

This information has not been peer-reviewed. Responsibility for the findings rests solely with the author(s).

Deposited research article

Preferred codons and amino acid couples in hyperthermophiles

Sávio Torres de Farias and Maria Christina Manhães Bonato

Addresses: Departamento de Biologia Molecular, Universidade Federal da Paraíba, João Pessoa-PB, 58059-900, Brazil.

Correspondence: Maria Christina Manhães Bonato. E-mail: Chrisjampa@aol.com, stfarias@yahoo.com.br

Posted: 19 July 2002

Received: 16 July 2002

Genome Biology 2002, **3(8)**:preprint0006.1-0006.18

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2002/3/8/preprint/0006>

This is the first version of this article to be made available publicly. This article has been submitted to *Genome Biology* for peer review.

© BioMed Central Ltd (Print ISSN 1465-6906; Online ISSN 1465-6914)



deposited research

AS A SERVICE TO THE RESEARCH COMMUNITY, GENOME **BIOLOGY** PROVIDES A 'PREPRINT' DEPOSITORY TO WHICH ANY ORIGINAL RESEARCH CAN BE SUBMITTED AND WHICH ALL INDIVIDUALS CAN ACCESS FREE OF CHARGE. ANY ARTICLE CAN BE SUBMITTED BY AUTHORS, WHO HAVE SOLE RESPONSIBILITY FOR THE ARTICLE'S CONTENT. THE ONLY SCREENING IS TO ENSURE RELEVANCE OF THE PREPRINT TO GENOME **BIOLOGY**'S SCOPE AND TO AVOID ABUSIVE, LIBELLOUS OR INDECENT ARTICLES. ARTICLES IN THIS SECTION OF THE JOURNAL HAVE **NOT** BEEN PEER-REVIEWED. EACH PREPRINT HAS A PERMANENT URL, BY WHICH IT CAN BE CITED. RESEARCH SUBMITTED TO THE PREPRINT DEPOSITORY MAY BE SIMULTANEOUSLY OR SUBSEQUENTLY SUBMITTED TO GENOME **BIOLOGY** OR ANY OTHER PUBLICATION FOR PEER REVIEW; THE ONLY REQUIREMENT IS AN EXPLICIT CITATION OF, AND LINK TO, THE PREPRINT IN ANY VERSION OF THE ARTICLE THAT IS EVENTUALLY PUBLISHED. IF POSSIBLE, GENOME **BIOLOGY** WILL PROVIDE A RECIPROCAL LINK FROM THE PREPRINT TO THE PUBLISHED ARTICLE.



Preferred codons and amino acid couples in hyperthermophiles

Running title: Preferences in hyperthermophiles

Authors:

Sávio Torres de Farias and Maria Christina Manhães Bonato

Address:

Departamento de Biologia Molecular
Universidade Federal da Paraíba
João Pessoa-PB, 58059-900, Brazil.

Correspondence:

M. Christina M. Bonato

Fax: 55-83-2167407.

Email: Chrisjampa@aol.com
stfarias@yahoo.com.br

Abbreviations: optimal growth temperature (OGT); hyperthermophiles (HT); moderate thermophiles (T); mesophiles (M); water accessible surfaces (WAS); pyrimidines (Y); purine (R); all nucleotides (N).

Abstract

Background: Most organisms grow at temperatures from 20 to 50⁰C but some prokaryotes, including Archaea and Bacteria, are capable of withstanding higher temperatures, from 60 to >100⁰C. What makes these cells so resistant to heat? Their biomolecules must be sufficiently stable, especially proteins, to work under these extreme conditions, but the bases for thermostability remains elusive.

Results: The preferential usage of certain couples of amino acids and codons in thermal adaptation was investigated, by comparative proteome analysis, using 28 complete genomes from 18 mesophiles, 4 thermophiles, and 6 hyperthermophiles. In the hyperthermophiles proteomes, whenever the percent of Glu (E) and Lys (K) increased, the percent of Gln (Q) and His (H) decreased, so that the E+K/Q+H ratio was > 4,5; in the mesophiles proteomes, it was < 2,5 and in the thermophiles an intermediary value was observed. The E+K/Q+H ratios for chaperonins, potentially thermostable proteins, were higher than their proteome ratios whereas, for DNA ligases, not necessarily thermostable, they followed the proteome ones. Analysis of codon usage revealed that hyperthermophiles preferred AGR codons for Arg in detriment of CGN codons, which were preferred by mesophiles.

Conclusions: The results suggested that the E+K/Q+H ratio may provide a useful mark for distinguishing hyperthermophilic, thermophilic and mesophilic prokaryotes and that the high percent of the amino acid couple E+K, consistently associated to the low percent of the pair Q+H, could contribute to protein thermostability. Second, the preference for AGR codons for Arg was a signature of all hyperthermophilics so far analyzed.

Background

Temperatures lower than 50°C are common in different habitats of the Earth and most organisms are mesophiles with an optimal growth temperature (OGT) at 24-40°C. Life at temperatures higher than 55-60°C can be found associated to low pH, high salinity or high-pressure environments, including members of Archaea and Bacteria. Evidence of eukaryotic life above 60°C is scarce (1). Prokaryotes data grouped moderate thermophiles with OGT in the range of 50°C to 70°C and hyperthermophiles with OGT above 80°C. The hyperthermophilic members of Archaea and Bacteria, which can grow at 80°C-105°C, are unable to propagate in temperatures lower than their OGT (2).

There are not many clues as to how life can thrive in extreme environments. In general the biochemistry of hyperthermophiles proteins is very similar to that of mesophiles. When protein sequences and three-dimensional structures are compared, there are no significant differences among molecules: the sequences of homologous proteins from hyperthermophiles and mesophiles are 40 to 80% similar; their three-dimensional structures are superposable; and they have the same catalytic mechanisms (3). Nevertheless, most enzymes from hyperthermophiles show an optimum catalytic activity above 100°C. Stability at high temperatures appears to be the result of very subtle synergistic and cooperative intra and inter molecular interactions, or from extrinsic protectants (4,5). Some findings relevant for explaining protein thermostability were: i) increase in the number of hydrogen bonds, and extension in the inter subunit ion-pairs networks (6,7); ii) increase in the number of charged amino acids (8,9); iii) decreased length of superficial loops and an increase in protein compactness (10,11). On the other hand, the fully functional and stable folded state of hyperthermophiles might require specific chaperones to be reached (3), as the chaperonin systems, which are included in the molecular chaperone family (12) of proteins.

In the present work, the proteome data from hyperthermophiles (HT), moderate thermophiles (T) and mesophiles (M) was compared, looking for characteristics which could be related to thermal adaptation, allowing distinguishing high and low-temperature organisms and proteins. So far two parameters were utilized: a) amino acid composition and amino acids couplets in each proteome, b) codon usage in the whole genome. The study was complemented using the same parameters in analyzes of two types of proteins: chaperonins and DNA ligases. These proteins were elected

for analysis considering their thermal stability and their presence in all organisms. Chaperonins are potentially thermostable in all OGT groups and amongst Hsps they are unique in being present in all three domains of life (13). On the other hand, DNA ligases are not necessarily thermostable in M but they are in HT and T. The results showed that high (E+K)/(Q+H) values were a characteristic of hyperthermophilic organisms and could be related to protein thermostability. Moreover, AGR codon bias for arginine was a signature for thermophiles and hyperthermophiles.

Results and Discussion

1. Proteome

The proteomes of 18 M, 4 T and 6 HT were analyzed for amino acid composition. In Table 1, the ratio (HT, T)/M corresponds to the average proportion of each amino acid in all HT and T (>60°C) organisms divided by their average proportion in all M (24-50°C). Only the amino acids with extreme ratios (> 1.30 and < 0.80) are listed. Preferred amino acids at high temperature were Glu, Lys, Tyr, Ile, and at mesophiles were Gln, His, Ala, Cys. Among the amino acids with extreme ratios we could highlight two couplets of preferred and avoid: E and Q, and K and H. These four amino acids are a homogenous group of high polar requirements, > 8,40 (14); are preferentially found in α -helices (15, 16); its second codon position is A; and, at least for Gln and His, the ranges in HT and T are not extensively superposed to the ranges in M.

The idea that the physico-chemical properties of the most frequent amino acids are reflected in the protein secondary structures constitutes the basic information for prediction methods. An analysis of prokaryotic and human proteins shows that Glu, Lys, Asp, Asn, Gln and His prefer α -helix and aperiodic structures, and that Ile, Leu, Val and Phe prefer β -sheets (16). Moreover, the measure of the water accessible surfaces (WAS) of amino acids calculated for 131 proteins from Bacteria and Archaea, showed that the WAS percentage of charged amino acids increased in hyperthermophiles, mainly due to Lys and Glu, and this increase was accompanied by a decrease in Gln, Ala and His (9). In several comparisons of organisms with different

life styles, the proteins sequences of hyperthermophiles included more charged (Glu, Arg, Lys) and fewer uncharged polar residues (Ser, Thr, Asn, Gln, His, Cys), increased residue hydrophobicity (Ile, Val), and increased residue volume (Tyr) (8,9,3).

Our comparison of 28 proteomes, from the tree domains of life, did not show correlation of Arg, Ser or Asn with temperature changes, but did show that a good indicator of lifestyle was the (HT, T)/M ratios, better than the mere abundance of amino acids (Table 1). So, considering the two couplets of preferred and avoided, the ratios E/Q and K/H were calculated for each one of the 28 proteomes. Figure 1 shows a clear correlation between lifestyle and (E+K)/(Q+H) ratios.

The ratio (E+K)/(Q+H) was always greater than 4.5 in HT, between 3.2 and 4.6 in T and lower than 2.5 in M showing that as the OGT decreases the proportion of E and K also decreases whereas the proportion of Q and H increases (Table 3). So, the (E+K)/(Q+H) ratios are an excellent indicator for lifestyles since the correlation between OGTs and preference for some amino acids couplets permits to discriminate HT, T and M across the three domains of life. Such results corroborate the proposition that adaptation to high temperatures might involve different mechanisms in HT and T (11).

Figure 2 shows that the differences observed in HT, T and M in relation to the relative concentrations of E, K, Q and H could not be attributed exclusively to the G+C content of the organisms. Codons for E and K are respectively GAR and AAR with a predominance of A over G whereas codons for Q and H are respectively CAR and CAY (no predominance). Although the higher E+K values were observed in hyperthermophiles, which show G+C content inferior to 50%, at similar G+C contents the organisms showed different E+K values. In M the G+C content runs from 35% to 67% (Figure 2, Table 3).

2. Chaperonins and DNA ligases analysis

(E+K)/(Q+H) ratio was calculated for the chaperonins GroEl (Bacteria), THS (Archaea) and TCP1 (Eukarya). As these proteins are thermostable they could exhibit a profile similar to that of HT and T proteomes even when they originate from mesophiles. Figure 3A shows that the (E+K)/(Q+H) ratios for chaperonins, in prokaryotes, are usually in the limits of thermophilic and hyperthermophilic proteome

values, that is above 2.5. The (E+K)/(Q+H) values, for chaperonin, ranged from 3.32 to 7.64 among prokaryotes (Table 3). The only exceptions were *V. cholerae* (2.1) and *Halobacterium* (1.84). Amongst eukaryotes the (E+K)/(Q+H) values for chaperonins were above 2.5, except for *A. thaliana* (Table 3). A possible relationship between (E+K)/(Q+H) ratio and protein thermostability, at least in prokaryotes, could be envisaged.

If the relative proportions of E, K, H and Q were related to protein stability we should expect that in mesophiles the (E+K)/(Q+H) values for DNA ligases were similar to that of their respective proteomes. Such profile would be strikingly different from that for chaperonins. Figure 3B shows that the (E+K)/(Q+H) values for DNA ligase follow the ones of the 28 proteomes. In HT and T, the values ranged from 3.8 to 9.4, and in M prokaryotes the values were, usually, lower than 2.5, except for the Eukarya group (Table 3). These results corroborate the idea that the relative concentrations of these amino acids couplets, in the composition of the peptides, could be associated to their thermostability.

3. Codon bias for Arginine

As the preference for one synonymous codon over another in organisms is a long known phenomenon, we looked for codon bias associated to temperature lifestyles. Striking differences, related to Arg (R) codons, were found between HT and T on one hand and M on another hand. Figure 4A shows that HT and T highly preferred AGR at expenses of CGN codons, which are preferred by M. One exception in prokaryotes was *Mycoplasma genitalium* with 58% of AGR and 48% of GCN and in eukaryotes the exceptions were *S. cerevisiae* and *C. elegans* with 69.5% and 55.8% of AGR, respectively (Table 4). Nevertheless the % of AGR in these organisms was never so high as in hyperthermophiles and thermophiles (Table 4).

Figure 4B shows that the % of Arg, in the proteome of these organisms, increases with the G+C content, and no correlation was found between the content of Arg and OTG. The preference for CGN codons cannot be attributed to the G+C content since very different values of CGN were found for similar values of G+C. Compare, for instance, the G+C content in *A. aeolicus* (43.2%), *A. fulgidus* (43.2%), *T. maritima* (46.1%), *Sulfolobus sp* (35.7%) *B. subtilis* (43.5%), *V. cholerae* (47.6),

Synechocystis (47.6) and *H. influenza* (38%) with their values for CGN usage, which are 8.0%, 9.4%, 18.0%, 8.9%, 63.0%, 91.5%, 80.0% and 91.0%, respectively (Table 4). The data reveal a superimposed correlation: at similar G+C contents the preference for the synonyms codons is determined by lifestyles, hot or non-hot.

Recently it was suggested that GC content drives codon usage and that the relationship between G+C content and codon and amino acid usage is ahistorical (17). Nevertheless, a broader connection with selective pressure based on the physicochemical properties of the amino acids might be considered. A transitional mutation in the second position of AGR converts Arg to Lys (AAR), which can be converted to Glu by a transitional mutation in the first position (GAR). All these amino acids have high polar requirements, 9.1, 10.0 and 12.5, respectively and, both, Lys and Glu are preferred amino acids in hyperthermophilic organisms. Also, a transitional mutation in the second position of CGY codons converts Arg to His (CAY), which can be converted to Gln (CAR) by a transversional mutation in the third position. Both, His and Gln, with similar polar requirements (8.6 and 8.4), are preferred amino acid in mesophilic organisms. The codons AGR (Arg), AAR (Lys) and GAR (Glu), are constituted of dinucleotides RR whereas the codons CGN (Arg), CAY (His) and CAR (Gln) are constituted mainly of YR and RY dinucleotides. A high correlation (higher than those expected from a random combination of mononucleotides), between RR or YY and hyperthermophiles, has been found (18). All the connections discussed above could help to explain the adaptation of organisms to high temperatures including preference for AGR and the high E+K/Q+H ratio in hyperthermophiles.

Conclusions

The results showed that *i*) the E+K/Q+H values for proteome discriminated organisms according their OGT: >4.5 for HT; 3.2 to 4.6 for T; <2.5 for ME; *ii*) the high percent of E and K associated to the low % of H and Q could be related to protein thermostability; *iii*) the AGR codon bias for Arg can be used as a signature for HT and T; *iv*) the E+K/Q+H ratio and codon bias for Arg are not apparently related to phylogeny. Members HT of the Bacteria domain show the same values as the HT members of the Archaea domain; the values for T organisms are related to their lifestyle (intermediate temperature) and not to their domain (Archaea) and the values for M are similar in Eukarya, Bacteria and Archaea.

Materials and Methods

Organisms:

Data from the complete genome of 28 organisms were analyzed: 06 hyperthermophiles (1-6), 04 moderate thermophiles (7-10) and 18 mesophiles (11-28), listed in Table 2.

Loci for chaperonins: 1. *Pyrococcus abyssi* (PAB2341), 2. *Pyrococcus horikoshii* (PH1256), 3. *Aquifex aeolicus* (AQ2200), 4. *Methanococcus jannashii* (MJ0999), 5. *Archeoglobus fulgidus* (AF2238), 6. *Thermotoga maritima* (TM0506), 7. *Methanobacterium thermoautotrophicum* (MTH0218), 8. *Thermoplasma Volcanium* (TVN0507), 9. *Thermus acidophilum* (Ta0980), 10. *Sulpholobus* (SSO0862), 11. *Escherichia coli* (B4143), 12. *Vibrio cholerae* (VC2664), 13. *Deinococcus radiodurans* (DR0607), 14. *Synechocystis* (SLR2076), 15. *Bacillus subtilis* (YKKC), 16. *Mycobacterium tuberculosis* (MT0456), 17. *Neisseria meningitidis* (NMB1972), 18. *Mycoplasma genitalium* (MG0392), 19. *Treponema pallidum* (TP0030), 20. *Pseudomonas aeruginosa* (PA4385), 21. *Haemophilus influenza* (HI0543), 22. *Xylella fastidiosa* (XF0615), 23. *Halobacterium sp* (CCTB), 24. *Saccharomyces cerevisiae* (YDR212W), 25. *Arabidopsis thaliana* (AT1g24510), 26. *Drosophila melanogaster* (gi7291004), 27. *Caenorhabditis elegans* (t21b10.7), 28. *Homo sapiens* (TCPB).

Loci for DNA ligases: 1. *Pyrococcus abyssi* (PAB2002), 2. *Pyrococcus horikoshii* (PH1622), 3. *Aquifex aeolicus* (AQ0633), 4. *Methanococcus jannashii* (MJ0171), 5. *Archeoglobus fulgidus* (AF0623), 6. *Thermotoga maritima* (TM0100), 7. *Methanobacterium thermoautotrophicum* (MTH1580), 8. *Thermoplasma Volcanium* (TVN1237), 9. *Thermus acidophilum* (Ta1148), 10. *Sulpholobus*(SSO0189), 11. *Escherichia coli* (B2411), 12. *Vibrio cholerae* (VC0971), 13. *Deinococcus radiodurans* (DR2069), 14. *Synechocystis* (gi1651660), 15. *Bacillus subtilis* (YERG), 16. *Mycobacterium tuberculosis* (MT3148), 17. *Neisseria meningitidis* (NMB0666), 18. *Mycoplasma genitalium* (MG254), 19. *Treponema pallidum* (TP0634), 20. *Pseudomonas aeruginosa* (PA1529), 21. *Haemophilus influenza* (HI1100), 22. *Xylella fastidiosa* (XI2556), 23. *Halobacterium sp* (LIG), 24. *Saccharomyces cerevisiae* (YOR005C). 25.

Arabidopsis thaliana (AT1g49250), 26. *Drosophila melanogaster* (gi7299616), 27. *Caenorhabditis elegans* (c07h6.1), 28. *Homo sapiens* (LIG4).

Data base

The 28 proteomes used in this work are listed in Table 2 and the data for the analysis of preferred amino for each proteome were obtained from the Institute of Genomic Research, TIGR (www.tigr.org) and from the Protein Extraction, Description, and Analysis Tool, PEDANT (<http://pedant.mips.biochem.mpg.de/>). The data for the analysis of codon usage for each proteome were obtained from the Kazusa DNA Research Institute (www.kazusa.or.jp) via Countcodon program.

E+K/H+Q ratio

E+K/H+Q ratio was calculated as follows: the percent of Glu plus the percent of Lys in a proteome (or ORF) was divided by the percent of His plus the percent of Gln in the same proteome (or ORF). The ORFs were deduced from the loci cited above.

Acknowledgements

We thank Dr. Romeu C. Guimarães for his critical reading of the manuscript; the Universidade Federal da Paraíba for enabling us to conduct these studies; CAPES that financed the postgraduate program on Genetics at UFPB and all the free sites of the web on genomics and proteomics.

References

1. Sicot FX, Mesnage M, Masselot M, Exposito JY, Garrone R, Deutsch J, Gaill F: **Molecular adaptation to an extreme environment: Origin of the stability of the Pompeii worm collagen.** *J Mol Biol* 2000, **302**:811-820.
2. Madigan MT, Martinko JM, Parker J: **Brock biology of microorganisms** 8th ed. Prentice-Hall Inc Upper Saddle River NJ, 1997.
3. Vieille C, Zeikus G: **Hyperthermophilic enzymes: sources, uses and molecular mechanisms for thermostability.** *Microb. Mol. Biol. Rev.* 2001, **65**:1-43.
4. Jaenicke R: **Do ultrastable proteins from hyperthermophiles have high or low conformational rigidity?** *Proc. Natl. Acad. Sci.* 2000, **97**:2962-2964.
5. Zierenberg R A, Adams MWW, Arp AJ: **Life in extreme environments: Hydrothermal vents.** *Proc Natl Acad Sci* 2000, **97**:12961- 12962.

6. Vogt G, Woell S, Argos P: **Protein thermal stability hydrogen bonds and ion pairs.** *J Mol Biol* 1997, **269**:631-643.
7. Ventriani C, Maeder DL, Tolliday N, Yip KSP, Stillman TJ, Britton K L, Rice D W, Klump HH, Robb FT: **Protein thermostability above 100⁰C: A key role for ionic interactions.** *Proc Natl Acad Sci* 1998, **95**:12300-12305.
8. Haney PJ, Badger HJ, Buldak GL, Reich CI, Woese CR, Olsen GJ: **Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species.** *Proc Natl Acad Sci* 1999, **96**:3578- 3583.
9. Cambillau C, Claverie J M: **Structural and genomic correlates of hyperthermostability.** *J Biol Chem* 2000, **275**:32383-32386.
10. Thompson MJ, Eisenberg D: **Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability.** *J Mol Biol* 1999, **290**:595-604.
11. Szilagyi A, Zavodszky P: **Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey.** *Structure Fold Des* 2000, **8**: 493-504.
12. Ellis RJ, van der Vies SM: **Molecular chaperones.** *Annu Rev Biochem* 1991, **60**: 321-347.
13. Macario AJL, Lange M, Ahring BK, Macario EC: **Stress genes and protein in the Archaea.** *Microbiol. Mol. Biol. Rev.* 1999, **63**:923-967.
14. Woese AR, Olsen GJ, Ibba M, Söll D: **Aminoacyl-tRNA synthetases, the genetic code and the evolutionary process.** *Microbiol Mol Biol Rev* 2000, **64**:202-236.
15. Chou PY, Fasman GD: **Conformational parameters for amino acids in helical, β -sheet, and random coil regions calculated from proteins.** *Biochemistry* 1974, **13**:211-222.
16. Chiusano ML, Alvarez-Valin F, Di Giulio M, Donofrio G, Ammirato G, Colonna G, Bernardi G: **Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code .** *Gene* 2000, **261**:63-69.
17. Knight RD, Freeland SJ, Landweber LF: **A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes.**
<http://genomebiology.com/2001/2/4/research/0010.1>.
18. Kawashima T, Amano N, Koike H, Makino S-I, Higuchi S, Kawashima-Ohya Y, Watanabe K, Yamazaki M, Kanehori K, Kawamoto T, *et al*: **Archaeal adaptation to higher temperatures, revealed by genomic sequences of *Thermoplasma volcanium*.** *Proc Natl Acad Sci* 2000, **97**:14257-62.

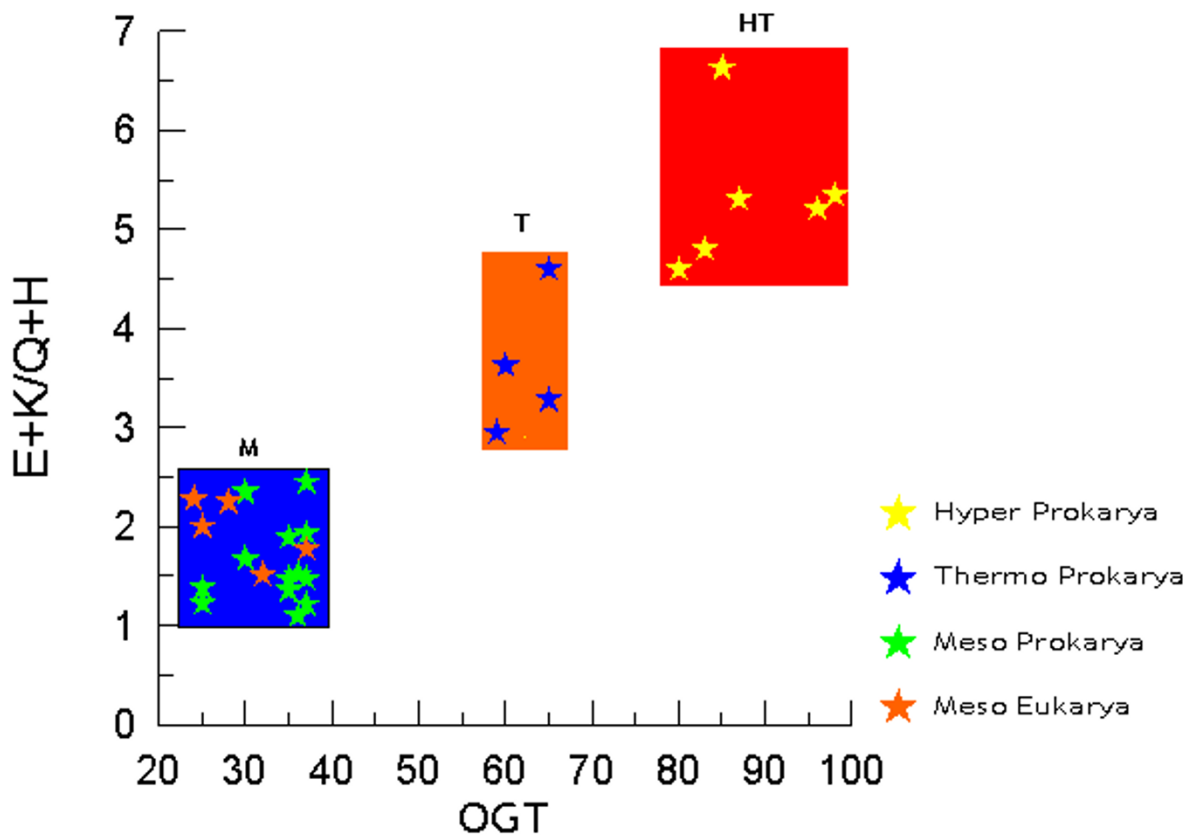


Figure 1.

E+K/Q+H values for the proteomes of 28 organisms versus optimal growth temperature (OGT). E+K/Q+H ratio = %E (Glu) + %K (Lys) divided by %H (His) + %Q (Gln) in the same proteome. Amongst HT, OGT ranges from 80-100°C, in T from 60-65°C and in M from 24-37°C. E+K/Q+H ratio discriminated hyperthermophilic, thermophilic and mesophilic lifestyles. The identification of the organisms is in Table 2.

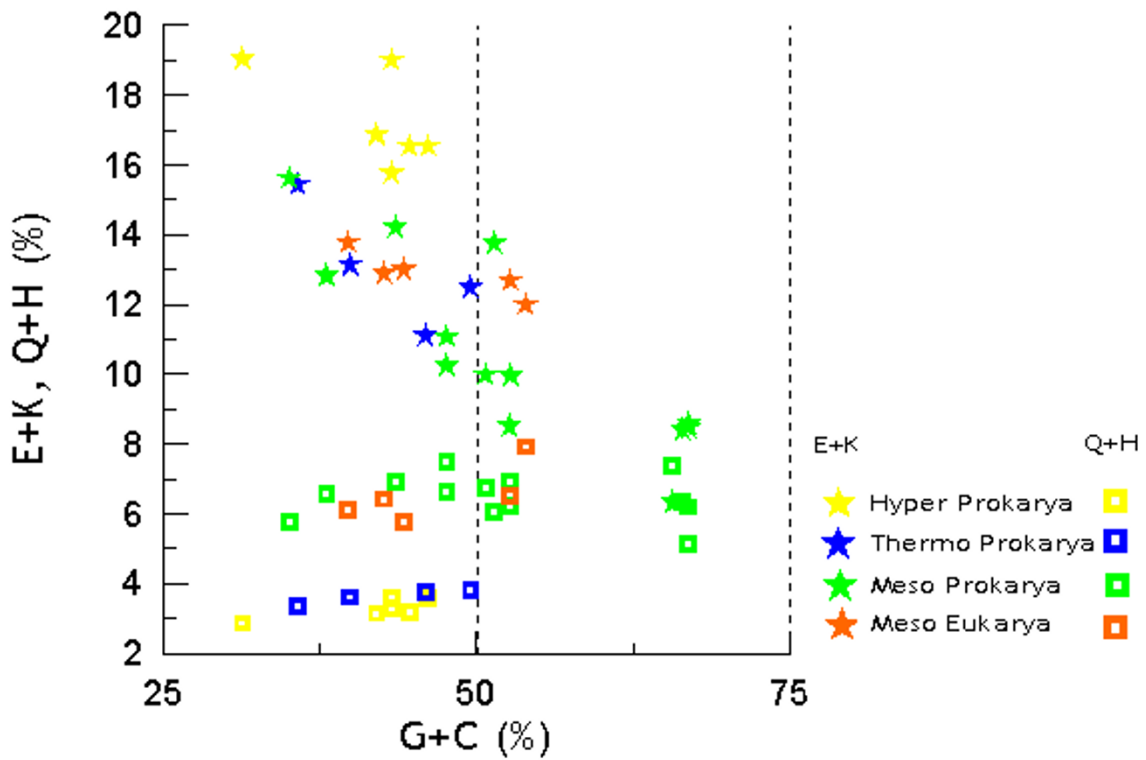


Figure 2.

Preferred amino acids couplets and G+C contents in HT, T and M. The percent of His (H) plus the percent of Gln (Q) and the percent of Glu (E) plus the percent of Lys (K), in each proteome, were plotted against G+C contents. The G+C content in HT and T was below 50%. Preference for E+K was not correlated to G+C content. The identification of the organisms is in Table 2.

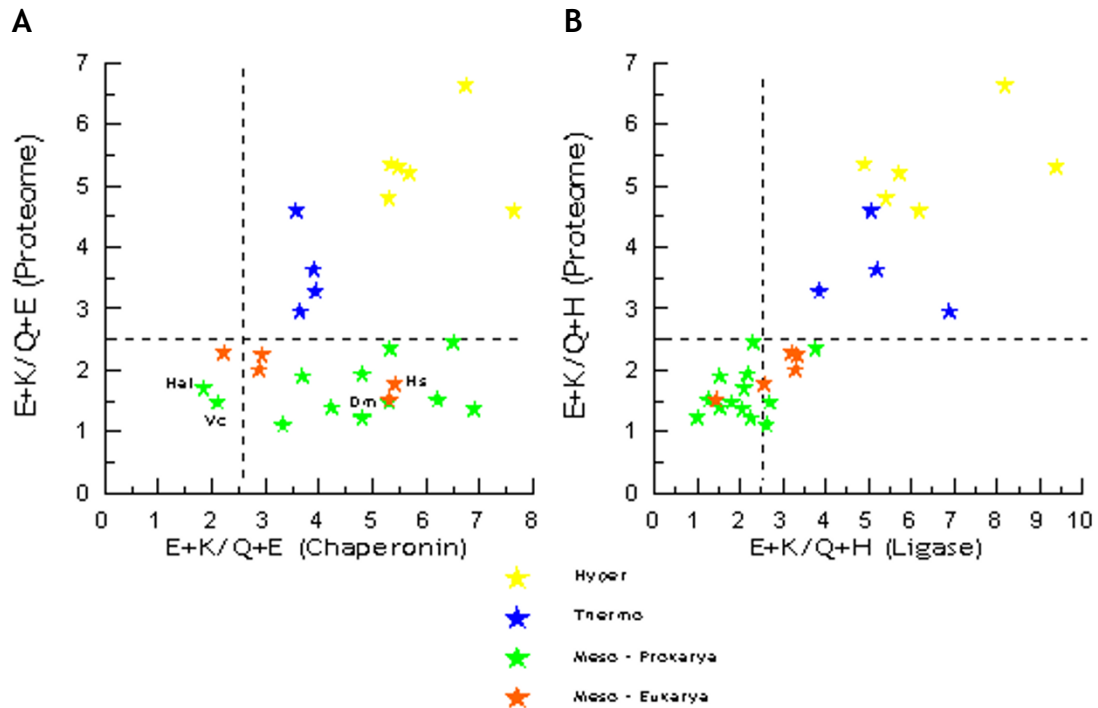


Figure 3.

The $(E+K)/(Q+H)$ values for chaperonins (3A) and ligases (3B) versus the $(E+K)/(Q+H)$ values for the proteomes of HT, T and M. The $(E+K)/(Q+H)$ ratios for the chaperonins and for the DNA ligases, described in Material and Methods, were calculated in each one of the 28 organisms. In mesophilic prokaryotes the ratios for chaperonins followed the hyperthermophilic values (> 2.5), except for *Vibrio cholerae* (Vc) and *Halobacterium sp* (Hal). In mesophilic eukaryotes, only *Homo sapiens* (Hs) and *Drosophila melanogaster* (Dm) followed the rule. The ratios for DNA ligases always followed the proteome values.

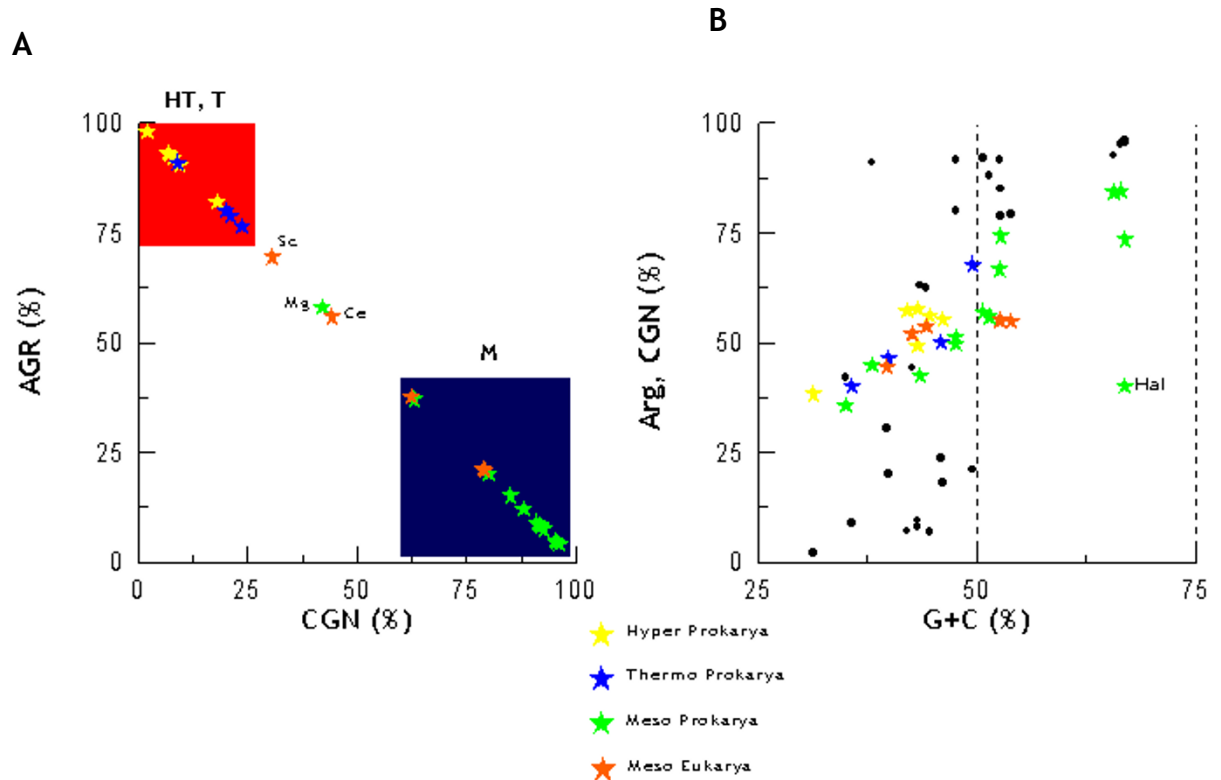


Figure 4.

Preferred codons for arginine (AGR versus CGN), according to life styles (A) and arginine utilization according to G+C content (B). A) HT and T preferred codons AGR whereas codons CGN were preferred by M. The mesophiles *Mycoplasma genitalium* (Mg) and *Caenorhabditis elegans* (Ce) showed no preference whereas *Saccharomyces cerevisiae* (Sc) preferred AGR codons. B) The percent of Arg in the proteome increased with the G+C content, except for the halophilic Archaea, *Halobacterium sp* (Hal). The preference for CGN codons (black circles) cannot be attributed to the G+C content.

Table 1.

Preferred and avoided amino acids from proteomes of 10 hyperthermophiles (HT) and thermophiles (T), and 18 mesophiles (M).

| Amino acids | Secondary structures* | Polar requirement | HT and T | | M | | Ratio (HT, T)/M |
|-------------|-----------------------|-------------------|--------------------|--------------------|--------------------|---------|-----------------|
| | | | Prokarya | Average (range) | Prokarya | Eukarya | |
| Gln (Q) | α -helix | 8.60 | 1.88 (1.44 - 2.08) | 4.19 (3.25 - 5.5) | 4.28 (3.97 - 5.24) | 0.44 | |
| His (H) | α -helix | 8.40 | 1.53 (1.43 - 1.9) | 2.29 (1.83 - 2.74) | 2.4 (2.15 - 2.68) | 0.65 | |
| Ala (A) | α -helix | 7.00 | 6.38 (5.44 - 7.84) | 9.83 (6 - 13) | 7.64 (6.19 - 11.1) | 0.73 | |
| Cys (C) | β -strand | 4.8 | 0.81 (0.56 - 1.27) | 1.09 (0.67 - 1.91) | 1.84 (1.27 - 2.21) | 0.62 | |
| Glu (E) | α -helix | 12.50 | 8.11 (5.53 - 9.6) | 5.8 (4.4 - 7.08) | 6.6 (6.39 - 6.99) | 1.34 | |
| Lys (K) | α -helix | 10.10 | 7.5 (4.5 - 10.4) | 4.56 (2.61 - 10) | 6.28 (5.6 - 7.29) | 1.49 | |
| Tyr (Y) | β -strand | 5.40 | 4.08 (3.2 - 4.82) | 2.74 (1.9 - 3.48) | 3.01 (2.76 - 3.35) | 1.45 | |
| Ile (I) | β -strand | 4.90 | 8.42 (7.18 - 10.5) | 5.5 (3.28 - 8.64) | 5.46 (4.43 - 6.51) | 1.52 | |

*Taken from 14; *Taken from 16; data on HT, T, M in %; only ratios >1.3 and <0.8 are shown.

Table 2.

Hyperthermophiles, Thermophiles and Mesophiles from Archaea (A), Bacteria (B) e Eukarya (E).

| Organisms | Domains | OGT (°C) | Organisms | Domains | OGT (°C) |
|---|---------|----------|---------------------------------------|---------|----------|
| 01. <i>Pyrococcus abyssi</i> | A | 96-98 | 2. <i>Pyrococcus horikoshii</i> | A | 95-98 |
| 03. <i>Aquifex aeolicus</i> | B | 85 | 4. <i>Methanococcus jannashii</i> | A | 80-85 |
| 05. <i>Archeoglobus fulgidus</i> | A | 83 | 6. <i>Thermotoga maritima</i> | B | 80 |
| 07. <i>Methanobacterium thermoautotrophicum</i> | A | 65 | 8. <i>Thermoplasma volcanium</i> | A | 60 |
| 09. <i>Thermus acidophilum</i> | A | 65 | 10. <i>Sulfolobus sp</i> | A | 65 |
| 11. <i>Escherichia coli</i> | B | 35-37 | 12. <i>Vibrio cholerae</i> | B | 35-37 |
| 13. <i>Deinococcus radiodurans</i> | B | 25-37 | 14. <i>Mycobacterium tuberculosis</i> | B | 25 |
| 15. <i>Neisseria meningitidis</i> | B | 30 | 16. <i>Mycoplasma genitalium</i> | B | 35-37 |
| 17. <i>Treponema pallidum</i> | B | 25-37 | 18. <i>Pseudomonas aeruginosa</i> | B | 35-37 |
| 19. <i>Haemophilus influenza</i> | B | 35-37 | 20. <i>Bacillus subtilis</i> | B | 35-37 |
| 21. <i>Synechocystis</i> | B | 35-37 | 22. <i>Xylella fastidiosa</i> | B | 25-30 |
| 23. <i>Halobacterium sp</i> | A | 25-37 | 24. <i>Saccharomyces cerevisiae</i> | E | 25-30 |
| 25. <i>Caenorhabditis elegans</i> | E | 20-24 | 26. <i>Arabidopsis thaliana</i> | E | 30-37 |
| 27. <i>Drosophila melanogaster</i> | E | 25 | 28. <i>Homo sapiens</i> | E | 37 |

Table 3.

E+K, and Q+H contents (%) in the 28 proteomes and E+K/Q+H ratios for the Proteome, Chaperonin and Ligase in each organism.

| Strains | Proteome | | | | Chaperonin | Ligase |
|-------------------------------|----------|--------|---------|--------|------------|--------|
| | E+K/Q+H | E+K(%) | Q+H (%) | G+C(%) | E+K/Q+H | |
| Hiper - Procarya | | | | | | |
| <i>P. abyssi</i> | 5.21 | 16.53 | 3.17 | 44.6 | 5.68 | 5.72 |
| <i>P. horikoshii</i> | 5.35 | 16.87 | 3.15 | 42.0 | 5.33 | 4.9 |
| <i>A. aeolicus</i> | 5.31 | 19.01 | 3.58 | 43.2 | 5.47 | 9.39 |
| <i>M. jannashii</i> | 6.63 | 19.04 | 2.87 | 31.3 | 6.73 | 8.18 |
| <i>A. fulgidus</i> | 4.8 | 15.76 | 3.28 | 43.2 | 5.3 | 5.4 |
| <i>T. matitima</i> | 4.6 | 16.53 | 3.59 | 46.1 | 7.64 | 6.18 |
| Termo - Procarya | | | | | | |
| <i>M. thermoautotrophicum</i> | 3.28 | 12.5 | 3.81 | 49.5 | 3.93 | 3.85 |
| <i>T. volcanium</i> | 3.63 | 13.13 | 3.61 | 39.9 | 3.9 | 5.19 |
| <i>T. acidophilum</i> | 2.95 | 11.1 | 3.76 | 45.9 | 3.64 | 6.89 |
| <i>Sulfolobus</i> | 4.6 | 15.46 | 3.36 | 35.7 | 3.57 | 5.06 |
| Meso - Procarya | | | | | | |
| <i>E. coli</i> | 1.47 | 9.98 | 6.75 | 50.7 | 5.3 | 2.7 |
| <i>V. cholerae</i> | 1.47 | 11.08 | 7.49 | 47.6 | 2.1 | 1.8 |
| <i>D. radiodurans</i> | 1.36 | 8.47 | 6.19 | 66.9 | 6.89 | 2.05 |
| <i>M. tuberculosis</i> | 1.1 | 6.32 | 5.75 | 65.6 | 3.32 | 2.62 |
| <i>N. meningitidis</i> | 1.89 | 11.75 | 6.22 | 51.4 | 3.69 | 1.53 |
| <i>M. genitalium</i> | 2.45 | 15.61 | 6.36 | 35 | 6.5 | 2.31 |
| <i>T. pallidum</i> | 1.51 | 9.95 | 6.57 | 52.7 | 6.21 | 1.26 |
| <i>P. aeruginosa</i> | 1.21 | 8.4 | 6.9 | 66.4 | 4.8 | 2.25 |
| <i>H. influenza</i> | 1.93 | 12.82 | 6.64 | 38 | 4.8 | 2.18 |
| <i>B. subtilis</i> | 2.35 | 14.22 | 6.05 | 43.5 | 5.32 | 3.76 |
| <i>Synechocystis</i> | 1.38 | 10.24 | 7.39 | 47.6 | 4.23 | 1.53 |
| <i>X. fastidiosa</i> | 1.23 | 8.51 | 6.92 | 52.6 | 4.8 | 1.00 |
| <i>Halobacterium sp</i> | 1.67 | 8.58 | 5.15 | 66.88 | 1.84 | 2.19 |
| Meso - Eucarya | | | | | | |
| <i>S. cerevisiae</i> | 2.25 | 13.79 | 6.12 | 39.7 | 2.93 | 3.33 |
| <i>C. elegans</i> | 2.00 | 12.88 | 6.42 | 42.6 | 2.88 | 3.3 |
| <i>A. thaliana</i> | 2.28 | 13.01 | 5.74 | 44.2 | 2.22 | 3.2 |
| <i>D. melanogaster</i> | 1.51 | 11.99 | 7.92 | 53.9 | 5.31 | 1.44 |
| <i>H. sapiens</i> | 1.77 | 12.68 | 7.17 | 52.66 | 5.41 | 2.57 |

Table 4.

Arginine and codon usage in each one of the 28 proteomes .

| Strains | CGN (%) | AGR (%) | Arg (%) | G+C(%) | OGT (°C) |
|-------------------------|---------|---------|---------|--------|----------|
| Hiper - Procarya | | | | | |
| <i>P. abyssi</i> | 7.0 | 93.0 | 5.73 | 44.6 | 96 |
| <i>P. horikoshii</i> | 6.8 | 93.2 | 5.60 | 42.0 | 98 |

| | | | | | |
|-------------------------------|-------|-------|------|-------|----|
| <i>A. aeolicus</i> | 8.0 | 92.0 | 4.92 | 43.2 | 87 |
| <i>M. jannashii</i> | 2.0 | 98.0 | 3.82 | 31.3 | 85 |
| <i>A. fulgidus</i> | 9.4 | 90.6 | 5.76 | 43.2 | 83 |
| <i>T. matitima</i> | 18.0 | 82.0 | 5.51 | 46.1 | 80 |
| Termo - Procarya | | | | | |
| <i>M. thermoautotrophicum</i> | 21.0 | 79.0 | 6.78 | 49.5 | 65 |
| <i>T. volcanium</i> | 20.0 | 80.0 | 4.62 | 39.9 | 60 |
| <i>T. acidophilum</i> | 23.62 | 76.38 | 5.0 | 45.9 | 59 |
| <i>Sulfolobus</i> | 8.92 | 91.02 | 4.0 | 35.7 | 65 |
| Meso - Procarya | | | | | |
| <i>E. coli</i> | 92.0 | 8.0 | 5.7 | 50.7 | 37 |
| <i>V. cholerae</i> | 91.5 | 8.5 | 4.97 | 47.6 | 35 |
| <i>D. radiodurans</i> | 95.5 | 4.5 | 7.35 | 66.9 | 35 |
| <i>M. tuberculosis</i> | 92.6 | 7.4 | 8.43 | 65.6 | 36 |
| <i>N. meningitidis</i> | 88.0 | 12.0 | 5.58 | 51.4 | 35 |
| <i>M. genitalium</i> | 42.0 | 58.0 | 3.57 | 35 | 37 |
| <i>T. pallidum</i> | 85.0 | 15.0 | 7.43 | 52.7 | 36 |
| <i>P. aeruginosa</i> | 95.2 | 4.8 | 8.46 | 66.4 | 37 |
| <i>H. influenza</i> | 91.0 | 9.0 | 4.48 | 38 | 37 |
| <i>B. subtilis</i> | 63.0 | 37.0 | 4.24 | 43.5 | 30 |
| <i>Synechocystis</i> | 80.0 | 20.0 | 5.11 | 47.6 | 25 |
| <i>X. fastidiosa</i> | 91.7 | 8.3 | 6.67 | 52.6 | 25 |
| <i>Halobacterium sp</i> | 96.1 | 3.9 | 4.0 | 66.88 | 30 |
| Meso - Eucarya | | | | | |
| <i>S. cerevisiae</i> | 30.5 | 69.5 | 4.44 | 39.7 | 28 |
| <i>C. elegans</i> | 44.2 | 55.8 | 5.19 | 42.6 | 25 |
| <i>^a thaliana</i> | 62.3 | 37.7 | 5.35 | 44.2 | 24 |
| <i>D. melanogaster</i> | 79.1 | 20.9 | 5.49 | 53.9 | 32 |
| <i>H. sapiens</i> | 78.75 | 21.25 | 5.5 | 52.66 | 37 |