A Universal Trend among Proteomes Indicates an Oily Last Common Ancestor

BI Journal Club 11.03.13
Aleksander Sudakov

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

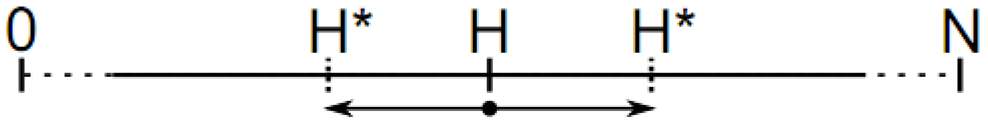
Glossary

Hydrophobic amino acids:
F Phenylalanine
I Isoleucine
L Leucine
V Valine

Charged AA:
E Glutamate
R Arginine
D Aspartate
K Lysine

LCA – last common ancestor
Oily – high hydrophobic amino acids (%FILV) content
Oil escape – decrease in %FILV over evolutionary time

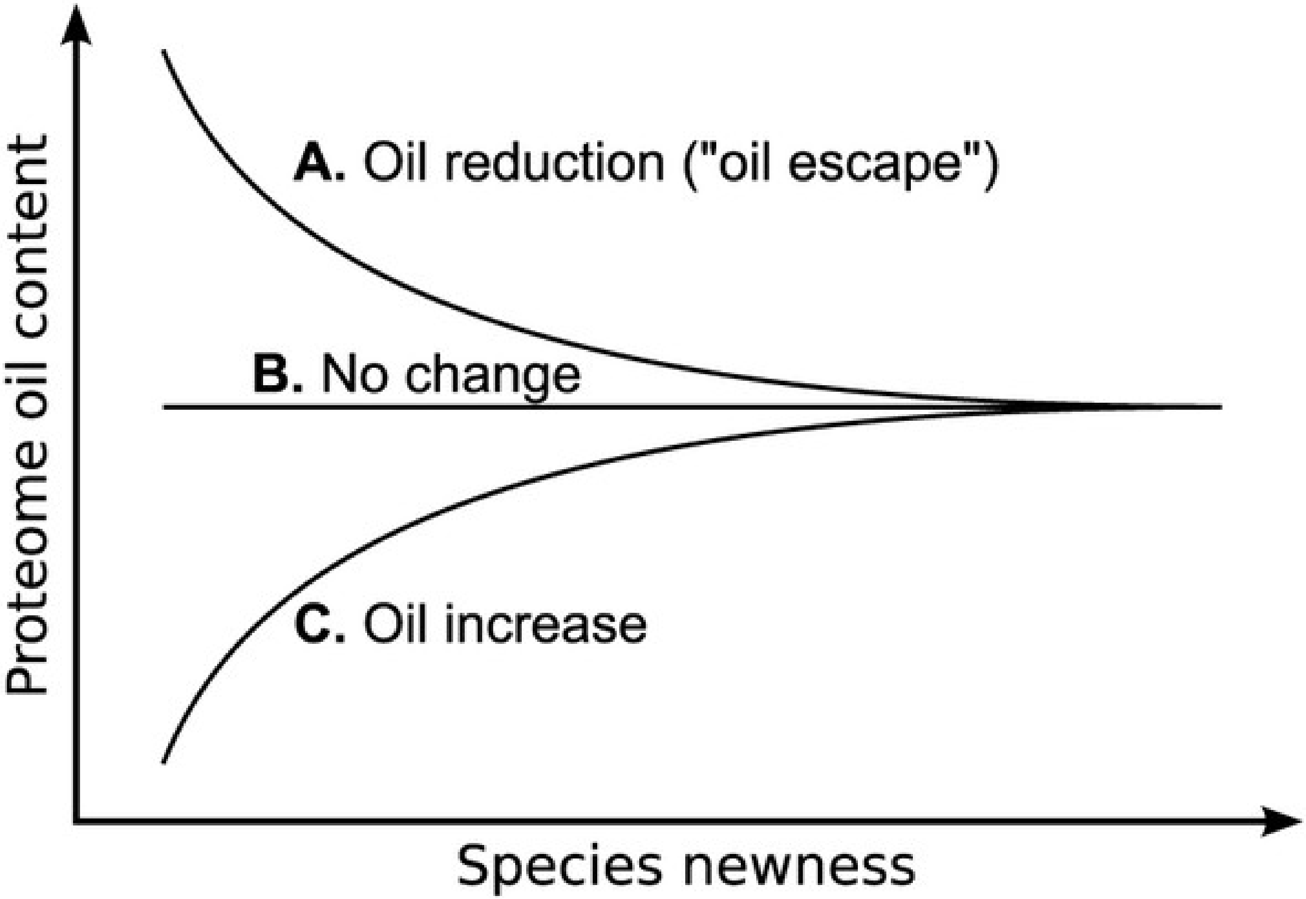
Time-dependent neutral drift



H = number of hydrophobic amino acids in proteome.

H – number of hydrophobic AA
N – total size of proteome

Assuming proteome size $2*10^6$, we would need $25,8*10^9$ mutations to achieve 8% change in H.


Proteome oil content
A. Oil reduction ("oil escape")
B. No change
C. Oil increase
Species newness

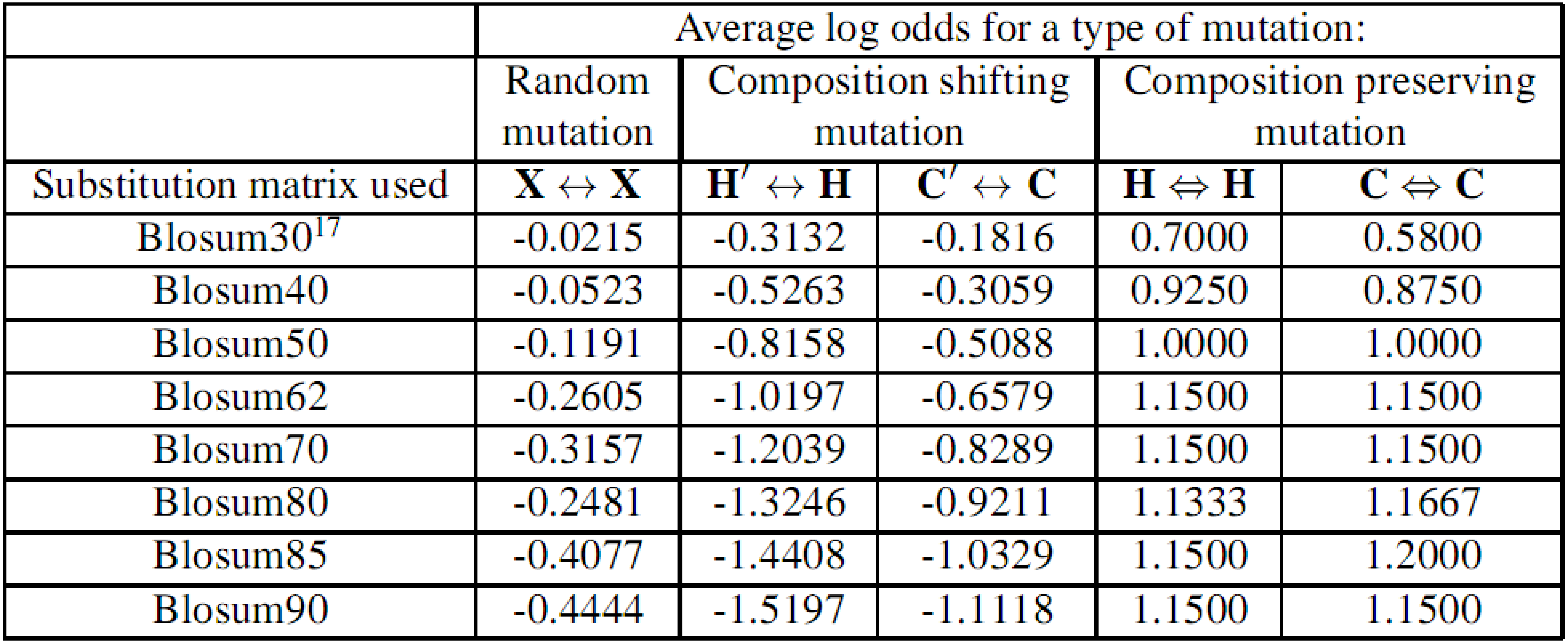
Mechanism of aminoacid drift

- Composition shifting substitutions within a genome are likely not made during *neutral drift*
- Main changes during *speciation events*
- Higher substitution rates, hitch-hiking effect (genes linked to favorable loci), reduced population sizes and higher fixation probabilities

Assumptions

- 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
- Importantly, it is especially expected that substitutions causing changes in oil content, due to being quite the opposite of neutral in fitness effects, 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

Mutations favor preservation of oil and charge



	Average log odds for a type of mutation:				
	Random mutation	Composition shifting mutation		Composition preserving mutation	
Substitution matrix used	$\mathbf{X} \leftrightarrow \mathbf{X}$	$\mathbf{H}' \leftrightarrow \mathbf{H}$	$\mathbf{C}' \leftrightarrow \mathbf{C}$	$\mathbf{H} \Leftrightarrow \mathbf{H}$	$\mathbf{C} \Leftrightarrow \mathbf{C}$
Blosum30[17]	-0.0215	-0.3132	-0.1816	0.7000	0.5800
Blosum40	-0.0523	-0.5263	-0.3059	0.9250	0.8750
Blosum50	-0.1191	-0.8158	-0.5088	1.0000	1.0000
Blosum62	-0.2605	-1.0197	-0.6579	1.1500	1.1500
Blosum70	-0.3157	-1.2039	-0.8289	1.1500	1.1500
Blosum80	-0.2481	-1.3246	-0.9211	1.1333	1.1667
Blosum85	-0.4077	-1.4408	-1.0329	1.1500	1.2000
Blosum90	-0.4444	-1.5197	-1.1118	1.1500	1.1500

biochemical impediments associated with drastically changing a protein's composition

Nodes and Tree of Life

- Species that are less diverged from LCA (e.g. bacteria) possess older proteins/proteomes than more diverged species (e.g. fruit fly) (less speciation events)
- We define species/proteome age or newness as the minimum number of *nodes* separating the species from the LCA in the tree of life
- ToL was obtained using NCBI Taxonomy's Common Tree algorithm

Proteome oil content over „evolutionary time“

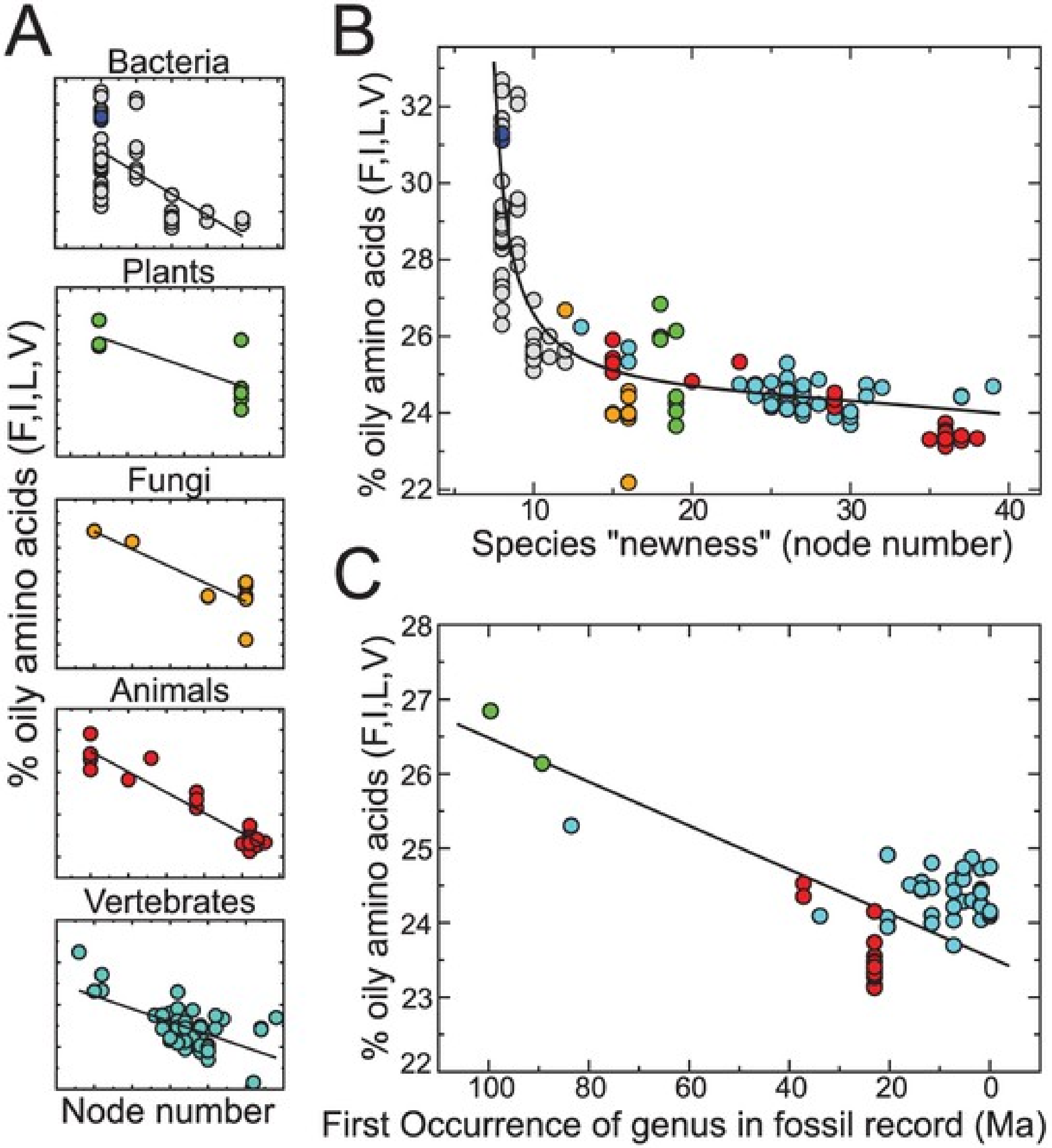


Figure 2. Proteome oil content reduces over "evolutionary time". Proteome databases, both individually (**A**; detailed in Figure S1 in Text S1) and cumulatively (**B**; number of data points, $N=152$), indicate a steady reduction in proteome oil content (%FILV) over evolutionary time (organism node 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

- The proteomes displayed a striking, universal relationship between the proteome oil content (%FILV), and the species node number
- This 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.

Proteome hydrophobicity trend

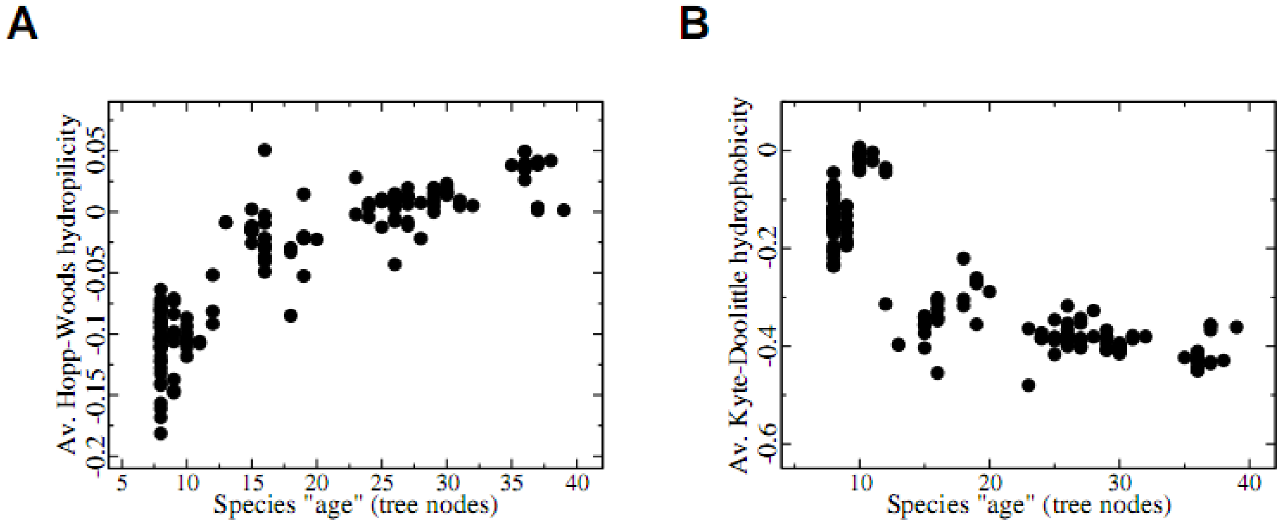


Figure S2: Reflecting the trends in the top four most hydrophobic residues (Fig. 2B), the proteome's average hydrophilicity increases over evolutionary time (**A**), while the proteome's average Kyte-Doolittle hydrophobicity decreases (**B**). In (**A**) the r and p-value for Spearman, Pearson and Kendal τ regression statistics are (0.8798, 2.6e-50), (0.8739, 7.7e-49) & (0.7003, 1.6e-37), respectively. For (**B**), those values respectively are (-0.8091, 1.9e-36), (-0.8266, 2.8e-39) & (-0.6197, 9.3e-30).

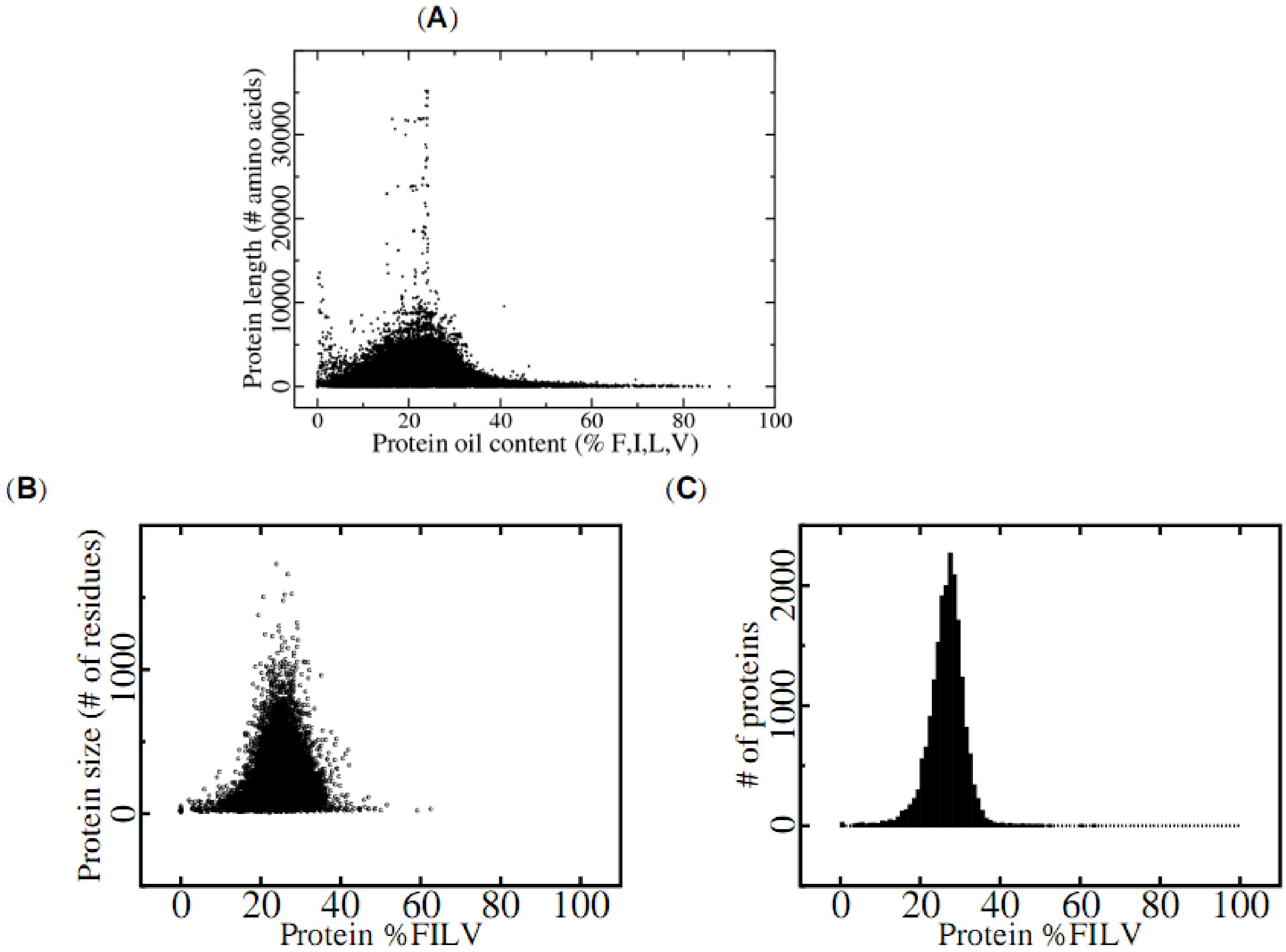


Figure S4: **Protein oil content is not a strongly controlled value.** A scatter plot of oil content vs. protein size in both the predicted proteomes (**A**) and a non-redundant set of the protein databank (**B**) indicates that an evolved protein's oil content depends little on its length (otherwise, one would expect the scatter plot to follow a curved trend). This indicates that, within the observed range, protein "oiliness" is not strongly controlled by protein design requirements, and so the drifts visible in Figs. 1 and 2 are functions of the proteome/protein's history rather than any specific protein design requirement. A histogram of protein oil content sourced from domains obtained from the SCOP database v1.75 is shown in (**C**) for reference.

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

SCOP Database

- Structural Classification Of Proteins
  database of known structural and evolutionary relationships amongst proteins of known structure
- Our “single protein” studies were performed on clusters of protein sequences homologous to “seed” protein domains listed in the SCOP database
- Within a cluster, each proteome was represented at most once, and homology was ascertained by BLAST-P's default values

Individual proteins „oil escape“

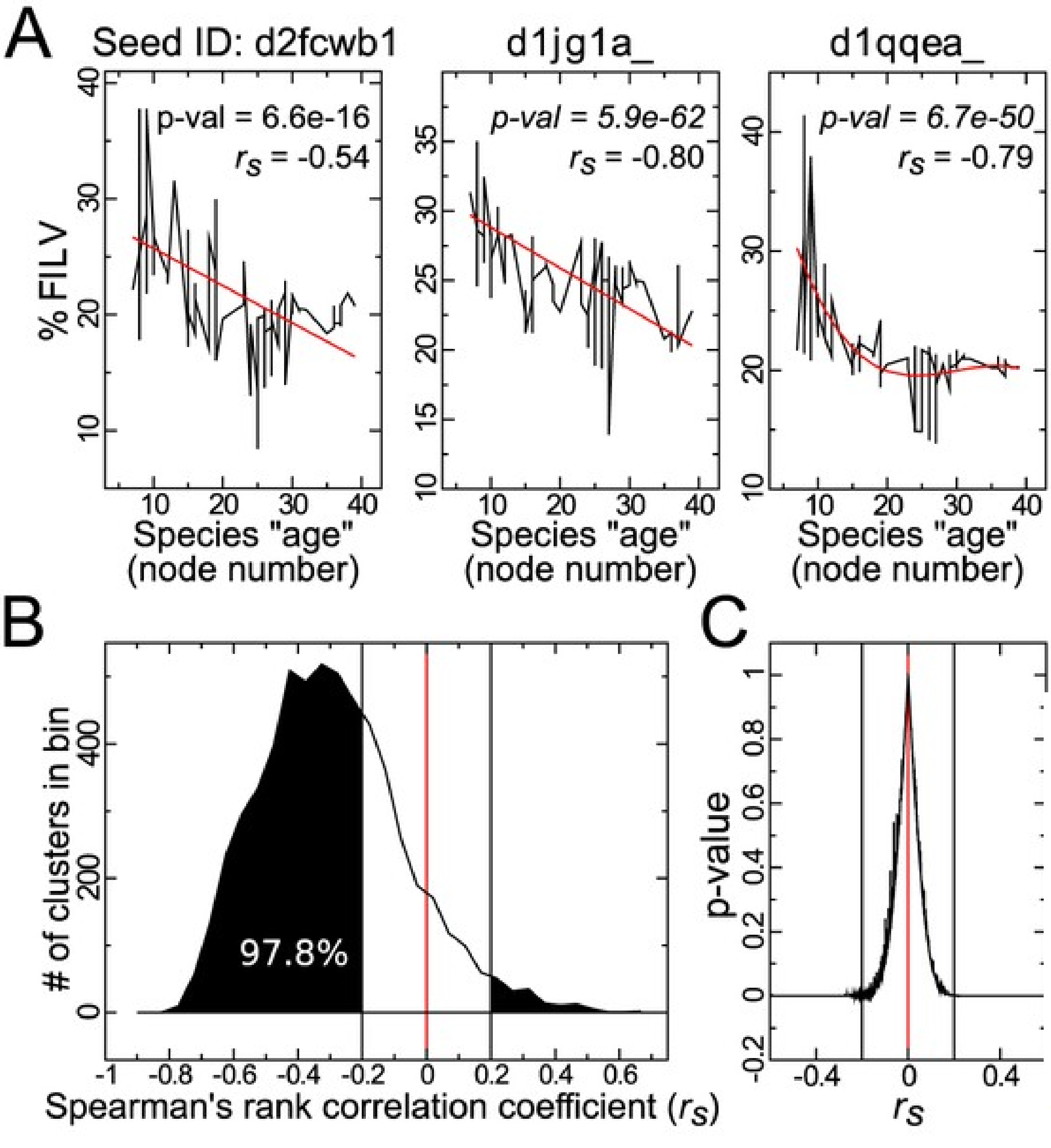
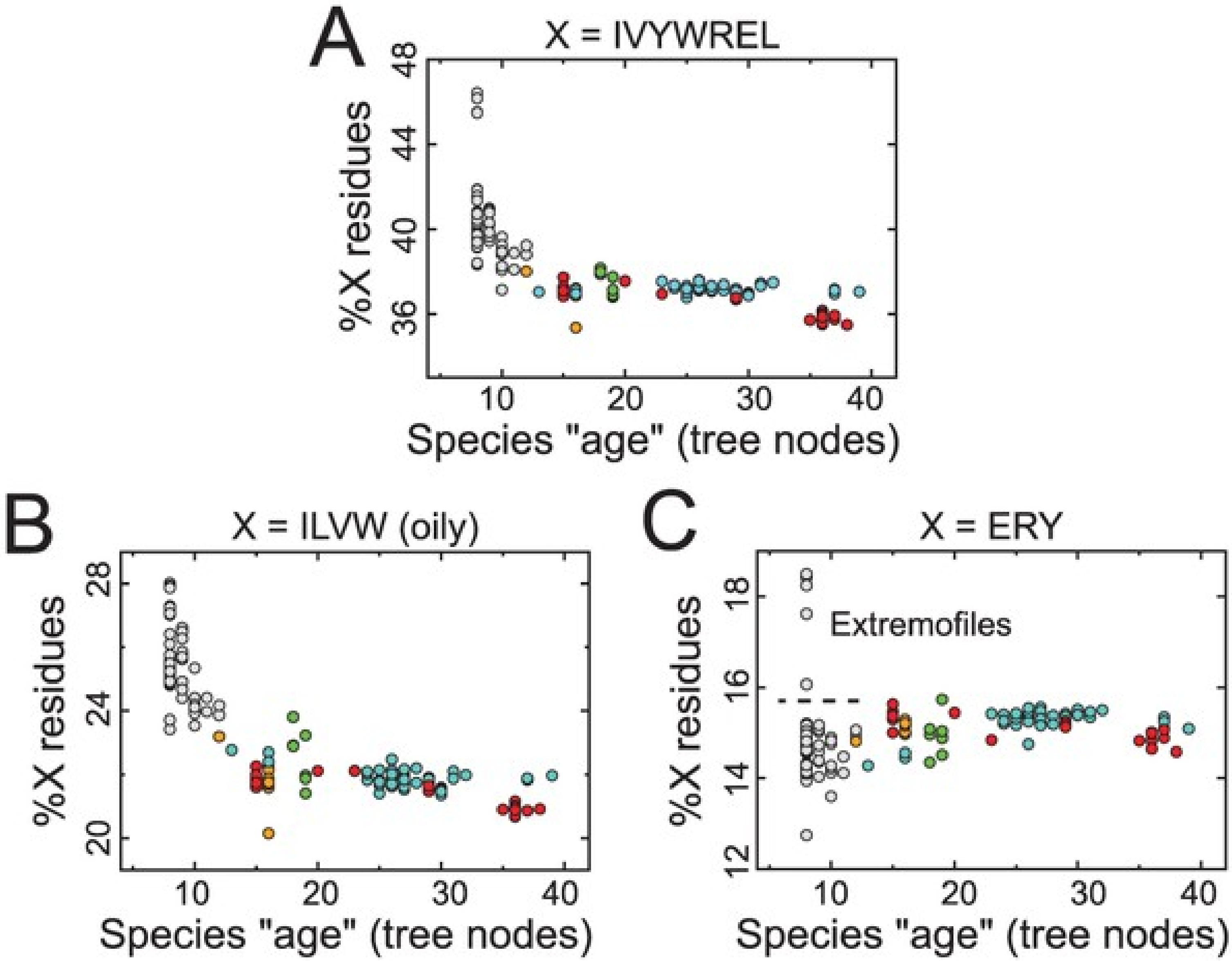


Figure 3. Most individual proteins display ''oil escape''. Panel **A** shows examples of oil escape among three homologous clusters (seeded by SCOP protein domains listed in Section S2 in Text S1). Similarly, a large majority of the homologous clusters (92.4% of the 5809 studied; see histogram **B**) 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.

IVYWREL reduction is result of oil content reduction



Oil escape is independent of other relationships

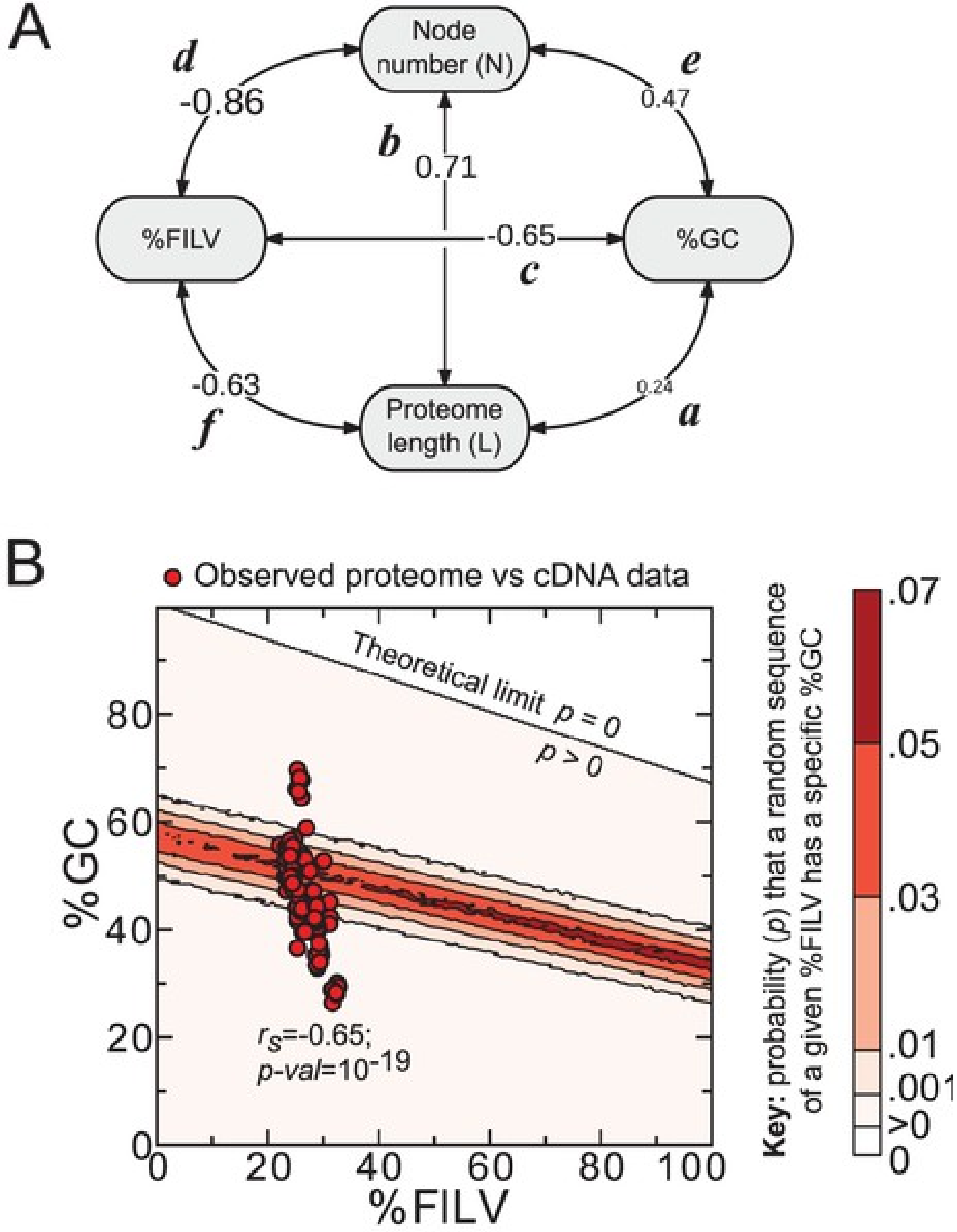


Figure 6. (A) Oil escape is independent of other relationships. All-versus-all Spearman correlation coefficients were calculated for all possible pairs involving *(i)* species cDNA %GC, *(ii)* proteome %FILV, *(iii)* node number N and *(iv)* proteome length L (individual graphs shown in Figure S12B in 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

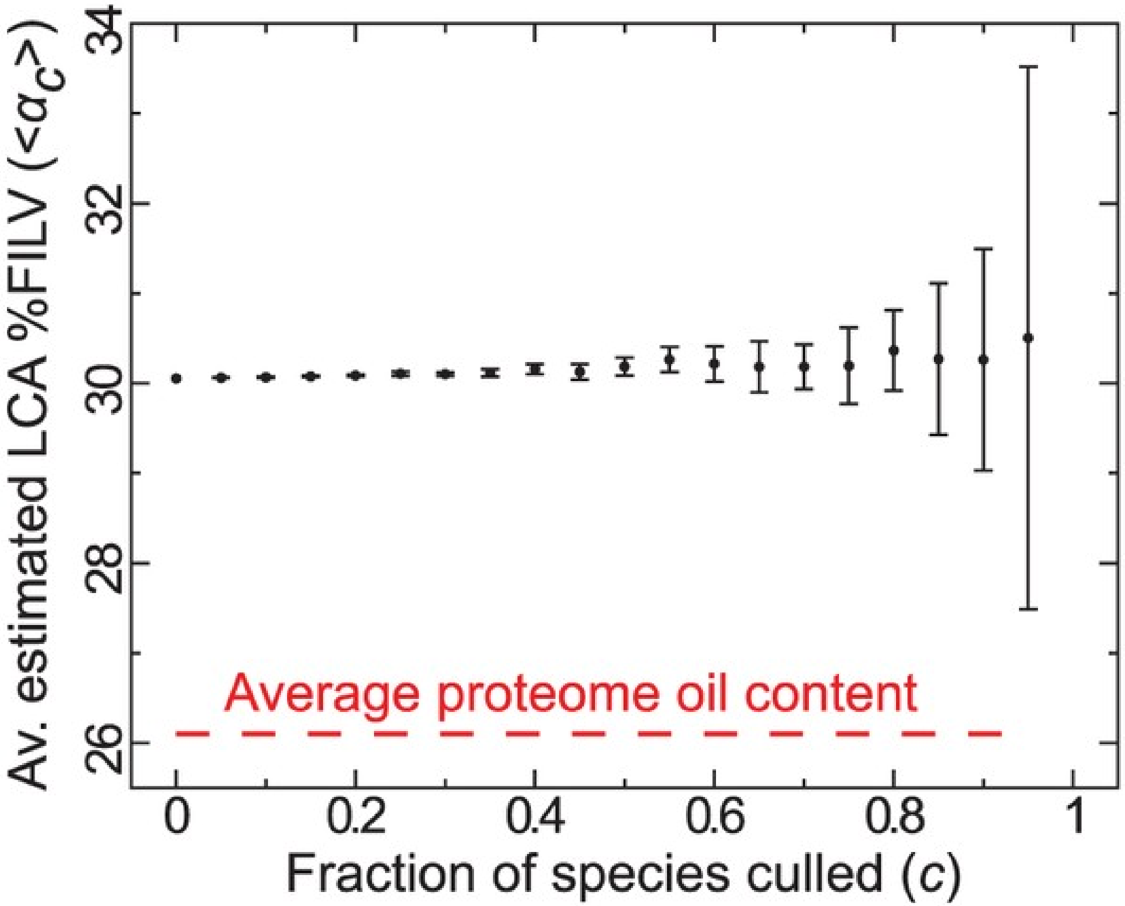
Estimating the ancestral oil content

- We use a previously described generalized mean squares model of evolution, which describes the character state **Y***i* of species i by

  $$\mathbf{Y}_i = \alpha + \beta \mathbf{T}_{i,2} + \varepsilon_i,$$

  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),
  εi is the random error, and
  T is an *n*x2 matrix whose first column elements all equal 1 and the second column elements depicts the species operational time/node number (i.e., **T***i1*=1 and **T***i2*=i’s operational time or node number). From the generalized least squares method, we can estimate both *α* and *β*

Estimated ancestral oil content vs fraction of species



Oil escape appears to be asymptotically slowing down

- We obtained asymptote 23,91%
- Expected percentage of codons that code for oily residues (FILV) 23,44% (alternate genetic codon tables do not change this value)
- LCA proteome originated as more oily than expected from sequence entropy considerations (if equal usage of codons is expected)

Conclusions

- Oil escape appears to unite the behavior of all tested proteomes spanning the domains of life
- Oil escape is passive, entropically driven molecular clock
- Oil escape occurs not only at the proteome level, but also at the individual protein composition level
- LCA proteome composition is predicted to have high oil content ~30%
- LCA may have unlikely composition from entropy standpoint

Thank you

© jaynie
AUSTRALIAN
Emu Oil
MOISTURISING CREAM
with Vitamin E
250g (8.8 OZ)