Was the universal common ancestry proved?

ARISING FROM D. L. Theobald *Nature* **465,** 219–222 (2010)

The question of whether or not all life on Earth shares a single common ancestor has been a central problem of evolutionary biology since Darwin[1]. Although the theory of universal common ancestry (UCA) has gathered a compelling list of circumstantial evidence, as given in ref. 2, there has been no attempt to test statistically the UCA hypothesis among the three domains of life (eubacteria, archaebacteria and eukaryotes) by using molecular sequences. Theobald[2] recently challenged this problem with a formal statistical test, and concluded that the UCA hypothesis holds. Although his attempt is the first step towards establishing the UCA theory with a solid statistical basis, we think that the test of Theobald[2] is not sufficient enough to reject the alternative hypothesis of the separate origins of life, despite the Akaike information criterion (AIC) of model selection[3] giving a clear distinction between the competing hypotheses.

Dawkins[4] argued that even though it may, at first, seem unlikely that such a complex structure as the eye evolved by selection, it could have been realized by a long sequence of small evolutionary steps driven by selection. Theobald[2] mentions that statistically significant sequence similarity can arise from factors other than common ancestry, such as convergent evolution due to selection, but such factors were not taken into account in his 'formal' test to reject the independent origins hypothesis.

Table 1 shows that the formal test provides support for a common origin of two putatively unrelated genes, mitochondrial *cytb* and *nd2*, with no homology. However, we believe that this result should not be regarded as evidence of the ultimate common ancestry of *cytb* and *nd2*. This raises a question mark as to the effectiveness of the formal test applied by Theobald[2]. It should be noted that, because alignment gives a bias for common ancestry, we did not make an alignment between *cytb* and *nd2*. To reject the separate origins hypothesis of the domains of life, it would be indispensable to develop a more 'biological' test to show that even by improving the model of the separate origins by taking into account biological factors such as the possibility of convergent evolution due to selection, the UCA hypothesis is still supported by the AIC. To do this, it is necessary to develop an entirely new methodological framework of molecular phylogenetics that is different from the conventional framework that neglects convergent and parallel evolution. Notably, there have been many reported cases of convergent and parallel evolution misleading molecular phylogenetic inference[5–9], and such a method is needed for molecular phylogenetics in general.

Table 1 | A formal test of the common ancestry between mitochondrial genes *cytb* and *nd2*

Test statistic	Score or value	Number of parameters
Common origin		
lnL (*cytb* + *nd2*)	−5,090.20	18
AIC	10,216.4	
Independent origin		
lnL (*cytb*)	−2,503.82	12
lnL (*nd2*)	−2,608.17	12
Total lnL	−5,111.99	24
AIC	10,271.97	

Nucleotide sequences of the mitochondrial genes *cytb* and *nd2* from cow, deer and hippopotamus were analysed by PAML[10] with the GTR + Γ model assuming the relations of ((cow, deer), hippopotamus) for the common origin model. The 5′-terminal 1,038 bp (excluding the initiation codon) were used without making further alignment between the two different genes. The common origin model gave a lower AIC value than the independent origin model. lnL, log-likelihood score.

Takahiro Yonezawa[1] & Masami Hasegawa[1,2]
[1]School of Life Sciences, Fudan University, Shanghai 200433, China.
e-mail: masamihase@gmail.com
[2]Institute of Statistical Mathematics, Tokyo 190-8562, Japan.

Received 2 July; accepted 24 August 2010.

1. Darwin, C. *On the Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life* (J. Murray, 1859).
2. Theobald, D. L. A formal test of the theory of universal common ancestry. *Nature* **465,** 219–222 (2010).
3. Akaike, H. in *Second International Symposium on Information Theory* (eds Petrov, B. N. & Csaki, F.) 267–281 (Akademiai Kiado, 1973).
4. Dawkins, R. *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe without Design* (Longman Scientific & Technical, 1986).
5. Stewart, C.-B., Schilling, J. W. & Wilson, A. C. Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* **330,** 401–404 (1987).
6. Christin, P. A., Salamin, N., Savolainen, V., Duvall, M. R. & Besnard, G. C_4 photosynthesis evolved in grasses via parallel adaptive genetic changes. *Curr. Biol.* **17,** 1241–1247 (2007).
7. Liu, Y. *et al.* Convergent sequence evolution between echolocating bats and dolphins. *Curr. Biol.* **20,** R53–R54 (2010).
8. Li, Y., Liu, Z., Shi, P. & Zhang, J. The hearing gene Prestin unites echolocating bats and whales. *Curr. Biol.* **20,** R55–R56 (2010).
9. Zhong, B., Yonezawa, T., Zhong, Y. & Hasegawa, M. The position of Gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. *Mol. Biol. Evol.* doi:10.1093/molbev/msq170 (2 July 2010).
10. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).

Author Contributions T.Y. and M.H. designed the work, T.Y. carried out the analysis, and M.H. wrote the manuscript.

Competing financial interests: declared none.

doi:10.1038/nature09482

Theobald reply

REPLYING TO T. Yonezawa & M. Hasegawa *Nature* **468,** doi:10.1038/nautre09482 (2010)

Yonezawa and Hasegawa[1] provide an example from two apparently unrelated families of nucleic acid coding sequences for which an Akaike information criterion (AIC) model selection test, similar to mine[2], chooses a common origin hypothesis. Although this may seem surprising, the coding sequences in this example were aligned in the same reading frame. The constraints of the genetic code are expected to induce correlations between these sequences (and among all coding sequences) that are not due to common ancestry. For instance, owing to codon bias and the structure of the genetic code, in these sequences the second codon position is biased towards T (about twofold over average), whereas the third position is usually an A (~50%) and rarely a G (~4%).

One can account for these correlations explicitly by using codon models (as implemented in PAML[3], codonFreq = 2 or 3) or standard amino acid models (as in PhyML[4]). With these more realistic models, independent ancestry is the strongly preferred hypothesis. Furthermore, the raw likelihoods and AIC scores increase significantly (by hundreds to thousands of logs), indicating that codon and amino acid models are greatly superior to the naive nucleotide models.

Yonezawa and Hasegawa[1] point out that I[2] did not explicitly test models in which selection or biophysical constraints generate sequence correlations among proteins with independent origins. Formal phylogenetic models accounting for such factors are currently unavailable; their development would be a welcome advance. Although these are important considerations for proteins with low sequence similarity, neither selection nor physical constraints alone can plausibly generate the high levels of sequence similarity (>55% average sequence identity) observed in the universal protein data set that I used[2,5]. The amount of adaptive convergence necessary to produce thousands of identical amino acids among 23 different proteins from completely independent beginnings is not comparable to the limited molecular convergence seen with, for example, homologous digestive lysozymes[6], in which already highly similar proteins (in function, structure and sequence) later acquired a handful of identical substitutions in parallel.

How could selection or biophysical constraints induce correlations among unrelated sequences? If certain similar amino acid sequences are necessary for performing specific functions (or for adopting a specific tertiary conformation that is necessary for function), then selection for function may 'lead' proteins with independent origins to neighbouring regions of sequence space. However, no particular protein sequence or fold is necessary for any given function. There are abundant examples of proteins with undetectable sequence similarity and different folds that perform the same biochemical and cellular functions[7]. For example, the proteases subtilisin, trypsin and carboxypeptidase have the same active site and mechanism, whereas papain, renin and thermolysin have different active sites and different mechanisms. All six proteases have radically different folds and sequences. Because different folds in general have different sequence requirements, proteins with the same function need not have similar sequences.

Even assuming that a certain protein fold is necessary for a given function, current molecular evidence indicates that sequence requirements for a fold are extremely low—nearly indistinguishable from random. This data comes from many independent sources from throughout biology.

Many large classes of proteins with identical folds have no detectable sequence similarity (for example, families of TIM barrels, carbonic anhydrases, OB-folds, SH3 domains, Rossmann folds and immunoglobulin domains). These proteins provide *prima facie* evidence that sequence requirements for any particular fold and function are nearly indistinguishable from random. Protein domains in the SCOP database[8] from different superfamilies yet with the same fold share ~9% sequence identity[9].

Identical folds with known independent origins have nearly random sequence similarity[9,10]. For example, unrelated proteins with the same fold from the MALISAM database share 8.5 ± 0.4% sequence identity[9,10]. This data can be used to estimate the correlations among independently evolved and created proteins with the same fold, and the correlations are nearly random. In the universal protein data set that I used[2], the average sequence correlation induced by common ancestry is roughly one log-likelihood per site for the most divergent proteins. In contrast, the correlations among independent proteins with the same fold are ~100 times weaker. From this we can estimate that model selection scores for common ancestry hypotheses will be many thousands of logs greater than competing selection hypotheses.

Even the most conserved proteins have not yet reached the limits of sequence space, which has been estimated to be near the random expectation for any given fold and function[11].

These arguments are largely circumstantial and informal. I have not tested all possible competing hypotheses, and my analysis will not be the "last word on common ancestry"[12]. I emphasize that I have in no sense provided an absolute 'proof' of universal common ancestry. One of the great advantages of the model selection framework that I presented is that if a novel model is proposed with a well-defined likelihood function, then we can easily compare it to the common ancestry models and see how it fares.

D. L. Theobald[1]

[1]Department of Biochemistry, Brandeis University, Waltham, Massachusetts 01778, USA.
e-mail: dtheobald@brandeis.edu

1. Yonezawa, T. & Hasegawa, M. Was the universal common ancestry proved? *Nature* **468,** 10.1038/nautre09482 (2010).
2. Theobald, D. L. A formal test of the theory of universal common ancestry. *Nature* **465,** 219–222 (2010).
3. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13,** 555–556 (1997).
4. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52,** 696–704 (2003).
5. 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).
6. Stewart, C. B., Schilling, J. W. & Wilson, A. C. Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* **330,** 401–404 (1987).
7. Omelchenko, M. V., Galperin, M. Y., Wolf, Y. I. & Koonin, E. V. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biol. Direct* **5,** 31 (2010).
8. Andreeva, A. *et al.* Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* **36,** D419–D425 (2008).
9. Cheng, H., Kim, B. H. & Grishin, N. V. Discrimination between distant homologs and structural analogs: lessons from manually constructed, reliable data sets. *J. Mol. Biol.* **377,** 1265–1278 (2008).
10. Cheng, H., Kim, B. H. & Grishin, N. V. MALISAM: a database of structurally analogous motifs in proteins. *Nucleic Acids Res.* **36,** D211–D217 (2008).
11. Povolotskaya, I. S. & Kondrashov, F. A. Sequence space and the ongoing expansion of the protein universe. *Nature* **465,** 922–926 (2010).
12. Steel, M. & Penny, D. Origins of life: Common ancestry put to the test. *Nature* **465,** 168–169 (2010).

doi:10.1038/nature09483