalities at the molecular level[6] (including fundamental biological polymers, nucleic acid genetic material, L-amino acids, and core metabolism) and the near universality of the genetic code[4,7]. Notably, these two traditional arguments for UCA are largely qualitative, and typical presentations of the evidence do not assess quantitative measures of support for competing hypotheses, such as the probability of evolution from multiple, independent ancestors.

The inference from biological similarities to evolutionary homology is a feature shared by several of the lines of evidence for common ancestry. For instance, it is widely assumed that high sequence resemblance, often gauged by an *E* value from a BLAST search, indicates genetic kinship[19]. However, a small *E* value directly demonstrates only that two biological sequences are more similar than would be expected by chance[20]. A Karlin–Altschul *E* value is a Fisherian null-hypothesis significance test in which the null hypothesis is that two random sequences have been aligned[20]. Therefore, an *E* value in principle cannot provide evidence for or against the hypothesis that two sequences share a common ancestor. (In fact, an *E* value cannot even provide evidence for the random null hypothesis.[21]) Sequence similarity is an empirical observation, whereas the conclusion of homology is a hypothesis proposed to explain the similarity[22]. Statistically significant sequence similarity can arise from factors other than common ancestry, such as convergent evolution due to selection, structural constraints on sequence identity, mutation bias, chance, or artefact manufacture[19]. For these reasons, a sceptic who rejects the common ancestry of all life might nevertheless accept that universally conserved proteins have similar sequences and are 'homologous' in the original pre-Darwinian sense of the term (homology here being similarity of structure due to ''fidelity to archetype'')[23]. Consequently, it would be advantageous to have a method that is able to objectively quantify the support from sequence data for common-ancestry versus competing multiple-ancestry hypotheses.

Here I report tests of the theory of UCA using model selection theory, without assuming that sequence similarity indicates a genealogical relationship. By accounting for the trade-off between data prediction and simplicity, model selection theory provides methods for identifying the candidate hypothesis that is closest to reality[16,17]. When choosing among several competing scientific models, two opposing factors must be taken into account: the goodness of fit and parsimony. The fit of a model to data can be improved arbitrarily by increasing the number of free parameters. On the other hand, simple hypotheses (those with as few ad hoc parameters as possible) are preferred. Model selection methods weigh these two factors statistically to find the hypothesis that is both the most accurate and the most precise. Because model selection tests directly quantify the evidence for and against competing models, these tests overcome many of the well-known logical problems with Fisherian null-hypothesis significance tests (such as BLAST-style *E* values)[16,21]. To quantify the evidence supporting the various ancestry hypotheses, I applied three of the most widely used model selection criteria from all major statistical schools: the log likelihood ratio (LLR), the Akaike information criterion (AIC) and the log Bayes factor (LBF)[16,17].

Using these model selection criteria, I specifically asked whether the three domains of life (Eukarya, Bacteria and Archaea) are best

[1]Department of Biochemistry, Brandeis University, Waltham, Massachusetts 01778, USA.

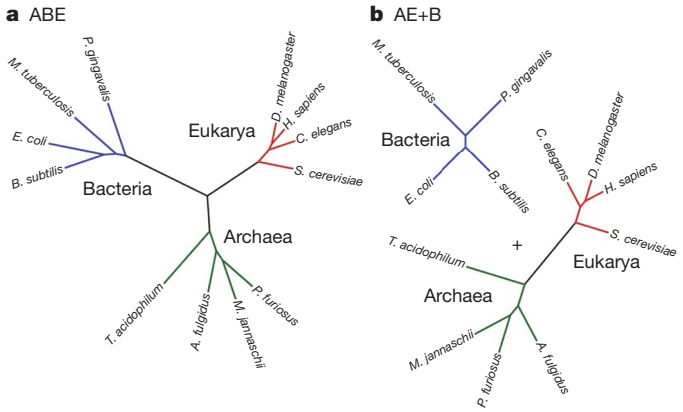

Figure 1 | Selected class I evolutionary hypotheses, excluding HGT. a, The model ABE, representing UCA of all taxa in the three domains of life. **b**, A competing multiple-ancestry model, AE+B, representing common ancestry of Archaea and Eukarya, but an independent ancestry for Bacteria. Trees shown are actual maximum likelihood estimates, with branch lengths proportional to the number of sequence substitutions.

described by a unified, common genetic relationship (that is, UCA) or by multiple groups of genetically unrelated taxa that arose independently and in parallel. As one example, a simplified model was considered for the hypothesis that Archaea and Eukarya share a common ancestor but do not share a common ancestor with Bacteria. This model (indicated by 'AE+B' in Fig. 1 and Table 1) comprises two independent trees—one containing Archaea and Eukarya and another containing only Bacteria. In these models the primary assumptions are: (1) that sequences change over time by a gradual, time-reversible Markovian process of residue substitution, described by a 20×20 instantaneous rate matrix defined by certain amino acid equilibrium frequencies and a symmetric matrix of amino acid exchangeabilities; (2) that new genetically related genes are generated by duplication during bifurcating speciation or gene duplication events; and (3) that residue substitutions are uncorrelated along different lineages and at different sites. The model selection tests evaluate how well these assumptions explain the given data set when various subsets of taxa and proteins are postulated to share ancestry, without any recourse to measures of sequence similarity.

The theory of UCA allows for the possibility of multiple independent origins of life[1–6]. If life began multiple times, UCA requires a 'bottleneck' in evolution in which descendants of only one of the independent origins have survived exclusively until the present (and the rest have become extinct), or, multiple populations with independent, separate origins convergently gained the ability to exchange essential genetic material (in effect, to become one species). All of the models examined here are compatible with multiple origins in both the above schemes, and therefore the tests reported here are designed to discriminate specifically between UCA and multiple ancestry, rather than between single and multiple origins of life. Furthermore, UCA does not demand that the last universal common ancestor was a single organism[24,25], in accord with the traditional evolutionary view that common ancestors of species are groups, not individuals[26]. Rather, the last universal common ancestor may have comprised a population of organisms with different genotypes that lived in different places at different times[25].

Table 1 | Class I hypotheses of single versus multiple ancestries

Hypothesis	$-\Delta K$	LLR	ΔAIC	LBF	ML evolutionary model
ABE	0	0	0	0	R-IGF
AE+B	17	6,569	6,586	6,889	(AE) R-IGF; (B) R-GF
AB+E	17	7,805	7,822	8,031	(AB) W-IGF; (E) R-GF
BE+A	18	8,192	8,210	8,488	(BE) R-IGF; (A) W-IGF
A+B+E	34	13,350	13,384	13,865	(E) R-GF; (B) R-GF; (A) W-IGF
ABE_{-M}+M	16	12,104	12,120	12,186	(ABE_{-M}) W-IF; (M) R-GF
ABE_{-H}+H	59	14,040	14,057	14,001	(ABE_{-H}) R-IGF; (H) empirical

Shown are the model section scores for class I hypotheses of single ancestry versus multiple ancestries, excluding HGT events. A, Archaea; B, Bacteria; E, Eukarya; H, Homo sapiens; M, Metazoa; ABE_{-M}, ABE without Metazoa; ABE_{-H}, ABE without H. sapiens. AE+B denotes a hypothesis of two independent ancestries, one tree for A and E together, and another separate tree for B. *K* denotes the total number of parameters in the model. All criteria are given as differences from ABE, so that larger values indicate less support for that model relative to ABE. LLR and ΔAIC scores correspond to the maximum likelihood (ML) estimates. For the ML evolutionary model, the first letter refers to the rate matrix: R, RtREV; W, WAG. The following letters denote models with additional parameters: I, invariant positions; G, gamma rate variation; F, empirical amino acid frequencies. The raw log likelihood for ABE is −126,299, and the marginal log likelihood is −126,713.

The data set consists of a subset of the protein alignment data from ref. 27, containing 23 universally conserved proteins for 12 taxa from all three domains of life, including nine proteins thought to have been horizontally transferred early in evolution[27]. The conserved proteins in this data set were identified based on significant sequence similarity using BLAST searches, and they have consequently been postulated to be orthologues. The first class of models I considered (presented in Table 1 and Fig. 1) constrains all the universally conserved proteins in a given set of taxa to evolve by the same tree, and hence these models do not account for possible horizontal gene transfer (HGT) or symbiotic fusion events during the evolution of the three domains of life. Hereafter I refer to this set of models as 'class I'. The class I model ABE, representing universal common ancestry of all taxa in the three domains of life and shown in Fig. 1a, can be considered to represent the classic three-domain 'tree of life' model of evolution[28].

Among the class I models, all criteria select the UCA tree by an extremely large margin (score differences ranging from 6,569 to 14,057), even though nearly half of the proteins in the analysis probably have evolutionary histories complicated by HGT. For all model selection criteria, by statistical convention a score difference of 5 or greater is viewed as very strong empirical evidence for the hypothesis with the better score (in this work higher scores are better)[16,17]. All scores shown are also highly statistically significant (the estimated variance for each score is approximately 2–3). According to a standard objective Bayesian interpretation of the model selection criteria, the scores are the log odds of the hypotheses[16,17]. Therefore, UCA is at least $10^{2,860}$ times more probable than the closest competing hypothesis. Notably, UCA is the most accurate and the most parsimonious hypothesis. Compared to the multiple-ancestry hypotheses, UCA provides a much better fit to the data (as seen from its higher likelihood), and it is also the least complex (as judged by the number of parameters).

The extraordinary strength of these results in the face of suspected HGT events suggests that the preference for the UCA model is robust to the extent of HGT. To test this possibility, the analysis was expanded to include models that allow each protein to have a distinct, independent evolutionary history. I refer to this set of models, which rejects a single tree metaphor for genealogically related taxa, as 'class II'. Representative class II models are shown in Fig. 2. Within each set of genealogically related taxa, each of the 23 universally conserved proteins is allowed to evolve on its own separate phylogeny, in which both branch lengths and tree topology are free parameters. For example, the multiple-ancestry model $[AE+B]^{II}$ comprises two clusters of protein trees, one cluster (AE) in which Archaea and Eukarya share a common ancestor but are genetically unrelated to another cluster (B) consisting only of Bacteria. Class II models are highly reticulate, phylogenetic networks that can represent very complex evolutionary mechanisms, including unrestricted HGT, symbiotic fusion events and independent ancestry of various taxa. Overall, the model selection tests show that the class II models are greatly preferred to the class I models. For instance, the class II UCA hypothesis ($[ABE]^{II}$) versus the class I UCA hypothesis (ABE) gives a highly significant LLR of 3,557, a ΔAIC of 2,633 and an LBF of 2,875. The optimal class II models represent an upper limit to the degree of HGT, as many of the apparent reticulations are probably due to incomplete lineage sorting, hidden paralogy, recombination, or inaccuracies in the evolutionary models. Nonetheless, as with the class I non-HGT hypotheses, all model selection criteria unequivocally support a single common genetic ancestry for all taxa. Also similar to the class I models, the class II UCA model has the greatest explanatory power and is the most parsimonious.

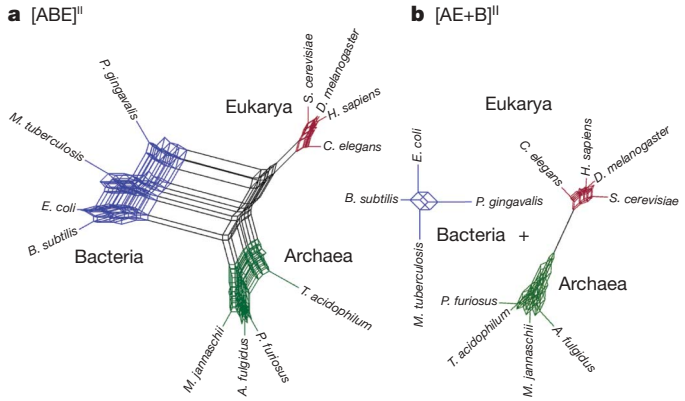

Figure 2 | Selected class II evolutionary hypotheses, including HGT. a, The reticulated model $[ABE]^{II}$, representing UCA. **b**, A competing network model of multiple ancestry, $[AE+B]^{II}$, representing common ancestry of Archaea and Eukarya, but a separate ancestry for Bacteria. Models are shown as phylogenetic networks (reticulate trees). The phylogenetic networks are derived from the maximum likelihood estimates of the 23 individual protein phylogenies using the evolutionary model parameters shown for ABE and AE+B in Table 1.

Several hypotheses have been proposed to explain the origin of eukaryotes and the early evolution of life by endosymbiotic fusion of an early archaeon and bacterium[29]. A key commonality of these hypotheses is the rejection of a single, bifurcating tree as a proper model for the ancestry of Eukarya. For instance, in these biological hypotheses certain eukaryotic genes are derived from Archaea whereas others are derived from Bacteria. The class II models freely allow eukaryotic genes to be either archaeal-derived or bacterial-derived, as the data dictate, and hence class II hypotheses can model several endosymbiotic 'rings' and HGT events. Because specific endosymbiotic fusion schemes can be represented by constrained versions of the unrestricted class II models, the endosymbiotic fusion hypotheses are nested within the class II hypotheses shown in Table 2. For nested hypotheses, the constrained versions necessarily have equal or lower likelihoods than the unconstrained versions. As a result, strict bounds can be placed on the LLR and ΔAIC scores for the constrained class II network models that represent specific endosymbiotic fusion or HGT hypotheses (see Methods and Supplementary Information). In all cases, these bounds show that multiple-ancestry versions of the constrained class II models are overwhelmingly rejected by the tests (model selection scores of several thousands), indicating that common ancestry is also preferred for all specific HGT and endosymbiotic fusion models. In terms of a fusion hypothesis for the origin of Eukarya, the data conclusively support a UCA model in which Eukarya share an ancestor with Bacteria and another independently with Archaea, and in which Bacteria and Archaea are also genetically related independently of Eukarya (see Table 3).

The proteins in this data set were postulated to be orthologous on the basis of significant sequence similarity[27]. Because the proteins are universally conserved, all of the taxa have their own specific versions of each of the proteins. It would be of interest to know how the tests respond to the inclusion of proteins that are not universally conserved, as omitting independently evolved proteins could perhaps bias the results towards common ancestry. Nevertheless, the inclusion of bona fide independently evolved genes has no effect on the likelihoods of the winning class II models, except in certain cases to strengthen the conclusion of common ancestry (for a formal proof, see the Supplementary Information). Many proteins probably do exist that have independent origins. For instance, in the Metazoa certain protein domains have probably evolved *de novo* that are not found in either Bacteria or Archaea[30]. However, the independent evolution of unique Metazoan proteins, by itself, is not evidence for or against UCA. The probability that the Metazoa would evolve a new protein domain is the same whether or not the Metazoa are related to Bacteria and Archaea. Therefore, omitting proteins with independent origins from the data set does not affect support for the UCA hypothesis versus multiple-ancestry hypotheses. In fact, including independently evolved proteins is expected to increase support for common ancestry for the subsets of taxa that share them (in this example, to increase support for common ancestry of the Metazoa).

Table 2 | Class II hypotheses of single versus multiple ancestries

Hypothesis	$-\Delta K$	LLR	ΔAIC	LBF
$[ABE]^{II}$	0	0	0	0
$[AE+B]^{II}$	391	7,642	8,033	8,124
$[AB+E]^{II}$	391	8,473	8,864	8,864
$[BE+A]^{II}$	414	8,829	9,243	9,333
$[A+B+E]^{II}$	782	14,481	15,263	15,369
$[ABE_{-M}+M]^{II}$	391	12,061	12,452	12,512
$[ABE_{-H}+H]^{II}$	391	14,141	14,532	14,126

Shown are model selection scores for class II hypotheses of single ancestry versus multiple ancestries, allowing for unlimited HGT and/or endosymbiotic fusion events. Abbreviations are as in the Table 1 legend. All criteria are listed as differences from $[ABE]^{II}$. All scores shown are highly statistically significant (the estimated variance for each score is approximately 3–6). The raw log likelihood for $[ABE]^{II}$ is −122,742, and the marginal log likelihood is −123,838.

Table 3 | Class I and class II hypotheses for selected subsets

Hypotheses	$-\Delta K$	LLR	ΔAIC	LBF
AB versus A+B	17	5,545	5,562	5,837
BE versus B+E	16	5,157	5,173	5,380
AE versus A+E	17	6,782	6,899	6,979
$[AB]^{II}$ versus $[A+B]^{II}$	391	6,008	6,399	6,505
$[BE]^{II}$ versus $[B+E]^{II}$	368	5,652	6,020	6,036
$[AE]^{II}$ versus $[A+E]^{II}$	391	6,839	7,230	7,245

Shown are model selection scores for class I and II hypotheses for selected subsets of the taxa. Single ancestry hypotheses are listed left, multiple-ancestry hypotheses right. Terms are as in Table 1.

As is common in phylogenetic practice, most gaps and poorly aligned regions were removed from the original data set used in this analysis[27], leaving only those sites that were thought to be homologous with high confidence. To explore the effect of these omitted sites, the model selection tests were performed on a similar data set, with the same proteins and species, in which all gaps were kept in the final alignment (see Supplementary Methods and Supplementary Tables 5–8). The inclusion of these gapped and poorly aligned regions in the analyses greatly increases the support for UCA in all cases (for instance, with the ABE versus AE+B test, the class I ΔAIC is 10,323 and the class II ΔAIC is 11,072).

What property of the sequence data supports common ancestry so decisively? When two related taxa are separated into two trees, the strong correlations that exist between the sequences are no longer modelled, which results in a large decrease in the likelihood. Consequently, when comparing a common-ancestry model to a multiple-ancestry model, the large test scores are a direct measure of the increase in our ability to accurately predict the sequence of a genealogically related protein relative to an unrelated protein. The sequence correlations between a given clade of taxa and the rest of the tree would be eliminated if the columns in the sequence alignment for that clade were randomly shuffled. In such a case, these model-based selection tests should prefer the multiple-ancestry model. In fact, in actual tests with randomly shuffled data, the optimal estimate of the unified tree (for both maximum likelihood and Bayesian analyses) contains an extremely large internal branch separating the shuffled taxa from the rest. In all cases tried, with a wide variety of evolutionary models (from the simplest to the most parameter rich), the multiple-ancestry models for shuffled data sets are preferred by a large margin over common ancestry models (LLR on the order of a thousand), even with the large internal branches. Hence, the large test scores in favour of UCA models reflect the immense power of a tree structure, coupled with a gradual Markovian mechanism of residue substitution, to accurately and precisely explain the particular patterns of sequence correlations found among genealogically related biological macromolecules.

METHODS SUMMARY

All analyses were performed with 12 taxa, four from each domain of life, from the previously described data set comprising 23 ubiquitous proteins[27]. Archaea: *Methanococcus jannaschii*, *Archaeoglobus fulgidus*, *Pyrococcus furiosus* and *Thermoplasma acidophilum*; Eukarya: *Drosophila melanogaster*, *Homo sapiens*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*; Bacteria: *Escherichia coli*, *Bacillus subtilis*, *Mycobacterium tuberculosis* and *Porphyromonas gingivalis*. Optimal models were determined using both maximum likelihood and Bayesian phylogenetic methods. For a hypothesis involving several independent trees, such as model AE+B, each tree in the model was allowed to have its own independent evolutionary model parameters (such as amino acid substitution matrix, shape parameter for the gamma rate distribution, fraction of invariant sites, and empirical amino acid background frequencies), if it improved the likelihood. For a multiple-tree model such as AE+B, the total likelihood is simply the product of the individual likelihoods from each independent tree. Similarly, in a Bayesian analysis the total marginal likelihood is the product of marginal likelihoods from each independent tree. The AIC was calculated as $\mathrm{AIC} = L - K$, where L is the log likelihood and K is the total number of parameters in the model. Note that this differs from some common versions of the AIC by a factor of -2, and thus a maximum is preferred; this version was chosen for ease of comparison with the other test scores. No assumptions were made about the positions of the roots of the trees, as all inferred trees are unrooted. For the class II models involving HGT, each protein was given its own branch length and topology parameters; all other parameters were identical to the analogous class I model. The class II models thus implicitly assume that HGT involves the exchange of entire protein-coding genes. All phylogenetic input files are available by request.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 28 August 2009; accepted 17 March 2010.

1. Sober, E. *Evidence and Evolution* Ch. 4 (Cambridge University Press, 2008).
2. Darwin, C. *On the Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life* Ch. 14 (J. Murray, 1859).
3. Raup, D. M. & Valentine, J. W. Multiple origins of life. *Proc. Natl Acad. Sci. USA* **80,** 2981–2984 (1983).
4. Crick, F. H. C. The origin of the genetic code. *J. Mol. Biol.* **38,** 367–379 (1968).
5. Sober, E. & Steel, M. Testing the hypothesis of common ancestry. *J. Theor. Biol.* **218,** 395–408 (2002).
6. Dobzhansky, T. Nothing in biology makes sense except in the light of evolution. *Am. Biol. Teach.* **35,** 125–129 (1973).
7. Hinegardner, R. T. & Engelberg, J. Rationale for a universal genetic code. *Science* **142,** 1083–1085 (1963).
8. Penny, D., Hendy, M. D. & Poole, A. M. Testing fundamental evolutionary hypotheses. *J. Theor. Biol.* **223,** 377–385 (2003).
9. Penny, D., Foulds, L. R. & Hendy, M. D. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature* **297,** 197–200 (1982).
10. Zuckerkandl, E. & Pauling, L. in *Evolving Genes and Proteins* (eds Bryson, V. & Vogel, H. J.) 97–166 (Academic Press, 1965).
11. Doolittle, W. F. The nature of the universal ancestor and the evolution of the proteome. *Curr. Opin. Struct. Biol.* **10,** 355–358 (2000).
12. How true is the theory of evolution? *Nature* **290** (Editorial), 75–76 (1981).
13. Popper, K. R. *Unended Quest: An Intellectual Autobiography* revised edn (Fontana, 1976).
14. Syvanen, M. On the occurrence of horizontal gene transfer among an arbitrarily chosen group of 26 genes. *J. Mol. Evol.* **54,** 258–266 (2002).
15. Woese, C. R. On the evolution of cells. *Proc. Natl Acad. Sci. USA* **99,** 8742–8747 (2002).
16. Burnham, K. P. & Anderson, D. R. *Model Selection and Inference: A Practical Information-Theoretic Approach* (Springer, 1998).
17. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90,** 773–795 (1995).
18. Futuyma, D. J. *Evolutionary Biology* 3rd edn (Sinauer Associates, 1998).
19. Murzin, A. G. How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **8,** 380–387 (1998).
20. Karlin, S. & Altschul, S. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA* **87,** 2264–2268 (1990).
21. Harlow, L. L., Mulaik, S. A. & Steiger, J. H. *What If There Were No Significance Tests? (Multivariate Applications)* (Lawrence Erlbaum, 1997).
22. Reeck, G. *et al.* "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell* **50,** 667 (1987).
23. Mindell, D. & Meyer, A. Homology evolving. *Trends Ecol. Evol.* **16,** 434–440 (2001).
24. Crick, F. H. C. in *Progress in Nucleic Acid Research* (eds Davidson, J. N. & Cohn, W. E.) 163–217 (Academic Press, 1963).
25. Doolittle, W. F. The practice of classification and the theory of evolution, and what the demise of Charles Darwin's tree of life hypothesis means for both of them. *Phil. Trans. R. Soc. Lond. B* **364,** 2221–2228 (2009).
26. Huxley, J. S. *Evolution: The Modern Synthesis* 2nd edn, 397–399 (G. Allen & Unwin, 1943).
27. Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. Universal trees based on large combined protein sequence data sets. *Nature Genet.* **28,** 281–285 (2001).
28. Woese, C. & Fox, G. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA* **74,** 5088–5090 (1977).
29. Poole, A. & Penny, D. Evaluating hypotheses for the origin of eukaryotes. *Bioessays* **29,** 74–84 (2007).
30. Chothia, C., Gough, J., Vogel, C. & Teichmann, S. A. Evolution of the protein repertoire. *Science* **300,** 1701–1703 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements I thank J. Felsenstein, P. Garrity, N. Matzke, C. Miller, C. Theobald and J. Wilkins for critical commentary.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Correspondence and requests for materials should be addressed to D.L.T. (dtheobald@brandeis.edu).

METHODS

Data sets. The original data set comprises 6,591 aligned amino acids from 23 ubiquitous proteins[27]: alanyl-tRNA synthetase, aspartyl-tRNA synthetase, glutamyl-tRNA synthetase, histidyl-tRNA synthetase, isoleucyl-tRNA synthetase, leucyl-tRNA synthetase, methionyl-tRNA synthetase, phenylalanyl-tRNA synthetase β subunit, threonyl-tRNA synthetase, valyl-tRNA synthetase, initiation factor 2, elongation factor G, elongation factor Tu, ribosomal protein L2, ribosomal protein S5, ribosomal protein S8, ribosomal protein S11, aminopeptidase P, DNA-directed RNA polymerase β chain, DNA topoisomerase I, DNA polymerase III γ subunit, signal recognition particle protein and rRNA dimethylase. The original data set was constructed by removing poorly aligned regions and most gapped columns from the CLUSTALW alignment[27]. I constructed a similar data set, using the same proteins from the same taxa, which retained the entire protein sequences. The proteins in this data set were independently aligned with ProbCons[31]. The resulting complete unmodified alignment comprised 25,411 columns, including gaps.

Likelihood phylogenetics. For the LLR and AIC tests, more than 1,800 competing biological models were fit to this data using the method of maximum likelihood and the program ProtTest 1.4 (ref. 32) (defaults) supplemented by independent runs with PhyML 2.4.5 (ref. 33). ProtTest calculates the maximum likelihood for 72 evolutionary models for each tree in each model: B, B-F, B-G, B-GF, B-I, B-IF, B-IG, B-IGF, C, C-F, C-G, C-GF, C-I, C-IF, C-IG, C-IGF, D, D-F, D-G, D-GF, D-I, D-IF, D-IG, D-IGF, J, J-F, J-G, J-GF, J-I, J-IF, J-IG, J-IGF, MM, MM-F, MM-G, MM-GF, MM-I, MM-IF, MM-IG, MM-IGF, MR, MR-F, MR-G, MR-GF, MR-I, MR-IF, MR-IG, MR-IGF, R, R-F, R-G, R-GF, R-I, R-IF, R-IG, R-IGF, V, V-F, V-G, V-GF, V-I, V-IF, V-IG, V-IGF, W, W-F, W-G, W-GF, W-I, W-IF, W-IG, and W-IGF, where the substitution matrices are coded as B = Blosum62, C = CtREV, D = Dayhoff, J = JTT, MM = MtMam, MR = MtREV, R = RtREV, V = VT, and W = WAG. The following letters denote models with further parameters: I = invariant positions, G = gamma distributed rate variation, F = empirical amino acid frequencies. For the class II HGT models, 23 different protein trees were calculated for each cluster of taxa proposed to be genealogically related. For example, the model $[AE+B]^{II}$ comprises 46 different trees—23 different protein trees for Archaea and Eukarya, and another 23 trees for Bacteria. The total log likelihood for a particular class II model is the sum of the log likelihoods for all the protein trees in the model.

Bayesian phylogenetics. All Bayesian analyses were calculated with the parallel version of MrBayes 3.1.2 (ref. 34) and used mixed-rate matrices and gamma-distributed rate variation across sites (16 categories). A uniform (0.0, 200.0) prior was assumed for the shape parameter of the gamma distribution, an unconstrained exponential prior (mean = 0.1) was assumed for the branch lengths, and a uniform prior was assumed for all topologies. Two independent Markov chain Monte Carlo (MCMC) analyses were performed (each with one cold and three heated chains), with all other parameters set to defaults. Convergence was inferred after the cold chain topologies had reached a standard deviation of split frequencies of less than 0.01 (generally never more than 10,000,000 generations). After convergence, the first half of the chain was discarded as 'burn in'. For the class II HGT models, the data were partitioned by protein, and all parameters (topology, branch lengths, state frequencies, amino acid substitution model and gamma shape) were unlinked across partitions.

Phylogenetic networks. Phylogenetic networks were computed and displayed with SplitsTree 4.10 (ref. 35), using the equal angle, consensus network algorithm (threshold = 0, to show all reticulations). The phylogenetic networks shown in Fig. 2 are derived from the maximum likelihood estimates of the 23 individual protein phylogenies using the evolutionary model parameters shown in Table 1.

Model selection test scores. LLR values were calculated directly from the likelihoods output by ProtTest and PhyML. The LLR test for non-nested hypotheses was used as previously described[36], which involves estimating the variance of a centred log likelihood using the per site likelihoods as output by PhyML. The number of parameters K was calculated as follows: one parameter per branch length for all trees in the model, where the number of branch lengths per tree is given by $2T-3$ (T is the number of taxa in a given tree); one parameter per tree if the number of invariant sites was estimated; one parameter per tree if the gamma-distribution shape parameter was estimated; 19 parameters per tree if the empirical amino acid frequencies were estimated. Marginal likelihoods for the Bayes factors were calculated with MrBayes[34] using the harmonic-mean estimator[17]. The LBF was calculated as the difference in the marginal-log likelihoods for each model.

Bounds on model selection scores. Consider three hypotheses: H_A, H_B and H_C. If H_B is a partially constrained hypothesis nested within H_C, then the following inequalities necessarily hold:

$$LLR_{A-B} \geq L_A - L_C \quad (1)$$

$$\Delta AIC_{A-B} \geq AIC_A - L_C \quad (2)$$

where $LLR_{A-B} = L_A - L_B$, $\Delta AIC_{A-B} = AIC_A - AIC_B$, and L_X is the log likelihood for hypothesis H_X. These inequalities follow directly from the definitions of the model-selection scores and the fact that the likelihood for a nested, constrained hypothesis is always less than or equal to the likelihood of the unconstrained hypothesis[16]. Derivations and discussion are provided in the Supplementary Materials. The inequalities are especially useful for the purposes of this work, where H_A is a UCA hypothesis and H_B and H_C are multiple-ancestry hypotheses.

31. Do, C. B., Mahabhashyam, M. S., Brudno, M. & Batzoglou, S. ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.* **15,** 330–340 (2005).
32. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21,** 2104–2105 (2005).
33. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52,** 696–704 (2003).
34. Altekar, G. Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20,** 407–415 (2004).
35. Huson, D. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23,** 254–267 (2006).
36. Vuong, Q. H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57,** 307–333 (1989).