A Universal Trend among Proteomes Indicates an Oily Last Common Ancestor

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

A Universal Trend among Proteomes Indicates an Oily Last Common Ancestor

Ranjan V. Mannige[1,2,3,4]*, Charles L. Brooks[2], Eugene I. Shakhnovich[1]

1 Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts, United States of America, **2** Department of Chemistry and Biophysics Program, University of Michigan at Ann Arbor, Ann Arbor, Michigan, United States of America, **3** Department of Molecular Biology, The Scripps Research Institute, La Jolla, California, United States of America, **4** Center for Theoretical Biological Physics, University of California San Diego, La Jolla, California, United States of America

Abstract

Despite progresses in ancestral protein sequence reconstruction, much needs to be unraveled about the nature of the putative last common ancestral proteome that served as the prototype of all extant lifeforms. Here, we present data that indicate a steady decline (oil escape) in proteome hydrophobicity over species evolvedness (node number) evident in 272 diverse proteomes, which indicates a highly hydrophobic (oily) last common ancestor (LCA). This trend, obtained from simple considerations (free from sequence reconstruction methods), was corroborated by regression studies within homologous and orthologous protein clusters as well as phylogenetic estimates of the ancestral oil content. While indicating an inherent irreversibility in molecular evolution, oil escape also serves as a rare and universal reaction-coordinate for evolution (reinforcing Darwin's principle of Common Descent), and may prove important in matters such as (i) explaining the emergence of intrinsically disordered proteins, (ii) developing composition- and speciation-based "global" molecular clocks, and (iii) improving the statistical methods for ancestral sequence reconstruction.

Citation: Mannige RV, Brooks CL, Shakhnovich EI (2012) A Universal Trend among Proteomes Indicates an Oily Last Common Ancestor. PLoS Comput Biol 8(12): e1002839. doi:10.1371/journal.pcbi.1002839

Editor: Nick Grishin, UT Southwestern Medical Center at Dallas, United States of America

Received May 3, 2012; **Accepted** October 28, 2012; **Published** December 27, 2012

Funding: This work was funded in part by NIH grant GM068670 to EIS (www.nih.gov). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rvmannige@lbl.gov

Introduction

What did the proteome of the first successful lifeform look like? This question is critical to the understanding of how life as we know it first began on earth. Despite the progresses in ancestral sequence reconstruction methods [1–3], predicting the proteome of the Last Common Ancestor (LCA) from today's proteomes is impeded by the inevitable accumulated errors implicit in any extrapolation method [4]. Also, while interesting predictions regarding ancestral genome behavior can be made [5,6], the sequence reconstruction methods utilized to make those predictions are fraught with precarious (or even potentially "flawed") presuppositions [7–11]. Here we re-address the question of what the composition of the LCA may have been, and provide an answer sourced from simple considerations that our LCA might have had a highly hydrophobic (oily) proteome, which has slowly been equilibrating over evolutionary time. The results are obtained independently of sequence reconstruction methods and previously published statistical techniques [1–3,5,6], which provides a method-independent snapshot into the molecular nature of the last common ancestor.

It is important to first introduce the notion that proteome *compositions* equilibrate at "glacial" speeds (over billions of years; Section S1.1 in Text S1, and discussed below), which will be important in extrapolating trends obtained from present proteomes to properties of the LCA's proteome. While mutations accumulate within a proteome at a relatively steady rate (to the order of about 1 substitution per billion years per nucleotide site [12]), the oil composition of that proteome–described as the cumulative percent composition within the sequence (%FILV) of the four most hydrophobic residues as per the Kyte-Doolittle scale [Phenylalanine (F), Isoleucine (I), Leucine (L) and Valine (V)]–is expected to change at a much slower rate due to reasons such as the proteome's massive size (Section S1.1 in Text S1) and the biochemical impediments associated with drastically changing a protein's composition (Section S1.2 in Text S1). Given this expected glacial drift/equilibration in proteome composition, in accordance with previous discussions [5], one can expect one of three general trends in the change of proteome oil content over even large evolutionary time (negative, neutral or positive correlations which is dependent on the LCA's original proteome oil content; Figure 1).

Also important to tracking changes in oil content over evolutionary time is the finding that, although all species have existed in *some* form for equal amounts of physical time (an expected outcome of common descent), their genomes are not equally *deviated* from the last common ancestor (Sections S1.3 and S1.4 in Text S1); the number of non-synonymous nucleotide substitutions accumulated since emerging from the LCA appears to be proportional to the number of speciation events that the species has encountered sans phylogenetic errors [13–15]. What this indicates is that, while a component of molecular evolution is due to the "constantly ticking" molecular clock caused by neutral or nearly neutral mutations [16–19], another component of genetic deviation from the LCA may be attributed to substitutions associated with

Author Summary

Although of importance to both evolution and protein design, the manner in which the first proteome came to be, and the actual features of the earliest ancestral proteomes are both unknown. Through the analysis of diverse proteomes, we provide glimpses into the composition of the last common ancestor (LUCA) of all lifeforms, which indicate that the earliest/last common ancestor had a proteome that was highly hydrophobic/oily. Notably, the evidence presented (a) indicates that proteomes of all species ranging from bacteria to mammals appear to adhere to the same universal constraint (''oil escape'') set into motion by the last common ancestor more than 3.5 billion years ago, (b) indicates the presence of a previously untapped global (composition-level) molecular clock, and (c) strengthens the non-equilibrium/directional view of amino acid substitutions that challenges central dogmas regarding reversibility in molecular evolution.

speciation events (roughly proportional to species node number in the tree of life; see Methods), even though the exact extent and mechanism of substitutions in this regime is not known [15]. Importantly, it is especially expected that substitutions causing changes in oil content, due to being quite the opposite of neutral in fitness effects (Section S1.2 in Text S1), are expected to dominantly occur not during neutral drift but during the non-neutral component of molecular evolution, i.e., in events such as speciation (Section S1.4 in Text S1).

In this report, we use species node number as a measure of evolutionary age or ''deviatedness'' from the LCA (also roughly proportional to organismal complexity) to study the changes in proteome and protein composition over evolutionary time. While this study may be susceptible to the various expected *local* inaccuracies involved in building the tree of life (ToL), the global trend–that lower node-number organisms are ''older'' genomes with less genomic deviation from the LCA–is a notion that is acceptable, and such a precedence has been set [13–15,20]. Additionally, given our interest in finding potential *low*-resolution or global trends over evolutionary time, the utility of node number is uniquely warranted.

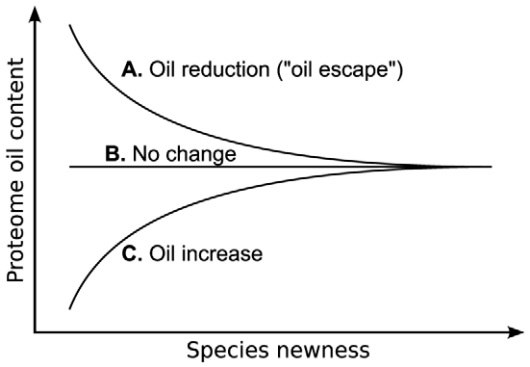

Figure 1. Three scenarios exist for the monotonic drift or evolution of oil content over time, predicated upon whether the last common ancestor's oil content is higher (A), equal (B) or lower (C) than the oil content that is expected from sequence entropy considerations.
doi:10.1371/journal.pcbi.1002839.g001

Results

A common trend (''oil escape'') observed across all proteomes

We obtained and studied all of the proteomes available within the Ensembl genome databases (272 diverse proteomes belonging to 152 distinct species sourced from all domains of life; listed in Section S5 in Text S1) for a relationship between a species node number and its proteome oil content. Remarkably, the proteomes displayed a striking, universal relationship between the proteome oil content (%FILV), and the species node number in both individual Ensembl databases (Figure 2A; annotated axes in each panel are different and elaborated in Figure S1A in Text S1) and the merged data (Figure 2B), which is unexpected given the high diversity of the proteomes studied and the coarse nature of the ToL. Other metrics for oil content (hydrophobicity scales) showed similar results (Figure S2 in Text S1); however, %FILV provided the strictest trend and so is kept as the main metric henceforth.

Given the highly diverse nature of the organisms in our collection, environmental variables are not expected to contribute strongly to this universal trend (also discussed later in Other Trends); Oil escape is also reasserted (albeit loosely) by paleobiological records, where the relationship between proteome oil content and genus-level First Occurrence (FO) records is also negative (Figure 2C; see Methods). The FO records are not

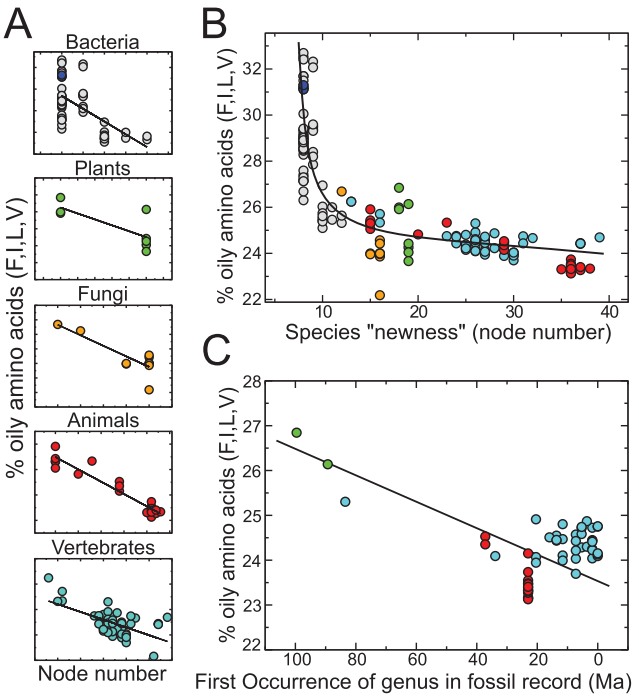

Figure 2. Proteome oil content reduces over ''evolutionary time''. Proteome databases, both individually (**A**; detailed in Figure S1 in Text S1) and cumulatively (**B**; number of data points, $N=152$), indicate a steady reduction in proteome oil content (%FILV) over evolutionary time (organism node number), with a Spearman rank correlation coefficient $r_s=0.84$ and probability $p-value=4.6\times10^{-43}$. Another metric for evolutionary age (paleobiology's ''First Occurrence'' records; **C**) reiterates this trend (Spearman $r_s=0.43$, $p-value=0.0012$, $N=43$; improved to $r_s=0.83$, $p-value=0.00014$ when binned per the abscissa), indicating the existence of a ''super-oily'' predecessor to all that exists. The three archeal proteomes obtained from Ensembl Bacteria are denoted by the blue-filled circles in (**A**, top) and (**B**).
doi:10.1371/journal.pcbi.1002839.g002

necessarily sufficient evidence for oil escape, but rather are referenced for their potential utility of a paleo-geological metric for time (FO dates) in replacing the more abstract node number, which, if possible, would make oil composition a "global" molecular clock (discussed later). Our hope is that Figure 2C may inspire the further collection and utilization of more FO data to that aim.

Oil escape adherence at the protein level

Here we show that "oil escape" occurs not only at the proteome level, but also at the individual protein composition level (which is evidenced by changes in oil content in groups of homologous, and later, orthologous, proteins over organism node space). Our "single protein" studies were performed on clusters of protein sequences homologous to "seed" protein domains listed in the SCOP database (v1.75, redundancy $\leq 10\%$) [21]. Within a cluster, each proteome was represented at most once, and homology was ascertained by BLAST-P's default values [22].

Interestingly, protein-resolution oil escapes, explicitly shown for three homologous clusters in Figure 3A and indicated by negative Spearman's correlation coefficients r_s, are not chance events as would be naïvely expected, but are the dominantly occurring trends among homologous clusters, with over 92.4% of the 5809 homologous protein clusters describing a decline in oil content over node space (Figure 3B). If we discount the statistically insignificant trends (i.e., if we omit those clusters showing two-tailed $p-value>0.05$ or $|r_s|<0.2$, see Figure 3C), then the clusters displaying oil escape further increases to a compelling 97.8% of the 4518 viable clusters (also, analysis of clusters specific to individual Ensemble databases also resulted in qualitatively identical results; Figure S13 in Text S1). Figure 3B cannot be explained by the addition of more hydrophilic domains to a sequence over time, since 87.8% of the 4222 statistically significant ($p-value<0.05$) *size-homogenized* clusters display oil escape (here, size-homogeneity was ensured by culling all sequences that differ in length from the seed sequence by 20%; data not shown). Also, the decrease in oil content over node number is not expected to occur predominantly due to the addition of hydrophilic loops over node number as homologous clusters do not display a bias towards negative relationships between oil content and protein length *within* homologous clusters (Figure S9 in Text S1).

The homologous clusters of proteins used in Figure 3 contain protein pairs related by both orthologous and paralogous relationships (i.e., the proteins can either be related by direct descent, or by duplication and then descent, respectively), and the paralogous relatives in the cluster may cause inaccuracies in gene classification at the functional level [23], although our broad protein-fold resolution of inquiry may not elicit such issues. Still, the same study was carried out among smaller clusters of purely orthologous proteins (COGs database [24]), which also predominantly displayed oil escape among clusters describing significant trends (Figure S10B in Text S1); interestingly, including paralogs into the clusters (Figure S10A in Text S1) reduces the extent of oil escape observed, which indicates that Figure 3 may even be underestimating the extent of oil escape actually occurring.

These data, when consolidated, indicate that a large majority of proteins (homologs, orthologs, and individual domains) display marked oil escape, which reiterates the proteome-level studies in Figure 2, all of which indicate that more ancient proteomes (and so, by extrapolation, the LCA) display oilier residues (and, incidentally, fewer "specificity incurring" residues; see Figure S3 in Text S1) than the newcomers.

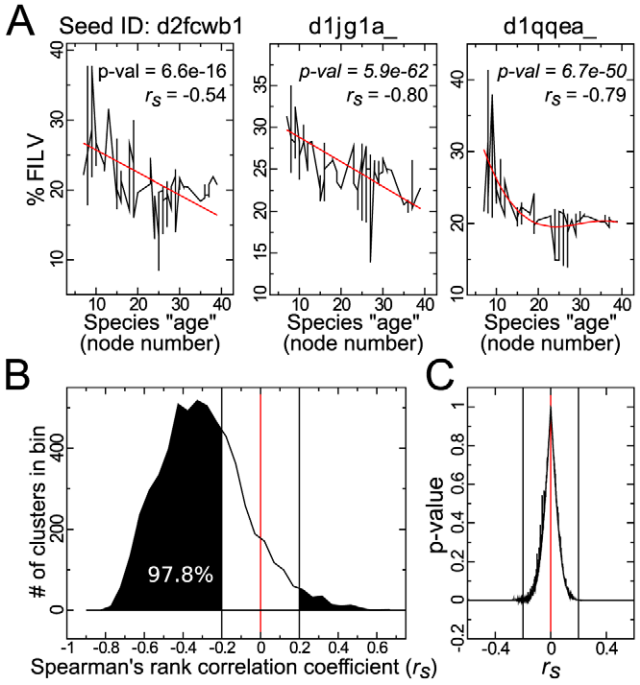

Figure 3. Most individual proteins display "oil escape". Panel **A** shows examples of oil escape among three homologous clusters (seeded by SCOP protein domains listed in Section S2 in Text S1). Similarly, a large majority of the homologous clusters (92.4% of the 5809 studied; see histogram **B**) undergo oil escape. Disregarding clusters with statistically irrelevant trends (defined by $p-value>0.05$ or $|r_s|<0.2$; **C**), the percent of protein clusters displaying oil escape rises to 97.8% (we also obtained similar results for homologous protein clusters limited to each of the individual Ensemble databases [41] –bacteria, plants, fungi,metazoa; Figure S13 in Text S1). The clusters obtained were high in diversity, with an organism node number range of 31.8 ± 0.42 and average size of $\sim 250 \pm 20$ sequences, each sourced from distinct proteomes.
doi:10.1371/journal.pcbi.1002839.g003

Use of comparative methods

Despite the rough (coarse grained) nature of the utilized tree of life (ToL), the trend evident in Figure 2 is interesting, since a common trend–oil escape–appears to unite the behavior of all tested proteomes spanning the domains of life, which offers a unique glimpse into the LCA's molecular composition. This, however, requires that the character trait–oil content–be dependent on the species' ancestral history rather than on external factors (such as non-homogeneous changes in environmental temperatures), i.e., species oil content must display strong dependence on the topology of the tree of life, which is akin to phylogenetic dependence (PD) in purely phylogenetic trees [25,26].

Using comparative methods, we confirm that while oil escape is dependent on the utilized ToL, that attribute is not sufficient to reiterate oil escape, which strengthens the notion of a directed Brownian evolution of oil content from an oily LCA. The dependence of a character state on a particular tree may be estimated using Pagel's λ metric [25,26], which normally ranges from 0 (no dependence) to 1 (strong dependence). We estimated (see method), from established methods [26], a strong bias of the pruned ToL on species oil content ($\lambda=0.99$; compared to $\lambda=0.14 \pm .08$ calculated from 1000 datasets with shuffled oil contents; Figure S14 in Text S1), which indicates that oil content is strongly dependent on a species' evolutionary history, and less

likely swayed (or biased) by "other" (phylogenetically independent) forces not associated with shared history (e.g., non-homogeneous, evolutionary pressures).

It is also important to note that PD, while useful in precluding non-tree-based (or non systematic) biases in our trend, is not sufficient to reproduce the oil escape trend. This is easily evidenced by studying a "neutral drift" version of the original tree, where branches are modified to ensure that all species ages are identical (root to tip lengths are identical); while this tree does not display oil escape (given that all species ages are identical and not node number dependent), just like in the original ToL, high dependence is observed ($\lambda = 0.99$; see Figure S14 in Text S1).

Reconstruction of the LCA state. From ancestral state estimation methods [27] (described in Methods), the "neutral drift" tree describes a random diffusion from an estimated ancestral state (LCA's estimated oil content $\alpha = 25.9\%$, rate of change $\beta = 0.03\%$ per node, and average error $\bar{\varepsilon} = n^{-1}\sum \varepsilon_i = -0.9$; notation borrowed from [27]) that is very close to today's average proteome oil content (26.1%), while the original "speciation" tree describes a steady drift from an oily LCA ($\alpha = 30.1\%$; $\beta = -0.18\%$ per node and $\bar{\varepsilon} = -0.7$), given the high α and substantially negative β. The "neutral tree" model of evolution by random diffusion is disregarded since the expected symmetric distribution of species about the predicted ancestral state is not observed (Figure S15 in Text S1). This leaves us with the "oil escape" model proposed in this paper (e.g., Figures 2,3), which may now be described as a "biased" Brownian motion in genome/proteome space at play particularly during speciation events.

Discussion

On tree completeness and bias

We discuss the two major problems that may potentially plague tree-related studies. First, the number of sequenced genomes (whose species populate the tree) is minuscule compared to the number of species observed in the biological universe (the incompleteness problem). Second, the genomes that are sequenced may be biased with respect to historic choice of model species, ease of handling new species, *et cetera* (the selection bias problem). Here we demonstrate how such biases do not disrupt our tree-related inferences.

Incompleteness problem. (i) It has been shown that the maximum likelihood comparative methodology used above is robust to incomplete phylogenies [26], which is reflected in our finding that the generalized least squares estimate for ancestral oil content is consistently reiterated (averaging at $\sim 30.2\% \pm 0.11$) even when up to 70% of the species collection is randomly culled (Figure 4). *(ii)* the ToL utilized in our studies is obtained by pruning NCBI's ToL [28,29] *without* collapsing single-children nodes [30], i.e., the node number (or shared path) of a species (or pair of species) within the pruned tree is equal to its corresponding value in the exhaustive NCBI tree (Section S6.1 in Text S1). Given the extensive species coverage in NCBI's taxonomy records (over 230,000 species are recorded [29]), the addition of newly sequenced species to our study will likely not modify the node number values for the extant species (or extant values V), and so, adding new species proteomes to our relatively incomplete set of species will likely have no affect on the tree topology. These points indicate that while adding new proteomes to our analysis may improve the statistical relevance of our findings, data incompleteness is not expected to strongly disrupt our general inferences.

Selection bias problem. To ascertain the importance of selection bias (of sequenced species) in our results, we reduced the sampling bias by coarse graining our dataset to the genus level (averaging the node numbers and %FILV's within genera), which, while losing some statistical significance, reiterates the oil escape trend with negative %FILV vs. node number Spearman r_s ($r_s = -0.6$; $p-value = 2.1 \times 10^{-8}$; 76 data points). Our studies on randomly culled species sets (Figure 4, expanded in Figure S16 in Text S1) also indicate that bias due to over-sampling of related species is not likely to drastically modify the primary observation of oil escape, since randomized culling of species will likely reduce sampling bias (as oversampled relatives are more likely to be culled than unique species), which does not change our estimated ancestral oil content until very high culling frequencies are approached. Finally, it is important to note that the robustness of our data and analysis to incompleteness and selection bias may be due to the highly diverse dataset sourced from varied environments and node space (for example, the following node numbers host diverse species: node number 29 hosts the zebra fish, mosquito, and monkey; node number 23 hosts the louse and the platypus).

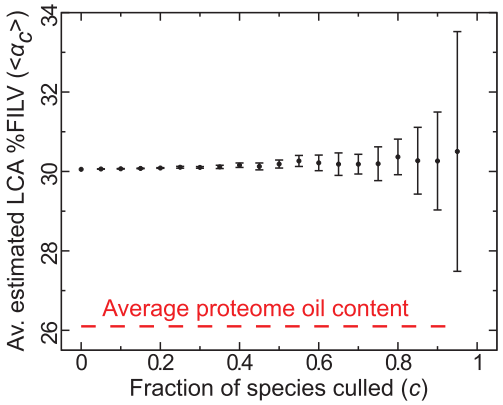

Figure 4. Estimated ancestral oil content vs. fraction (c) species culled. We obtained generalized least squares estimates of ancestral states using previously published methods [27], which indicate that the ancestral state estimates remain at the same average value ($\sim 30.1\%$) even when the species dataset is randomly culled down to 95% (100 randomly culled sets per c). Interestingly, stastical relevance is only lost at $c > 0.70$, where the variance in estimated oil content (shown as error bars for 100 randomly culled sets per c) exceeds 0.25. This figure in expanded form is available in Figure S16 in Text S1.
doi:10.1371/journal.pcbi.1002839.g004

In conclusion to the previous sections, oil escape has been evidenced from simple regression analysis of whole proteomes (Figure 2), protein clusters (Figure 3), and by comparative methods. Finally, given that the biases due to incompleteness and sampling of the sequenced species and are not expected to negate/diminish the general trend of oil escape (see text, Figure 4, and Figures S15 and S16 in Text S1), we expect that oil escape is significant as a trend and merits discussion.

Other trends

It must be noted that while other phylogenetically dependent traits *may* modulate oil escape, given the highly diverse nature of the tree, the possibility of explaining the entire trend with other niche based relationships/dependences is low. Figures S4 and S9 in Text S1 already indicate that oil escape can not be satisfactorily described by the increase of the protein length or the addition of loops or intrinsically disordered regions within globular proteins. Here, we will discuss trends in amino acid and composition previously reported in the literature to ensure that other potential relationships are also independent of the oil escape seen here, i.e., oil escape is a previously unreported trend.

(i) Optimal growth temperatures (OGT). A previous report has established a strong positive correlation ($r=0.93$) between proteome %IVYWREL (single-letter amino acid codes used) and a prokaryote's optimal growth temperature or OGT [31] (the relationship is reiterated in Figure S5A in Text S1). At first glance, species age (node number) and proteome %IVYWREL appears to also correlate well ($r_s=-0.823$, $p-value=1.2\times10^{-38}$; see Figure 5A), which could have indicated an interesting relationship between change in living conditions (temperature) and the progression of complex lifeforms. However, the utility of %IVYWREL as an OGT "thermometer" is contingent upon equal representation (and positive correlation) of both high and low hydrophobic amino acid groups within IVYWREL (i.e., %ILVW and %ERY are also expected to positively correlate with each other; see Figure S5B–F in Text S1); given the lack of positive correlation between %ILVW and %ERY ($r_s=-0.47$, $p-value=9.4\times10^{-10}$), the trend in Figure 5A is merely a result of oil escape, which is also indicated by the opposing relationships in Figure 5B,C; the trend in Figure 5A appears to be caused primarily by oil escape.

(ii) Oil escape is not a spurious relationship. With the advent of genome and proteome databases, a number of additional interrelationships associated with genome/proteome composition have been reported. Of particular interest to the focus of this paper (oil escape) are the reported relationships between %GC content and genome size [32], organismal complexity and genome/proteome length [33], and %GC and hydrophobicity [34–38]. It is important to exclude the possibility that oil escape may be a consequence of potentially more strongly evidenced relationships. For example, could %GC content, which is known to be a correlate with hydrophobicity [39,40], be a stronger correlate with node number? Or are other sets of correlations transitively inducing the effect of oil escape?

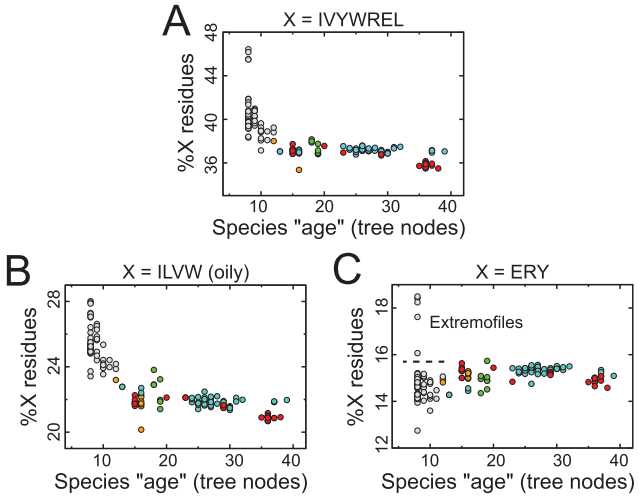

Figure 5. A universal reduction in %IVYWREL over time is a result of strongly reducing oil content. At first glance, it appears as though the combined fraction of IVYWREL amino acids in a proteome ($F_{IVYWREL}$), which is a correlate of prokaryote Optimal Growth Temperature in prokaryotes [31], is also inversely correlated with node number or evolutionary time for both prokaryotes and eukaryotes (**A**; $r_s=-0.82$, $p-value=1.2\times10^{-38}$, $N=152$). However, this effect is primarily caused by the hydrophobic residues of the set (IVWL; **B**; $r_s=-0.86$,$p-value=1.9\times10^{-46}$), while the rest of the amino acids (ERY) show the opposite mildly *increasing* relationship with respect to node number (**C**; $r_s=0.4$,$p-value=2.4\times10^{-9}$). We can more certainly conclude that the trend in $F_{IVYWREL}$ seen in (**A**) is only a result of "oil escape" given the following information: (i) both groups (IVWL and ERY) are required to be positively correlated in order for IVYWREL to be predictive of OGT (Figure S5A–D in Text S1), which is not the case (for each proteome, the Spearman correlation between F_{IVWL} and F_{ERY} is $r_s=-0.47$, $p-value=9.4\times10^{-10}$; also see **B** and **C**), and (ii) oil content alone is not well correlated with OGT (Figure S5E,F in Text S1), and so the two trends of "oil escape" and decrease in OGT can easily be unambiguously distinguished. The extremofile outliers in (**C**) are: *Pyrococcus* (*P*) *abyssi* (abscissa value: 8, ordinate value: 18.4), *P kodakaraensis* (8, 18.5), *P furiosus* (8, 18.2), *P horikoshii* (8, 17.6), and *Bacillus halodurans* (8, 16.1).
doi:10.1371/journal.pcbi.1002839.g005

In order to tackle these questions, we calculated all possible relationships between species cDNA %GC (also obtained from Ensembl genomes [41]), proteome oil content (%FILV), species node number (N), and proteome length L (a proxy for organismal complexity in prokaryotes [33,42–44]), which can be described as a complete correlation network (Figure 6A) where labeled edges (***a***–***f***) indicate Spearman correlation coefficients between properties shown as nodes. It is evident from this graph that oil escape (***d***) displays a Spearman correlation that is highest in magnitude (and consequently, also lowest in p-value). While this may hint at the independence of oil escape from other variables, only statistical tests [45] are able to to strike out the possibility that other variables in Figure 6A are incapable of causing oil escape. This requires the introduction of a statistical criterion that establishes transitivity (and hence the possibility of causality [46]) to a triplet of variables [45,47]. Given three variables i, j and k, and their Spearman or Pearson correlation coefficients r_{ij}, r_{ik}, and r_{jk}, one can show that if $r_{ij}^2+r_{ik}^2>1$, then r_{jk}'s sign must be commensurate to the two other relationships, i.e., r_{jk}'s sign must equal that of $r_{ij}r_{ik}$ [45,47]. However, when this criterion is not met, the trend between j and k (indicated by r_{jk}) can not be caused by the two other correlations. Therefore, while causality may be incapable of being proven using correlations, the *lack* of causality or transitivity can be shown in some situations. Given this criterion, we can conclude that none of the observed adjacent relationships in Figure 6A are able to *cause* oil escape ***d*** (since, from Figure 6A, both $\boldsymbol{b}^2+\boldsymbol{f}^2$ and $\boldsymbol{c}^2+\boldsymbol{e}^2$ are less than 1).

(iii) GC content vs node number: a possibly independent trend. So far, we have shown that oil escape (***d***) is likely not caused by any other pair of correlations in Figure 6A. However, we are still left with the question of which of the remaining correlations are caused by others. For example, while a strong relationship is indicated between species %FILV and %GC (***c***), the relationship is possibly not direct, which we can be shown by a practical implementation of "proof by contradiction". First, we assume that a non-spurious negative relationship exists between oil content and %GC, i.e., the expected biases [39,40] in the codon table (or genetic structure) [48,49] are the main causes of the %FILV-%GC relationship. With this assumption, we are able to quantify the expected %FILV-%GC relationship by generating, for each $\%FILV=f\in\{0,1,2,\ldots,99,100\}$, 10,000 randomly assembled genes that each code for a hundred amino acid peptide with oil content f. From these pairs of gene %GC and corresponding protein %FILV, we created a probability distribution that is represented as a contour plot in Figure 6B (the additional "limiting" contour line dividing the impossible, $p=0$, from possible, $p>0$, was added by analyzing the extremum possibilities from the codon table). Importantly, the relationship observed between proteome %FILV and cDNA %GC (red circles in Figure 6B) displays no correspondence with the expected probability distribution. Also, when sampling generated sequences taken between the observed $22<\%FILV<33$ range, the expected

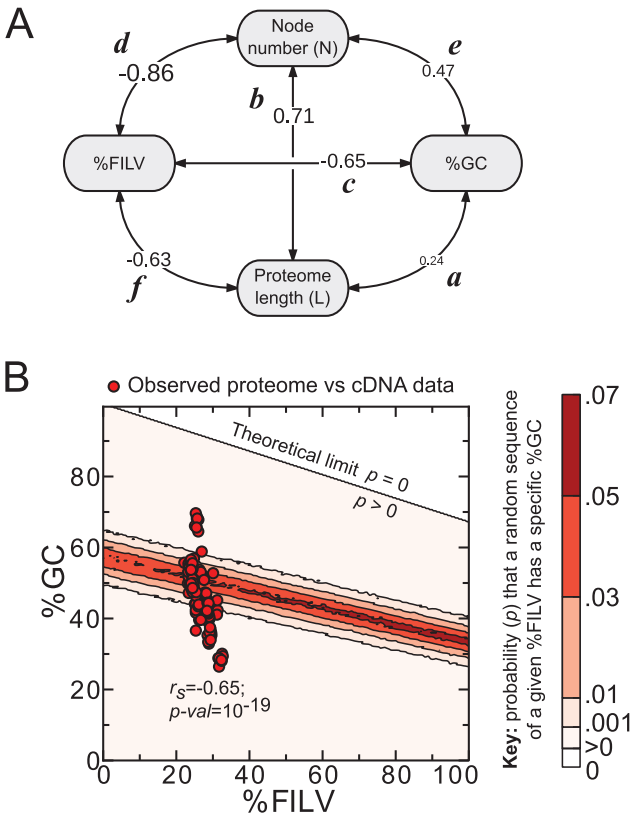

Figure 6. (A) Oil escape is independent of other relationships. All-versus-all Spearman correlation coefficients were calculated for all possible pairs involving *(i)* species cDNA %GC, *(ii)* proteome %FILV, *(iii)* node number N and *(iv)* proteome length L (individual graphs shown in Figure S12B in Text S1). The results indicate that oil escape (***d***) can not be caused by the other variables (see discussion). Relationships between other variables as well as effects of population size on compositions are also discussed in the text. (**B**) **Change in %GC per node number may be an independent trend.** Finally, despite the strong correlation between %FILV and %GC (***d***), the relatively strong relationship between %GC and node number is expected to be independent of oil escape due to the incongruence between expected (contour plot in **B**) and observed correlations (observation shown as red circles in **B**). P-values for each of the regressions ***(a)*** through ***(d)*** are all statistically acceptable with values approximating 0.0032, 4×10^{-24}, 3×10^{-19}, 9×10^{-46}, 6×10^{-18}, and 2×10^{-9}, respectively.
doi:10.1371/journal.pcbi.1002839.g006

Spearman correlation between %FILV and %GC is -0.25 ($p-value<0.001$), which indicates a relatively weak relationship (as indicated by the broad contour distribution in Figure 6B) that is significantly distinct from the observed $r_s=-0.65$ ($p-value<0.001$). These incongruencies between expected and observed relationships indicate that the observed relationship between %FILV (or hydrophobicity [35]) and %GC is not directly caused by the codon table, and, therefore, the relationship between %GC and node number observed is independent of oil escape. We leave the exploration of the cause of trend (***e***) to future research.

(iv) A note on population size. Effective population size is an important factor in population genetics [50], particularly because deleterious mutations that result in offspring of lower fitness are more likely to be maintained in smaller effective population sizes [51]. It is therefore important to establish whether population size is capable of affecting changes in proteome and genome composition. Given the sparse statistics on effective population sizes, we chose to use proteome length (L) as an inverse metric for population size, which is strongly indicated in previous reports [33,42–44]. It is important to recognize that L can be utilized to relate population size to %FILV or %GC only if the relationships display transitivity [45,47] via L, which is a property that still requires verification in future studies (in the absence of transitivity, all relationships involving *population size* must be discounted). The first correlation that we can verify is the expected positive relationship [33,42,44] between organismal complexity (node number N) and proteome length L (***b***), and, assuming transitivity, a negative relationship between complexity (via N) and population size (via L) [33,42–44].

We now direct our attention to the relationship between population size and %GC. Assuming transitivity, %GC content displays only a weak correlation with population size through L (Figure 6A ***a***). While the actual trend (Figure S12B in Text S1) does not qualitatively reproduce the non-monotonic relationship between between %GC and population size reported from simulations [51], the trend (especially when calculated exclusively for bacteria studies, where $r_s=0.47$ ($p-value=2\times10^{-4}$) do match the %GC versus L studies in bacteria [32]. The reason for the incongruence with simulation [51] may be due to the possible lack of transitivity of population size and %GC through L. However, assuming transitivity, our data supports the notion that genomes with high AT content display a fitness advantage compared to genomes with low AT content; this implies a mutational bias towards high AT content, which is universally observed even in high-GC bacteria [52].

Finally, we will evidence that, despite the apparent correlations, %FILV is not likely to be strongly or directly dependent on L and therefore (assuming transitivity) population size. While %FILV does display a negative *correlation* with L, there is high probability that the result is spurious and caused by relationships ***b*** and ***d***, since $\boldsymbol{b}^2+\boldsymbol{d}^2>1$, making ***b*** and ***d*** transitive [45,47], while $\boldsymbol{b}^2+\boldsymbol{f}^2$ and $\boldsymbol{d}^2+\boldsymbol{f}^2$ are both less than 1. The potentially causal relationship is also indicated by the stronger correlation of L with N (Figure 6A ***b***) and a high phylogenetic dependence of L on the tree of life and therefore node number ($\lambda=0.72$, where $\lambda=1$ indicates the strongest dependence and trees with shuffled leaf attributes yield $\lambda\approx0.14$). Therefore, our studies are unable to discern a strong relationship between species population size and proteome %FILV. The lack of this relationship is possibly due to the lack of transitivity between the involved proxy variables. Regardless, the measure of fitness of a protein is not likely to be directly related to the thermodynamic stability of the protein, but to the capacity to maintain a particular near-ground-state ensemble of structures, which is not strongly dependent on only hydrophobicity, but dependent on both hydrophobicity and the buried polar interactions that impart greater folding specificity [53–55], and defray, in part, the entropic cost of hydrophobic collapse [56].

To conclude this section, we find that oil escape (%FILV vs node number) is indeed a significant relationship that is not expected to be caused by other variables such as %GC content, genome length (L), population size (through L), and optimal growth temperatures.

Oil escape appears to be asymptotically slowing down

Besides displaying a strong monotonic trend ($r_s=0.846$), the scatter plot that describes oil escape (Figure 2B) is also modeled well (with correlation coefficient $r\approx0.88$ for two instances described below) by the following asymptotic closed form:

$$\%FILV = \phi_1 + (\phi_2 - \phi_1)e^{-e^{(\phi_3)}ln(N)} \quad (1)$$

Here, N, ϕ_1, ϕ_2, and ϕ_3 are the node number, asymptote (in $\%FILV$), ordinate-intersect, and "rate constant" of the trend, respectively. Keeping all ϕ's variable, and fitting to the data in Figure 2B, we obtained the asymptote $\phi_1 = 23.91\%$ (with $r = 0.879$; Section S4 in Text S1), which, despite the coarse grained quality of node number (and the tree of life), is remarkably close to the expected % of codons ($\%FILV^c$) that code for oily residues (from the universal codon table, $\%FILV^c = 23.44\%$). Also, constraining $\phi_1 = \%FILV^c$ still results in a similarly high $r = 0.877$ (Section S4 in Text S1). These findings strengthen the idea of asymptotic decline in oil content, where the LCA's proteome originated as more oily than expected from sequence entropy considerations (if one expects equal distributions of codon usage), following which later organisms asymptotically approach more moderate values ($\%FILV^c = 23.44$). It is important to note that, for our data, alternative genetic codon tables do not change the value of $\%FILV^c = 23.44$ [48,49], making this a universal maximum for sequence entropy. As a footnote, while one of the six alternative nuclear codon tables has a lower $\%FILV^c$ of ~ 21.88, this "outlier table" is only utilized by a subset of the *Candida* genus of yeasts that is not utilized/represented in our data. Finally, the nearing of later proteomes' oil content towards the "maximal entropy" value ($\%FILV^c = 23.44$) may have interesting implications regarding changes in the rates of molecular evolution.

Oil escape is a passive (entropically driven) drift

The list of species used in our proteome/protein-level studies are diverse, sourced from all domains of life and displays a large range of cellular makeup, body types, biochemistries, evolutionary niches, etc. Given this diversity, we conjecture that oil escape is not driven by an adaptive pressure, as the niche diversity precludes such a broad spectrum pressure. Additionally, given that previous trends such as dependence of proteome composition on oil content do not explain the oil escape trend, we provide a hypothesis based not on extant adaptive pressures but on the features of the last common ancestor (LCA). It is important to note that oil escape must be a trace (fossil) drift from the LCA which would have predominantly begun *after* the production of the first fit proteome with acceptable but relatively high oil content (this is especially the case as all observed oil contents today are expected to be within the range of "acceptable" oil content, which is relatively quite broad, $\sim 10-40\%FILV$; Figure S4 in Text S1). The subsequent reduction in oil content should be considered to be more of a passive drift driven by the impetus to maximize sequence information entropy through evolutionary time, i.e., adaptive pressures and neutral drift would not be driving forces in oil escape, although coupling of the passive drift with either phenomenon is possible. Still, oil escape is not likely to be coupled to neutral drift, which can primarily account for substitutions of nearly-neutral fitness. Alternatively, oil escape may easily be coupled to mutations occurring during speciation events, partly due to the putative increase in substitution rates [13–15,57–64], fixation probabilities [65], and hitchhiking [66–68] of composition-changing mutations with adaptive mutations occurring during speciation events. It is important to reiterate that while oil escape itself might be coupled to adaptive processes, the impetus to shift the oil content of a sequence is driven by the need to maximize information entropy and not adaptive forces.

Oil escape as a molecular clock

This passive drift is slow in effect, as change in oil content (and most other compositions, like charge content) is impeded by the biochemical requirements of a protein (Sections S1.2 and S1.4 in Text S1) and the proteome's massive size (Section S1.1 in Text S1), and therefore is not expected to be "washed away" by a random mutational walk over billions of years (Section S1.1 in Text S1). This ensures the persistence of oil escape even billions of years after the initiation of the trend, i.e., one may consider oil escape as a trace fossil.

These considerations indicate that the oil escape trend summarized in Figure 2B appears to be a rare, universal "reaction coordinate" for evolution, adhered to (at least roughly) by all species tested ranging from bacteria to animals, which may serve as a unique *composition*-level or "global" molecular clock that, if calibrated, may augment the utility of sequence-based "local" molecular clocks [69]. The calibration itself would require more species-level paleobiological first occurrence data for species whose genomes are sequenced. Figure 2C, while rough and statistically sub-optimal, provides a proof of concept for such a calibration project.

It is also important to note that features such as variable substitution rates [70,71] may pose inherent problems to the utility of oil escape as a molecular clock. However, the finding of a genome "core" that describes, even for bacteria, genes with conservative substitution rates, lower involvement in HGTs, and higher reliability in reconstructing robust trees [72–80] indicates that the selection of appropriate protein collections for utility in the global molecular clock may ameliorate a number of inherent problems associated with a global molecular clock. Finally, oil escape is a function of the number of substitutions associated with node number and therefore speciation, making it distinct from neutral molecular evolution. Whether mutations associated with speciation events can be used as a metric for evolutionary age remains to be seen and warrants further research.

Effects of horizontal gene transfer and recombination

The possibility of recombination and horizontal gene transfer (HGT) [81–84] between species indicates that distinct genes may have distinct evolutionary histories, resulting in the tree of life being rendered, in one extreme formulation [72,73], as a *network* of life [74]. Such a free exchange of genes between species would mean a much quicker equilibration of any compositional inequities between species and hence a drastic degradation of the oil escape signal. However, the following points indicate that HGTs are not significantly detrimental to the oil escape signal: (a) complex organisms, while not completely immune to HGT, are generally not affected by this mode of evolution [85], (b) a large percent ($\sim 80-95\%$) of the genes/proteins that are crucial to the core functioning of an organism are not replaced or affected by HGT [72–74], while "accessory" or niche-associated genes partake most in HGTs [74]. The observation of oil escape in both bacterial and complex organism databases (hosting fungi, plants, and metazoa) and orthologous groups spanning all domains of life (mimicking core genes/proteins) indicates that, while horizontal gene transfer may contribute to the noise in the oil escape signal, the complete degradation of the signal is not evident.

While tree reconstructions based on one or few genes may be prone to errors due to recombination and HGT events, utilization of a high volume of independent reference points (such as a whole proteins) results in trees that are robust to recombination and

HGT events [78–80]. This indicates that the NCBI tree of life, which is expertly compiled from a highly diverse set of phylogenetic and taxonomic data, has a very low chance of being drastically affected by HGT and recombination.

Rise of disordered proteins

Intrinsically disordered proteins (IDPs) are low-hydrophobicity proteins that remain unfolded for most of their existence [86]. Recently, a strong connection has been recognized between the rise of complex organisms and the increase in the incidence of IDPs in genomes [87]. The driving forces behind both the increase in organismal complexity and the increase in the incidence of IDPs are still unknown, although they are believed to not be driven by adaptive evolution [88]. It is interesting then that oil escape, when considered as a shift in Gaussian *distributions* of oil content within a proteome, is able to predict the gradual increase in IDP incidence by the gradual decrease in oil content. Particularly, the shift of the distribution of oil content from prokaryote to eukaryote (as observed in Figure S11 in Text S1) appears to push the left tail/fringe of the distribution beyond a hydrophilicity threshold that allows for IDPs to exist [86]. Also, given that bacterial proteomes alone also exhibit statistically-significant oil escape (Figure S1 in Text S1), we can conclude that oil escape is not *caused* by the increase in IDPs (and hence organismic complexity), while the increase in the incidence of IDPs may be driven by oil escape. To our knowledge, oil escape from a relatively oily ancestor is the first non-adaptive explanation of the emergence of IDPs, which fits well with the assertion that no adaptive processes can account for the increase in organismal complexity [88].

Is molecular evolution irreversible?

Given the enormity of sequence space, the chances of reverting to a distant common ancestor are abysmal, a notion that helps paint a one-directional picture of molecular and hence organismal evolution (see, e.g., work by Bridgham *et al.* [89]). To add to this unidirectionality, Jordan *et al.* have provided an interesting observation [5] indicating that there might even be a directionality or "irreversibility" to the *types* of mutations incurred within a protein (e.g., they reported that $L \rightarrow F$ residue substitutions occur more often than $F \rightarrow L$ substitutions), which contradicts the general notion of reversibility or symmetry in point mutations (where $L \rightarrow P$ and $P \rightarrow L$ substitutions are equally probable and the substitution matrix is symmetric). While this paper has been challenged [7–10] primarily due to the reconstruction methods utilized within the paper, we find from a much simpler methodology that such a directionality *may* exist, albeit in not exactly the same form as reported previously (see Figure S7 in Text S1).

Our last common ancestor may not be the most "likely" one

Finally, to underline the dangers involved in extrapolating from present day sequences to a putative common ancestor by statistical methods, it is important to recognize that the LCA's proteome composition may have not been the most "likely" sequence from a sequence entropy standpoint; as indicated in Figure 2B, the ancestral sequence may have high oil content, indicating a less-likely *low* sequence entropy. This raises an interesting and unforeseen pitfall in the backwards extrapolation of the last common ancestor's sequence using methods involving maximum likelihood methods: while backwards-extrapolation may hold ground when used in less diverged groups, when extending further back in evolutionary time, we might be presently exploring a biophysically meaningful but historically meaningless section of sequence space. Such a pitfall may be circumvented by applying a "field" (or mutational bias) to a prediction method.

Concluding remarks

Looking forward, the notion of oil escape has varied implications. Most importantly, oil escape–a universal trend relating oil content to tree node number–evidences the emergence of all known organisms from an oily last common ancestor, and provides a biophysical explanation to the emergence of intrinsically disordered proteins [86]. Additionally, the existence of at least two seemingly *independent* constraints on genome/proteome composition (evidenced in Figure 6) indicates a robustness of evolving genomes placed under multiple *universal* constraints. Finally, the universal fossil trend in proteomes today, aside from serving as the first potential *global* molecular clock, provides glimpses into the earliest proteome, and may provide validity to the possible irreversibility of evolution [5].

Methods

Databases used

The predicted proteome and cDNA sequences used were both obtained from Ensemble genome databases [41] which, at the time of procurement, hosted 272 diverse proteomes belonging to 152 distinct species sourced from all domains of life (listed in Section S5 in Text S1). Both proteome and cDNA sequences were obtained from all predicted genes, known or otherwise [41]. Only amino acids that belong to the natural 20 amino acid repertoire were used in our proteome calculations. Similarly, only nucleotides A, T, G and C were used in our cDNA calculations.

A metric for species evolvedness

Expanding on the general idea that species that are less diverged from the last common ancestor (e.g., bacteria) possess older proteins/proteomes than more diverged species (e.g., fruit fly) [13–15,20], we define a species/proteome's modernity or newness as the minimum number of nodes– *node number*–separating the species or organism from the LCA (root) in the tree of life (ToL). The pruned tree of life (Section S6.2 in Text S1), containing all studied species (Section S5 in Text S1), was obtained using NCBI Taxonomy's Common Tree algorithm [28,29] (accessed from the interactive tree of life [30] with the selected option of leaving "internal nodes expanded"). This tree classification system is based on expert but heuristic gathering of phylogenetic, taxonomic and other biological information (excerpt from the NCBI taxonomy website: "…[The] database does not follow a single taxonomic treatise but rather attempts to incorporate phylogenetic and taxonomic knowledge from a variety of sources, including the published literature, web-based databases, and the advice of sequence submitters and outside taxonomy experts"; http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter = howcite), and is therefore not biased towards specific classification methods. For example, while distinct issues are associated with ignoring horizontal gene transfer [74] and recombination [81] in phylogenetic tree reconstruction, a robust and faithful tracing of lineages is possible with the utilization of larger sequence datasets sourced from whole proteomes [78–80].

First occurrence (FO) records

The paleobiological FO records (listed in parenthesis in Section S5 in Text S1) are obtained from the online paleobiology database (pbdb.org). Due to the relative sparseness of the online records, we associated a species in our collection to its *genus* FO record, which

makes this metric only a coarse grained indicator of evolutionary age (e.g., the mosquito species *Anopheles gambiae*, was linked to the genus *Anopheles*, which has the FO date of 23.030 Ma or million years before current date).

Estimating the λ metric phylogenetic dependence

This section summarizes the phylogenetic method introduced by Pagel [25] and expanded by Freckleton *et al.* [26]. The ToL may be described as a variance-covariance matrix $\mathbf{V}$, where elements $\mathbf{V}_{i,j}$ denote the evolutionary path lengths shared between species i and j (so, $\mathbf{V}_{i,i}$ indicates species i's node number in the original tree and a constant in the "neutral drift" tree). Also, we may set a (column) vector $\mathbf{Y}$ which contains the character traits of the species in the tree (in our case, $\mathbf{Y}_i$ is equal to the oil content of species i). The extent to which a given character state $\mathbf{Y}_i$ depends on its position in the phylogeny may be assessed using an off diagonal multiplier λ, where the variance-covariance matrix is transformed to $\mathbf{V}(\lambda)$ by multiplying all off diagonal elements ($\mathbf{V}_{i,j\neq i}$) by λ where normally $0\leq\lambda\leq1$. The estimated λ will be that which maximizes the likelihood function $p(\lambda)$ (obtained from Equation 4 in Freckleton et al. [26]) for a given $\mathbf{V}$ (tree) and $\mathbf{Y}$ (character set; for examples of such λ-searches, see Figure S14 in Text S1). A maximum likelihood estimate of $\lambda=1$ would indicate that the character state is evolving according to the Brownian model of evolution on the phylogeny, while $\lambda=0$ (where only the diagonal elements in $\mathbf{V}$ remain non-zero) indicates a character trait that is independent of the given phylogeny and shared histories.

For convenience, Equation 4 in Freckleton et al., which is a joint-normal probability density, is repeated here:

$$p(\lambda)=\frac{|\mathbf{V}|^{-1/2}}{(2\pi\sigma^2)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\alpha)^T\mathbf{V}(\lambda)^{-1}(\mathbf{Y}-\mathbf{X}\alpha)\right] \quad (2)$$

where α is the character state at operational time 0, σ^2 is the variance of the Brownian noise introduced in the character state per time unit, and $\mathbf{X}$ is an $n\times1$ "design matrix" of ones which describes a Brownian mode of evolution (setting $\mathbf{X}_i=\mathbf{V}_{i,i}^{-1}$ or $\mathbf{X}_i=\mathbf{V}_{i,i}$, which describes a simplified "biased" Brownian model of evolution, does not significantly change our estimated $\lambda's$). The maximum likelihood estimate for α is

$$\hat{\alpha}=(\mathbf{X}^T\mathbf{V}(\lambda)^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{V}(\lambda)^{-1}\mathbf{Y})$$

and the unbiased (restricted maximum likelihood) estimate for variance σ^2 is

$$\sigma^2=\frac{1}{(n-1)}(\mathbf{Y}-\mathbf{X}\hat{\alpha})^T\mathbf{V}(\lambda)^{-1}(\mathbf{Y}-\mathbf{X}\hat{\alpha}).$$

By maximizing Equation 2, our maximum likelihood λ for a given tree may be estimated. For both the original and "nearly neutral" trees we obtained $\lambda's$ of 0.99 (Figure S14 in Text S1), indicating strong phylogenetic dependence. Note that the estimated $\alpha's$ are likely not accurate, since it's value is directly controlled by the "design matrix" $\mathbf{X}$, whose model ($\mathbf{X}_i=1$) is over-simplified to an unbiased Brownian diffusion about the ancestral state α.

Estimating the ancestral oil content

We use a previously described generalized mean squares model of evolution [27], which describes the character state $\mathbf{Y}_i$ of species i by $\mathbf{Y}_i=\alpha+\beta\mathbf{T}_{i,2}+\varepsilon_i$, where α is the character state of the ancestor (LCA) at operational time 0, β is the estimated rate of change of the character state per operational time unit (e.g., node number), ε_i is the random error, and $\mathbf{T}$ is an $n\times2$ matrix whose first column elements all equal 1 and the second column elements depicts the species operational time/node number (i.e., $\mathbf{T}_{i,1}=1$ and $\mathbf{T}_{i,2}=\mathbf{V}_{i,i}=i$'s operational time or node number) [27]. From the generalized least squares method [27], we can estimate both α and β by solving for

$$(\beta,\alpha)=(\mathbf{T}^T\mathbf{V}^{-1}\mathbf{T})^{-1}(\mathbf{T}^T\mathbf{V}^{-1}\mathbf{Y}) \quad (3)$$

where $\mathbf{V}$ is the variance-covariance matrix of the given tree (as above; i.e., $\mathbf{V}_{i,i}=\mathbf{T}_{i,2}$). Also, the error ε_i for species i may then be obtained from $\varepsilon_i=\mathbf{Y}_i-\alpha-\beta\mathbf{T}_{i,2}$.

While the validity of the previous λ-method of obtaining α was predicated by the choice of the design matrix *and* the statistical inaccuracies of the tree (caused by it's coarse-grained nature), the current reconstruction method does away with the Brownian diffusion model, and so is only dependent on statistical inaccuracies of the tree. However, given that, it is safer to use such α estimates as qualitative checks for hypothesis/data validity (e.g., in Figure 4) rather than an absolute prediction of the ancestral state.

Supporting Information

Text S1 Supporting discussions, figures and data. This document contains discussions and additional analysis that support the findings and claims made in the main article. Additionally, the complete list of species utilized along with the tree of life utilized are listed at the end.
(PDF)

Acknowledgments

RVM thanks Alana Canfield, Dr. Scott Wylie, Professor Ron Hills, and Dr. Isaac Yonemoto for helpful discussions.

Author Contributions

Conceived and designed the experiments: RVM CLB EIS. Performed the experiments: RVM. Analyzed the data: RVM CLB EIS. Contributed reagents/materials/analysis tools: CLB EIS. Wrote the paper: RVM EIS.

References

1. Alvarez-Valin F, Clay O, Cruveiller S, Bernardi G (2004) Inaccurate reconstruction of ancestral GC levels creates a "vanishing isochores" effect. Mol Phylogenet Evol 31: 788–793.
2. Arndt PF (2007) Reconstruction of ancestral nucleotide sequences and estimation of substitution frequencies in a star phylogeny. Gene 390: 75–83.
3. Ianchette M, Diallo AB, Green ED, Miller W, Haussler D (2008) Computational reconstruction of ancestral dna sequences. Methods Mol Biol 422: 171–184.
4. Williams PD, Pollock DD, Blackburne BP, Goldstein RA (2006) Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol 2: e69.
5. Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, et al. (2005) A universal trend of amino acid gain and loss in protein evolution. Nature 433: 633–638.
6. Povolotskaya IS, Kondrashov FA (2010) Sequence space and the ongoing expansion of the protein universe. Nature 465: 922–926.
7. Goldstein RA, Pollock DD (2006) Observations of amino acid gain and loss during protein evolution are explained by statistical bias. Mol Biol Evol 23: 1444–1449.
8. McDonald JH (2006) Apparent trends of amino acid gain and loss in protein evolution due to nearly neutral variation. Mol Biol Evol 23: 240–244

9. Hurst LD, Feil EJ, Rocha EPC (2006) Protein evolution: causes of trends in amino-acid gain and loss. Nature 442: E11–2; discussion E12.
10. Misawa K, Kamatani N, Kikuno RF (2008) The universal trend of amino acid gain-loss is caused by cpg hypermutability. J Mol Evol 67: 334–342.
11. Arenas M, Posada D (2010) The effect of recombination on the reconstruction of ancestral sequences. Genetics 184: 1133–1139
12. Hartl DL, Clark AG (2006) Principles of population genetics. Sunderland, Massachusetts: Sinauer Associates, Inc. Publishers. pp. 317–386.
13. Webster AJ, Payne RJH, Pagel M (2003) Molecular phylogenies link rates of evolution and speciation. Science 301: 478.
14. Pagel M, Venditti C, Meade A (2006) Large punctuational contribution of speciation to evolutionary divergence at the molecular level. Science 314: 119–121.
15. Venditti C, Pagel M (2010) Speciation as an active force in promoting genetic evolution. Trends Ecol Evol 25: 14–20.
16. Kimura M, Ota T (1974) On some principles governing molecular evolution. Proc Natl Acad Sci U S A 71: 2848–2852
17. Kimura M (1979) Model of effectively neutral mutations in which selective constraint is incorporated. Proc Natl Acad Sci U S A 76: 3440–3444.
18. Kimura M (1981) Possibility of extensive neutral evolution under stabilizing selection with special reference to nonrandom usage of synonymous codons. Proc Natl Acad Sci U S A 78: 5773–5777
19. Ohta, Gillespie (1996) Development of neutral and nearly neutral theories. Theor Popul Biol 49: 128–142
20. Vishnoi A, Kryazhimskiy S, Bazykin GA, Hannenhalli S, Plotkin JB (2010) Young proteins experience more variable selection pressures than old proteins. Genome Res 20: 1574–1581
21. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) Scop: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.
22. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
23. Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39: 309–338
24. Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The cog database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28: 33–36.
25. Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401: 877–884
26. Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: a test and review of evidence. Am Nat 160: 712–726
27. Pagel M (1997) Inferring evolutionary processes from phylogenies. Zoologica Scripta 26: 331–348
28. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2009) Database resources of the national center for biotechnology information. Nucleic Acids Res 37: D5–15
29. Federhen S (2012) The ncbi taxonomy database. Nucleic Acids Res 40: D136–D143.
30. Letunic I, Bork P (2007) Interactive tree of life (itol): an online tool for phylogenetic tree display and annotation. Bioinformatics 23: 127–128
31. Zeldovich KB, Berezovsky IN, Shakhnovich EI (2007) Protein and dna sequence determinants of thermophilic adaptation. PLoS Comput Biol 3: e5.
32. Bentley SD, Parkhill J (2004) Comparative genomic structure of prokaryotes. Annu Rev Genet 38: 771–792.
33. Schad E, Tompa P, Hegyi H (2011) The relationship between proteome size, structural disorder and organism complexity. Genome Biol 12: R120.
34. Sueoka N (1962) On the genetic basis of variation and heterogeneity of dna base composition. Proc Natl Acad Sci U S A 48: 582–592.
35. Lobry JR (1997) Influence of genomic g+c content on average amino-acid composition of proteins from 59 bacterial species. Gene 205: 309–316
36. Gu X, Hewett-Emmett D, Li WH (1998) Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. Genetica 102–103: 383–391.
37. Singer GA, Hickey DA (2000) Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol 17: 1581–1588.
38. D'Onofrio G, Jabbari K, Musto H, Bernardi G (1999) The correlation of protein hydropathy with the base composition of coding sequences. Gene 238: 3–14.
39. Bernardi G, Bernardi G (1985) Codon usage and genome composition. J Mol Evol 22: 363–365
40. Bastolla U, Moya A, Viguera E, van Ham RCHJ (2004) Genomic determinants of protein folding thermodynamics in prokaryotic organisms. J Mol Biol 343: 1451–1466.
41. Kersey PJ, Lawson D, Birney E, Derwent PS, Haimel M, et al. (2010) Ensembl genomes: extending ensembl across the taxonomic space. Nucleic Acids Res 38: D563–D569.
42. Lynch M, Conery JS (2003) The origins of genome complexity. Science 302: 1401–1404
43. Yi S, Streelman JT (2005) Genome size is negatively correlated with effective population size in ray-finned fish. Trends Genet 21: 643–646.
44. Lynch M (2006) The origins of eukaryotic gene structure. Mol Biol Evol 23: 450–468.
45. Langford E, Schwertman N, Owens M (2001) Is the property being positively of correlated transitive? The American Statistician 55: 322–325
46. Sugihara G, May R, Ye H, Hsieh CH, Deyle E, et al. (2012) Detecting causality in complex ecosystems. Science 338: 496–500.
47. Lipovetsky S (2002) Comment on "Is the property of being positively correlated transitive?". The American Statistician 56: pp. 341–342.
48. Osawa S, Jukes TH, Watanabe K, Muto A (1992) Recent evidence for evolution of the genetic code. Microbiol Rev 56: 229–264
49. Jukes TH, Osawa S (1993) Evolutionary changes in the genetic code. Comp Biochem Physiol B 106: 489–494.
50. Wright S (1931) Evolution in mendelian populations. Genetics 16: 97–159.
51. Mendez R, Fritsche M, Porto M, Bastolla U (2010) Mutation bias favors protein folding stability in the evolution of small populations. PLoS Comput Biol 6: e1000767.
52. Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6: e1001115.
53. Lumb KJ, Kim PS (1995) A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. Biochemistry 34: 8642–8648.
54. Gonzalez L, Woolfson DN, Alber T (1996) Buried polar residues and structural specificity in the gcn4 leucine zipper. Nat Struct Biol 3: 1011–1018.
55. Bolon DN, Mayo SL (2001) Polar residues in the protein core of escherichia coli thioredoxin are important for fold specificity. Biochemistry 40: 10047–10053.
56. Fernàndez A, Kardos J, Goto Y (2003) Protein folding: could hydrophobic collapse be coupled with hydrogen-bond formation? FEBS Lett 536: 187–192.
57. Mindell D, Sites J, Graur D (1990) Mode of allozyme evolution: Increased genetic distance associated with speciation events. J evol Biol 3: 125–131.
58. Barraclough TG, Savolainen V (2001) Evolutionary rates and species diversity in flowering plants. Evolution 55: 677–683.
59. Lanfear R, Ho SYW, Love D, Bromham L (2010) Mutation rate is linked to diversification in birds. Proc Natl Acad Sci U S A 107: 20423–20428.
60. Lanfear R, Bromham L, Ho SYW (2011) Reply to englund: Molecular evolution and diversification–counting species is better than counting nodes when the phylogeny is unknown. Proc Natl Acad Sci U S A 108: E85–E86.
61. Eo SH, DeWoody JA (2010) Evolutionary rates of mitochondrial genomes correspond to diversification rates and to contemporary species richness in birds and reptiles. Proc Biol Sci 277: 3587–3592.
62. Webster AJ, Payne RJH, Pagel M (2004) Response to comments on "molecular phylogenies link rates of evolution and speciation". Science 303: 173c.
63. Venditti C, Meade A, Pagel M (2006) Detecting the node-density artifact in phylogeny reconstruction. Syst Biol 55: 637–643
64. Venditti C, Pagel M (2008) Model misspecification not the node-density artifact. Evolution 62: 2125–2126.
65. Templeton AR (2008) The reality and importance of founder speciation in evolution. Bioessays 30: 470–479.
66. Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23: 23–35.
67. Barton NH (2000) Genetic hitchhiking. Philos Trans R Soc Lond B Biol Sci 355: 1553–1562
68. Smith JM, Haigh J (2007) The hitch-hiking effect of a favourable gene. Genet Res 89: 391–403.
69. Kumar S (2005) Molecular clocks: four decades of evolution. Nat Rev Genet 6: 654–662.
70. Subramanian S, Lambert DM (2011) Time dependency of molecular evolutionary rates? yes and no. Genome Biol Evol 3: 1324–1328.
71. Yue JX, Li J, Wang D, Araki H, Tian D, et al. (2010) Genome-wide investigation reveals high evolutionary rates in annual model plants. BMC Plant Biol 10: 242.
72. Lawrence JG, Hendrickson H (2003) Lateral gene transfer: when will adolescence end? Mol Microbiol 50: 739–749.
73. Kurland CG, Canback B, Berg OG (2003) Horizontal gene transfer: a critical view. Proc Natl Acad Sci U S A 100: 9658–9662.
74. Ragan MA, McInerney JO, Lake JA (2009) The network of life: genome beginnings and evolution. introduction. Philos Trans R Soc Lond B Biol Sci 364: 2169–2175.
75. Lawrence JG, Ochman H (1998) Molecular archaeology of the escherichia coli genome. Proc Natl Acad Sci U S A 95: 9413–9417.
76. Wolfgang MC, Kulasekara BR, Liang X, Boyd D, Wu K, et al. (2003) Conservation of genome content and virulence determinants among clinical and environmental isolates of pseudomonas aeruginosa. Proc Natl Acad Sci U S A 100: 8484–8489.
77. Dobrindt U, Agerer F, Michaelis K, Janka A, Buchrieser C, et al. (2003) Analysis of genome plasticity in pathogenic and commensal escherichia coli isolates by use of dna arrays. J Bacteriol 185: 1831–1840.
78. Fitz-Gibbon ST, House CH (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res 27: 4218–4222.
79. Tekaia F, Lazcano A, Dujon B (1999) The genomic tree as revealed from whole proteome comparisons. Genome Res 9: 550–557.
80. Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ (2001) Universal trees based on large combined protein sequence data sets. Nat Genet 28: 281–285.
81. Schierup MH, Hein J (2000) Consequences of recombination on traditional phylogenetic analysis. Genetics 156: 879–891.
82. Thomas CM, Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol 3: 711–721.

83. Syvanen M (1994) Horizontal gene transfer: evidence and possible consequences. Annu Rev Genet 28: 237–261.
84. Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284: 2124–2129.
85. Koonin EV, Aravind L, Kondrashov AS (2000) The impact of comparative genomics on our understanding of evolution. Cell 101: 573–576.
86. Uversky VN (2002) Natively unfolded proteins: a point where biology waits for physics. Protein Sci 11: 739–756.
87. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337: 635–645.
88. Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A 104 Suppl 1: 8597–8604.
89. Bridgham JT, Ortlund EA, Thornton JW (2009) An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature 461: 515–519.