Increased Frequency of Cysteine, Tyrosine, and Phenylalanine Residues Since the Last Universal Ancestor

Увеличение частоты цистеина, тирозина и фенилаланина со времен последнего универсального общего предка (LUA)

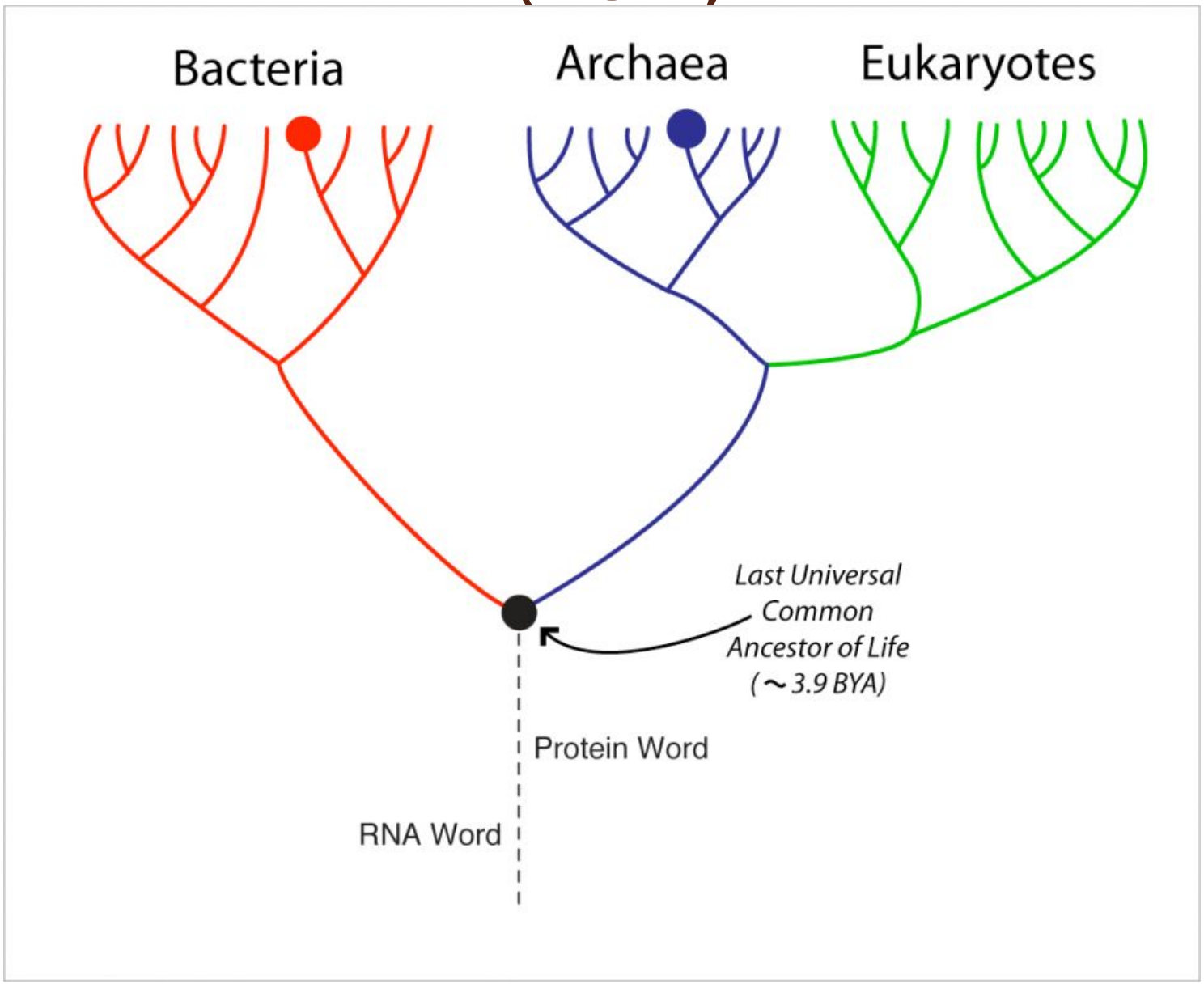
Dawn J. Brooks and Jacques R. Fresco

The Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544

Тихонов Андрей Владимирович

Кафедра генетики СПбГУ

Last Universal Ancestor (LUA)



Виды, взятые в исследование:

- *Aquifex aeolicus*, Aae;
- *Archaeoglobus fulgidus*, Afu;
- *Aeropyrum pernix*, Ape;
- *Bacillus subtilis*, BAC;
- *Chlamydia pneumoniae*, CLA;
- *Campylobacter jejuni*, Cje;
- *Deinococcus radiodurans*, Dra;
- *Escherichia coli K12*, ENT;
- *Helicobacter pylori J9*, HPY;
- *Halobacterium* sp. *NRC-1*, Hbs;
- *Haemophilus influenzae*, Hin;
- *Mycoplasma pneumoniae*, MYC;
- *Methanococcus jannaschii*, Mja;
- *Methanobacterium thermoautotrophicum*, Mth;
- *Mycobacterium tuberculosis*, Mtu;
- *Neisseria meningitidis*, Nme;
- *Pseudomonas aeruginosa*, Pae;
- *Pyrococcus horikoshii*, Pyr;
- *Rickettsia prowazekii*, Rpr;
- *Borrelia burgdorferi*, SPI;
- *Saccharomyces cerevisiae*, Sce;
- *Synechocystis* PCC6803, Ssp;
- *Thermoplasma acidophilum*, Tac;
- *Thermotoga maritima*, Tma;
- *Vibrio cholerae*, Vch;
- *Xylella fastidiosa*, Xfa.

19 бактерий
+ **6** архей
+ **1** эукариот = **26** видов

Набор белков, включенных в исследование

59 семейств COG

(clusters of orthologous groups)

TABLE II

COG protein families included in the LUA protein set

COG0013 alanyl-tRNA synthetase, COG0030 dimethyladenosine transferase, COG0060 isoleucyl-tRNA synthetase, COG0495 leucyl-tRNA synthetase, COG0143 methionyl-tRNA synthetase, COG0016 phenylalanyl-tRNA synthetase α-subunit, COG0072 phenylalanyl-tRNA synthetase β-subunit, COG0442 prolyl-tRNA synthetase, COG0081 ribosomal protein L1, COG0244 ribosomal protein L10, COG0080 ribosomal protein L11, COG0102 ribosomal protein L13, COG0093 ribosomal protein L14, COG0200 ribosomal protein L15, COG0197 ribosomal protein L16/L10E, COG0256 ribosomal protein L18, COG0090 ribosomal protein L2, COG0091 ribosomal protein L22, COG0089 ribosomal protein L23, COG0087 ribosomal protein L3, COG0088 ribosomal protein L4, COG0094 ribosomal protein L5, COG0097 ribosomal protein L6, COG0051 ribosomal protein S10, COG0100 ribosomal protein S11, COG0048 ribosomal protein S12, COG0099 ribosomal protein S13, COG0184 ribosomal protein S15P/S13E, COG0186 ribosomal protein S17, COG0185 ribosomal protein S19, COG0052 ribosomal protein S2, COG0092 ribosomal protein S3, COG0522 ribosomal protein S4 and related proteins, COG0098 ribosomal protein S5, COG0049 ribosomal protein S7, COG0096 ribosomal protein S8, COG0103 ribosomal protein S9, COG0172 seryl-tRNA synthetase, COG0441 threonyl-tRNA synthetase, COG0532 translation initiation factor 2 (GTPase), COG0180 tryptophanyl-tRNA synthetase, COG0525 valyl-tRNA synthetase, COG0202 DNA-directed RNA polymerase α-subunit/40-kDa subunit, COG0085 DNA-directed RNA polymerase β-subunit/140-kDa subunit, COG0250 transcription antiterminator, COG0258 5′-3′ exonuclease (including N-terminal domain of Pol I), COG0592 DNA polymerase III β-subunit, COG0468 RecA/RadA recombinase, COG0550 topoisomerase IA, COG0459 chaperonin GroEL (HSP60 family), COG0533 metal-dependent proteases with possible chaperone activity, COG0201 preprotein translocase subunit SecY, COG0541 signal recognition particle GTPase, COG0552 signal recognition particle GTPase, COG0112 glycine hydroxymethyltransferase, COG0125 thymidylate kinase, COG0237 dephospho-CoA kinase, COG0575 CDP-diglyceride synthetase, COG0012 predicted GTPase

Частота каждой аминокислоты в исследованных видах

TABLE III

Frequency of each amino acid in conserved and non-conserved sequence residues and in the entire protein set, pooled among the 26 species

Amino acids that consistently fall within the top half of the mutability ranking (see Table IV) are assigned high (H) relative mutability; those consistently in the bottom half, low (L) relative mutability; otherwise, undetermined (?) relative mutability.

Amino acid	Conserved	Non-conserved	Protein set	Conserved / Non-conserved	Relative mutability
Ala	0.0814	0.0821	0.0820	0.99	H
Cys	0.0039	0.0085	0.0074	0.45	L
Asp	0.0561	0.0551	0.0553	1.02	?
Glu	0.0779	0.0784	0.0782	0.99	?
Phe	0.0331	0.0388	0.0374	0.85	L
Gly	0.1320	0.0562	0.0738	2.35	L
His	0.0208	0.0192	0.0195	1.09	?
Ile	0.0593	0.0697	0.0673	0.85	H
Lys	0.0611	0.0799	0.0755	0.77	?
Leu	0.1079	0.0847	0.0901	1.27	?
Met	0.0128	0.0265	0.0233	0.48	?
Asn	0.0241	0.0402	0.0365	0.60	H
Pro	0.0629	0.0360	0.0423	1.74	L
Gln	0.0167	0.0375	0.0327	0.45	H
Arg	0.0710	0.0592	0.0620	1.20	L
Ser	0.0261	0.0563	0.0493	0.46	H
Thr	0.0385	0.0520	0.0488	0.74	H
Val	0.0801	0.0783	0.0787	1.02	H
Trp	0.0113	0.0097	0.0100	1.17	L
Tyr	0.0231	0.0317	0.0297	0.73	L

Относительная изменчивость аминокислот

согласно Dayhoff *et al.*, 1978; Jones *et al.*, 1992; Gonnet *et al.*, 1992

TABLE IV

Rank order of relative mutability of the amino acids (from most to least mutable) based on empirical data of Dayhoff et al. (1), Jones et al. (11), and Gonnet et al. (12)

	Dayhoff	Jones	Gonnet
Most mutable	Asn	Ser	Ser
	Ser	Thr	Ala
	Asp	Asn	Thr
	Glu	Ile	Gln
	Ala	Ala	Lys
	Thr	Val	Val
	Ile	Met	Glu
	Met	His	Asn
	Gln	Asp	Ile
	Val	Gln	Leu
	His	Arg	Met
	Arg	Glu	Asp
	Pro	Lys	Arg
	Lys	Pro	His
	Gly	Leu	Gly
	Tyr	Phe	Phe
	Phe	Gly	Pro
	Leu	Tyr	Tyr
	Cys	Cys	Cys
Least mutable	Trp	Trp	Trp

The probability of observing amino acid **j** in a specific genome:

$$p_j = \lambda\left(\sum_i x_i y_i z_i\right)$$

- where *i* represents each codon assigned to amino acid ***j;***
- ***xi, yi,*** and ***zi*** represent the frequency of occurrence of the first, second, and third nucleotides, respectively, of codon ***i*** within coding sequences of that genome;
- and **λ** is a constant such that the sum over all amino acids is equal to one.
- The normalization constant **λ** compensates for probabilities assigned to stop codons.

Частота Cys, Tyr, Phe (наблюдаемая и предсказанная) у 26 исследованных видов

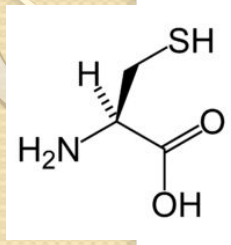
TABLE V

Percent frequency of cysteine, tyrosine, and phenylalanine observed and predicted by neutral evolution in proteomes of 26 species

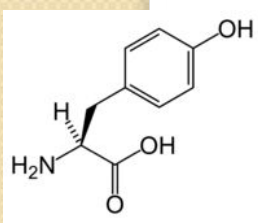
Species abbreviations are as follows: *Aquifex aeolicus*, Aae; *Archaeoglobus fulgidus*, Afu; *Aeropyrum pernix*, Ape; *Bacillus subtilis*, BAC; *Chlamydia pneumoniae*, CLA; *Campylobacter jejuni*, Cje; *Deinococcus radiodurans*, Dra; *Escherichia coli K12*, ENT; *Helicobacter pylori J9*, HPY; *Halobacterium* sp. *NRC-1*, Hbs; *Haemophilus influenzae*, Hin; *Mycoplasma pneumoniae*, MYC; *Methanococcus jannaschii*, Mja; *Methanobacterium thermoautotrophicum*, Mth; *Mycobacterium tuberculosis*, Mtu; *Neisseria meningitidis*, Nme; *Pseudomonas aeruginosa*, Pae; *Pyrococcus horikoshii*, Pyr; *Rickettsia prowazekii*, Rpr; *Borrelia burgdorferi*, SPI; *Saccharomyces cerevisiae*, Sce; *Synechocystis* PCC6803, Ssp; *Thermoplasma acidophilum*, Tac; *Thermotoga maritima*, Tma; *Vibrio cholerae*, Vch; *Xylella fastidiosa*, Xfa.

Species	Cysteine		Tyrosine		Phenylalanine	
	Proteome observed	Proteome predicted	Proteome observed	Proteome predicted	Proteome observed	Proteome predicted
Aae	0.0079	0.0266	0.0415	0.0357	0.0516	0.0260
Afu	0.0118	0.0309	0.0365	0.0296	0.0459	0.0251
Ape	0.0094	0.0311	0.0335	0.0222	0.0275	0.0209
BAC	0.0080	0.0301	0.0348	0.0376	0.0449	0.0321
CLA	0.0160	0.0328	0.0326	0.0457	0.0474	0.0463
Cje	0.0122	0.0296	0.0368	0.0593	0.0600	0.0535
Dra	0.0067	0.0267	0.0230	0.0136	0.0316	0.0127
ENT	0.0117	0.0332	0.0286	0.0298	0.0389	0.0293
HPY	0.0110	0.0303	0.0368	0.0449	0.0542	0.0394
Hbs	0.0075	0.0263	0.0255	0.0148	0.0312	0.0128
Hin	0.0104	0.0321	0.0315	0.0479	0.0447	0.0456
MYC	0.0075	0.0292	0.0323	0.0440	0.0559	0.0385
Mja	0.0128	0.0277	0.0438	0.0516	0.0426	0.0401
Mth	0.0121	0.0283	0.0322	0.0285	0.0365	0.0225
Mtu	0.0088	0.0295	0.0208	0.0148	0.0296	0.0152
Nme	0.0103	0.0288	0.0298	0.0275	0.0412	0.0237
Pae	0.0100	0.0271	0.0254	0.0144	0.0356	0.0136
Pyr	0.0063	0.0303	0.0384	0.0387	0.0460	0.0322
Rpr	0.0110	0.0285	0.0389	0.0603	0.0488	0.0536
SPI	0.0073	0.0270	0.0429	0.0608	0.0619	0.0511
Sce	0.0130	0.0285	0.0380	0.0453	0.0450	0.0385
Ssp	0.0100	0.0335	0.0291	0.0343	0.0401	0.0347
Tac	0.0060	0.0294	0.0464	0.0330	0.0470	0.0272
Tma	0.0071	0.0286	0.0358	0.0332	0.0519	0.0262
Vch	0.0105	0.0334	0.0296	0.0352	0.0407	0.0347
Xfa	0.0119	0.0340	0.0262	0.0277	0.0347	0.0291
Mean	0.0099	0.0298	0.0335	0.0358	0.0437	0.0317

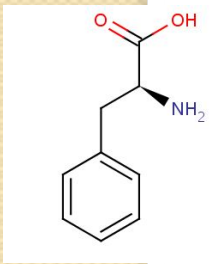
Сравнение частот Cys, Tyr и Phe



- Frequency of **Cys** is significantly less than that predicted in all 26 species ($p << 0.01$), the mean over all species being 1/3 of that that predicted.



- Frequency of **Tyr** is less than predicted in only 15 of the species ($p = 0.28$, which is not statistically significant), and the mean observed frequency of Tyr 0.0335, is close to that predicted, 0.0358.



- Frequency of **Phe** is higher than predicted in 25 species ($p << 0.01$), the mean observed frequency, 0.0437, being 40% higher than predicted.

- Принято считать, что **Cys**, **Tyr**, **Phe** появились в генетическом коде позже других.
- Эти кодоны вводились в генетический код постепенно.
- Вероятно, что увеличение частоты **Cys** продолжается до сих пор, и еще не достигнута его равновесная частота.
- С другой стороны, частота **Phe** превысила частоту, предсказанную теорией нейтральной эволюции. Наблюдается положительный отбор для Phe, который изначально встречался крайне редко.
- То же самое можно утверждать и про **Tyr**, наблюдаемая частота которого существенно не отличается от предсказанной, согласно теории нейтральной эволюции.