Genome-scale analysis of protein function distribution of extremophilic archaea using COG database

Roberto Ubidia-Incio
Universidad Norbert Wiener
Sofía CA. Rodríguez
Universidad Nacional Mayor de San Marcos
Karen L Orozco
Universidad Nacional Mayor de San Marcos
Víctor Rojas-Zumaran
Hospital Nacional Docente Madre Niño San Bartolome
Jeel Moya-Salazar (jeel.moya@uwiener.edu.pe)
Universidad Norbert Wiener

Article

Keywords: Comparative genomics, extremophiles, COGs, life beyond Earth, bioinformatics

Posted Date: April 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1552117/v1

Abstract

Extremophilic archaea have a variety of adaptations that they have developed in order to cope with their environmental conditions. In this study, we analyze the distribution of protein-coding genes of the archaea listed in NCBI's COG database in order to observe changes in the way these genes have been distributed among functional categories for the different types of extremophilic groups. We found that the category T (Signal transduction mechanisms) presents variations that are more related to the extremophilic adaptations than other categories, and that some metabolic categories seem to be related indirectly to the extreme environments in which these groups thrive. These findings could be useful in studying how the distribution of protein functions affects life in unexplored extreme environments on Earth or elsewhere, and when designing modified organisms that can be used for human exploration support in these environments.

Introduction

Life on Earth appeared approximately 3.5 Gy ago [1], at that point conditions on Earth were very different from those of today [2] and the first living beings possibly appeared in an anoxygenic thermophilic environment [3]. Since then, life has adapted to many different types of climates, some of them so harsh that until recently it was thought that life couldn't thrive in these conditions and they have led to the hypothesis that life might thrive on similar environments beyond Earth [4], and are used as models to understand how life might be if we found it in the different harsh environments we can find in the Solar System or elsewhere.

There have been different studies on the special strategies extremophiles utilize in order to adapt to their environments, many of these involving mainly the cell membrane and the cell wall composition as hyperthermophiles, halophiles and psychrophiles [5]. In other cases, they have multiple copies of genes involving repair mechanisms as radiotolerant bacteria[6] or have evolved proteins that are active at these conditions as halophiles [7], psychrophiles [8], and hyperthermophiles [9]. Each type of extremophile usually has similar adaptations regardless of their taxonomic distance that may have been gained through processes such as horizontal gene transfer or convergent evolution [10,11,12,13].

These changes are usually involved with extra signaling pathways working on the detection of environmental changes that when activated induce the changes previously mentioned [14,15], adding additional proteins or copies of the genome making it easier to repair or avoid any important damage [16], or having different groups or proteins that activate at different environmental conditions [17]. We can visualize these changes at the genomic level by looking at the number of genes distributed for the different functions they accomplish, and comparing groups of organisms with different adaptation [12]. In this sense comparative genomics provide us with tools that help us to understand how information is distributed over the genome. This information can easily be obtained from different databases available on the internet, and some of them have already organized the information from organisms with

sequenced genomes, such as COGs database [18], which has divided the genomes of every sequenced bacteria and archaea in functional categories including genetic expression mechanisms categories (J, A, K, L, B), metabolic categories (C, G, E, F, H, I, P, O), and other non-metabolic categories (D, Y, V, T, M, N, Z, W, U, O, X) [16]. A complete list of these categories can be seen in Table 1.

The aim of this study was to visualize and compare the distribution of protein-coding genes of different groups of extremophiles among these functional categories and observe which of these categories are given more importance through evolution when exposed to different harsh conditions.

Results

Taxonomic and functional categories

The list of arcahea was initially divided into three major groups according to the taxonomic groups listed on the COG database to observe the distribution of extremophiles and other archaea among these groups. Crenarchaeaota had 21 species, all of them thermophilic or hyperthermophilic, seven acidophilic, and one metalotolerant. Euryarchaeaota had 56 species, 11 hyperthermophilic, two psychrophilic, 21 halophilic, six alkaliphilic, two radiotolerant, seven acidophiles and 21 methanogens. Finally, Thaumarchaeaota (THAU) had only four species, two of them psicrophilic and the other two non-extremophilic.

We obtained a series of line graphs for each type of extremophile, for each archaeal clade, and for the whole group of archaea; and tables with the mean values and standard deviations (resumed in Table 2) used to generate the graphics. A table listing the mean values and the standard deviation of the number of COGs and percentages in each functional category for each extremophilic group can be seen Table 2 and a comparative graphic representing the data in the line graphs for the whole group of archaea and for every extremophilic group can be seen in Figure 1 which was used for an initial exploration of the differences among functional categories.

Two tables were prepared using the data obtained for every group. These (Table 3 and table 4) show the standard deviations of every extremophilic and taxonomic group separated by sizes for the most divergent functional categories observed in Figure 1 (K, L, T, O, C, E, H, P for quantities and L, P, J, T, C, G, E, H, P for percentages). We used these two tables to contrast the functional categories with higher variation among groups with the variation within each group so we could identify changes that could only be related to the extremophilic groups.

We can observe that as expected some of the functional categories which have the highest differences among all archaea have little variation in certain groups. Metabolic categories remain with high variability among most groups (CEHP for Q; CGEH for %) while K (Transcription) and L (Replication, recombination and repair) have little variation for Q, especially in extremophilic groups; and for % the categories with less variation vary from group to group. In table 4 we can see only the categories that presented the highest differences among all archaea. Categories with standard deviation values < 10 o <1% will be

referred as categories of interest. The factorial analysis allowed grouping into four protein components. The first is made up of the COGs T, L, D, K, P, H, X, B, U, O, R, S, Z, M, N, E, F, Q, I, W, G, and V. The second is comprised of COGs J and A, and the third consist only of COG W ($p<0.05$) (Data in Suppl 1).

In order to observe variations specific for any extremophilic group among the categories of interest identified, we combined the line graphs of each extremophilic group with the taxonomic groups they belong to.

Hyperthermophiles and Thermophiles (HT)

In figure 2 we compare line graphs of the HT group with the two taxonomic groups to which they belong, Crenarchaeaota (CREN) and Euryarchaeaota (EURY). We can observe a very similar amount and proportion of COGs of the T category (Signal Transduction Mechanisms) for HT and CREN without a displacement of the HT line towards EURY in spite of HT having 11 species in the EURY group which is one third of the total HT species. We can also observe that the other categories of interest for HT (K: Transcription, L: Replication, recombination and repair, O: Post translational modification, C: Energy production and conversion, P: Inorganic ion transport and metabolism) have similar values between HT and CREN but with a little displacement to the line of EURY.

Halophiles (HL)

We have that the 21 halophiles (HL) included in this study belong to the euryarchaeota group (EURY) comprising about one third of the group (21 of 56 species). Here we can see that categories of interest are coincident between the two groups and we can only observe relatively high differences for categories T (Transduction) and E (Transport and metabolism of aminoacids); the distribution of COGs, as seen in the percentages, is almost identical for both groups with a little increase in the category T.

Acidophiles

Acidophiles are evenly distributed in Crenarchaeaota (7 species that are also thermophiles) and Euryarchaeaota (7 species, from which 5 are also thermophiles) (Figure 4)

Psychrophiles

Psychrophiles show more divergent lines when compared with the Thaumarchaeaota group (THAU), although only for the amount of COGs per category, we can observe an increase in the amount of COGs for K, L, T C, E, H and P and an increase in the distribution of COGs for T (Figure S1).

Alkaliphiles, Radiotolerants, and Metalotolerants

The six alkaliphilic archaea in this study belong to the Euryarchaeaota group, and in figure S2 we can observe little difference between these groups. We also show the graphs for radio tolerant and metalo tolerant archaea (figures S3 and S4), but as there are only 2 radiotolerant archaea and 1 metalotolerant archaeon we cannot make any assumptions on the distribution of their COGs.

Methanogens

Methanogens are not extremophiles but comprises almost all the non-extremophilic archaea listed in the COG database and were used as a control group for this study. There are 24 methanogens, all of them in the Euryarchaeaota group, and as we can see in the graphs in Figure S5.

Discussion

When comparing all the line graphs obtained for each extremophilic and taxonomic group, and for the whole group of archaea we can observe differences in the distribution of certain functional categories. The standard deviation serves as a way to differentiate among variations that belong to clades and variations among extremophilic groups. A big standard deviation in any extremophilic group could be explained due to differences between the groups to where different organisms with the same extreme adaption belong. If we see organisms with the same adaptation that are in different clades but there is not a big standard deviation, we may be seeing a change that could be part of the adaptation to an extreme condition, but we have to compare with the data for the whole group of archaea and of the clades involved in that group to ensure that this little difference is not something common for every archaea. We can observe this for example when we look at categories L (Replication, recombination and repair mechanisms) and T (Signal transduction mechanisms), which have a great variability among archaea (Fig. 1), but are restricted to a small range for every type of extremophile, thus, we can say that these changes in the amount of proteins specialized for these tasks are related to the adaptation to the conditions where these archaea thrive.

As every archaea included in the CREN group is a thermophile we can assume that the T category has almost no differences because the distribution of COGs that they present are important for the thermophilic adaptation as we can corroborate in the literature which states that thermophiles arrange the composition of their membranes sending signals when their membrane receptors sense changes in the temperature [14]. The other categories of interest may also play roles in the adaptation to high temperatures as they also seem to have a more or less conserved number of proteins. K, L and O correspond to basic and very important functions for the organisms. These organisms must have proteins that have evolved to function normally at high temperatures [20,21], and thus must have been conserved and protected with little chance to mutate or duplicate in order to avoid losing stability. C and P correspond to metabolic functions related to energy production and ion metabolism, these functions must be related with the membrane modifications and the nutrient availability in the harsh environment where these microorganisms thrive [22].

We can relate the increase observed in Fig. 3 for transductional proteins in halophiles to their adaptation to the halophilic environment which includes signaling pathways that regulate the enter and exit of ions [5], and also the uptake of nutrients which has been observed to be regulated depending on the nutrient availability in species as *Halobacterium salinarum* [23].

When compared with Euryarchaeota, psychrophiles show almost no differences, not even for the category T which may indicate a relation between these COGs and the psychrophilic adaptation, but we only analyzed four species of psychrophilic archaea, two of them in EURY and the other two in THAU, it is a very low number of organisms so we can't make strong assumptions, but we can make the observation that regardless of being evenly distributed among two taxonomic groups, psychrophiles may resemble one more than the other. The lack of differences among PS and EURY may be due to the little number of species, which may represent a bias, or due to the fact that Euryarchaeaota englobes many extremophilic groups and most of them have adaptations related to signal transduction. Psychrophiles also have adapted to sense and transduce changes in the temperature, and make changes on its membrane to endure the low temperatures [8,14,15].

Despite the equivalent distribution of acidophiles between Crenarchaeaota and Euryarchaeaota, the line graphs show that the amount of COGs and proportions for acidophiles are very similar to those of Crenarchaeata, in a similar way that thermophiles showed. Here we may assume that thermophiles and acidophiles have evolved similarly as most of the acidophiles listed in the COG database have also evolved in thermophilic environments. We can't make assumptions as six species is too little to consider it as representative, but we can observe that in the same way as with other extremophilic groups, there is a difference in the T category, and also in the metabolic categories E and P. There almost no difference between the mean values of these methanogens and the Euryarchaeaota group although for extremophilic groups with similar taxonomic distributions such as halophiles we can observe bigger differences.

In general, we can observe that although the amount of protein-coding genes have a high variability among certain groups through the archaea domain, especially for Euryarchaeaota that comprises many different groups, variations on the distribution (percentages) are usually low, indicating a conservation in the proportion of genes entrusted for the different functions in these organisms [24]. As an exception we can refer to the J category (Translation, ribosomal structure and biogenesis) which show little variability in the number of COGs for the whole archaeal group, but a high variability in the proportion among groups (Fig. 1). This may be due to the importance in the conservation of these genes and the functions of their products that have led to little chance for processes that could greatly alter their numbers.

The category that has shown the most interesting relation with the extremophilic adaptations is T (Signal transduction mechanisms). As we have seen for some groups, extremophiles have adapted to their environments by recognizing different signals from the outside that will lead to the activation of several responses that will help the organisms to survive in those environments. Furthermore, variations in the metabolic categories could not be attributed directly to an extremophilic adaptation because of the high variability that they present (represented by error bars in Figs. 2 to 4), but to their sources of energy, most of these archaea obtain nutrients from minerals in their environments, the availability of these minerals may change depending on the environmental conditions and that may be the principal reason why we observe that there are more significative changes in the metabolic categories related to inorganic ion (P), coenzyme (H), aminoacid (E), carbohydrates (G) transport and metabolism, and energy metabolism.

We also have to pay some attention to R and S categories which includes genes that have not been placed in the other categories due to a lack of knowledge about their exact functions. These two categories comprise about 15% of all the COGs for all the archaea studied and there may be many genes related to the adaptation to extreme environments and that may produce some changes to the graphs if they were classified into their corresponding category, giving a better view of the changes in the distribution of genes among groups, as described for other microorganisms [12,19].

This study has a limitation. There is only a limited number of species with genomes completely sequenced so while we are using a relatively big number of species for this analysis (83 species) we cannot be sure if it is actually representative for all groups as we could only find five psicrophilic archaea, six alkalophilic archaea, one that was resistant to high concentrations of metals and two with resistance to high exposure to radiation.

In conclusion, the information obtained in this study is intended to provide clues about which functions had been given more importance when archaea had to adapt to different extreme environments and how the distribution of the functions of proteins may have affected their adaptation to these extreme conditions. This may be important regarding studies of unexplored extreme environments on Earth and beyond, and synthetic biology because these findings may be useful for the design of organisms needed as support for human exploration in such environments, or even for the colonization of these environments in the case of terraformation in worlds with harsh conditions such as Mars or any of the satellites that are hypothesized to have water oceans that could support life if they don't have it already.

Methods

Study design, techniques for collecting and processing data

An analytic-correlational, cross-section study was performed to analyze the distribution of protein-coding genes in all the archaea species registered in the COG database (http://www.ncbi.nlm.nih.gov/COG/). The full list of COGs for every archaeal species registered in the COG database was copied into Microsoft Excel 2010 (Redmond, USA), where the amount of COGs for each functional category was counted and percentages were calculated from these amounts in relation to the total number of COGs of the species [12].

A bibliographic search was performed for the whole list of sequenced archaea in the COG database in order to clasificate them in one or more groups of extremophiles. These were clasificated in six extremophilic groups (hyperthermophiles and thermophiles, halophiles, psychrophiles, acidophiles, alkaliohiles, metalotolerants and radiotolerants) and a non-extremophilic group (methanogens) that was used as a control group. Also, we grouped them into taxonomic groups (Crenarchaeota, Euryarchaeota, Thaumarchaeota and the whole Archaea group) in order to compare distributions among extremophile groups and taxonomic groups.

The data for amounts/quantities and percentages of COGs of the analyzed species was then distributed among the extremophilic and taxonomic groups. Mean values for quantities and percentages with their respective standard deviations were calculated and line graphs were created based on these data.

Data analysis

All the data obtained from the mean values and standard deviations, and the graphs were compared in order to identify differences among groups that are related to the extremophilic adaptations. First an exploratory comparison between the values for the complete set of archaea and the values for each extremophilic group was done (Figure one) to identify the most divergent values for the functional categories. Then comparisons between extremophilic and taxonomic groups were done in order to identify the changes in functional categories that can only be related to an extremophilic adaptation.

The statistical analysis was done using the statistical analyzer IBM SPSS v.24.0 (Armonk, USA) for Windows. Barlett and Kaiser–Meyer–Olkin test sphericity tests was employed for correlating variables and Equamax Rotation for the factorial analysis considering a p-value of 0.05 as a threshold.

Declarations

Acknowledgements

We are grateful with Carla Gallo and Gonzalo Moscoso for their scientific support and for their help reviewing this work, and with Ronald Torres-Martinez for his statistical advice.

Author contributions statement

R.U-I., J.M-S., and V.R-Z. made equal contributions to the conception of the work, all authors were involved at different stages in the design and implementation of the work. R.U-I., J.M-S. and S.C.R wrote the main manuscript text, and K.LO. and R.U-I prepared all figures and tables. All authors reviewed the manuscript.

Data availability statement

The datasets analysed during the current study are available in the NCBI'S Phylogenetic classification of proteins encoded in complete genomes (http://www.ncbi.nlm.nih.gov/COG/). All other data generated and analysed during the current study are available in this published article and its Supplementary Information files.

References

1. Schoff, J. W., & Packer, B. M. Early Archean (3.3-Billion to 3.5 Billion-Yer-Old) Microfossils from Warrawoona Group, Australia. Science **237**, 10–73 (1987).

2. Nisbet, E., & Fowler, C. M. The Evolution of the atmosphere in the Archaean and early Proterozoic. Chin. Sci. Bull. **56**, 4–13 (2011).
3. Mulkidjanian, A. Y., Bychkov, A. Y., Dibrova, D. V., Galperin, M. Y., & Koonin, E. V. Origin of first cells at terrestrial,anoxic, geothermal fields. PNAS **109**, 821–830 (2012).
4. Rothschild, L. J., & Mancinelli, R. L. Life in extreme environments. Nature **409**, 1092–1101 (2001).
5. Driessen, A. J., van de Vossenberg, J. L., & Konings, W. N. Membrane composition and ion-permeability in extremophiles. FEMS Microbiol. Rev. **18**, 139–148 (1996).
6. Arrage, A. A., Phelps, T. J., Benoit, R. E., Palumbo, A. V., & White, D. C. Bacterial sensitivity to UV light as a model for ionizing radiation resistance. J. Microbiol. Methods **18**, 127–136 (1993).
7. Fendrihan, S, *et al.* Extremely halophilic archaea and the issue of long-term microbial survival. Rev Environ Sci Biotechnol. **5**, 203–218 (2006).
8. Novototskaya-Vlasova, K. A., Petrovskaya, L. E., Rivkina, E. M., Olgikh, D. A., & Kirpichnikov M. P. Characterization of a Cold-Active Lipase from Psychrobacter cryohalolentis K5T and Its Deletion Mutants. Biokhimiya **78**, 501–512 (2013).
9. Kim, T. D., Ryu, H. J., Cho, H. I., Yang, C., & Kim, J. Thermal Behavior of Proteins: Heat-Resistant Proteins and Their Heat-Induced Secondary Structural Changes. Biochemistry **39**, 14839–14846 (2000).
10. Soucy, S. M., Huang, J., & Gogarten, J. P. Horizontal gene transfer: building the web of life. Nature Rev. Genet. **16**, 472–482 (2015).
11. López-García, P., Zivanovic, Y., Deschamps, P., & Moreira, D. Bacterial gene import and mesophilic adaptation in archaea. Nature Rev. Microbiol. **13**, 447–456 (2015).
12. Moya-Salazar, J., & Ubidia-Incio, R. Genome-scale analysis of protein functions and evolution from acute diarrheal disease-causing bacteria using COG database and Tax Plot. Rev. Latinoamer. Patol. Clin. **62**, 206–219 (2015).
13. Ubidia-Incio, R., & Moya-Salazar, J. New phylogenetic analysis method using TaxPlot, and its application on Acute Diarrheal Disease-causing bacteria. The Biologist Lima **15**, 15–22 (2017).
14. Kates, M., Pugh, E.L., & Ferrante, G. Regulation of Membrane Fluidity by Lipid Desaturases. Membrane Fluidity. **12**, 379–395 (1984)
15. Los, D.A., & Murata, N. Structure and expression of fatty acid desaturases. Biochmicaet Biophysica Acta. **1394**, 3–15 (1998).
16. Makarova, K. S., *et al.* Genome of the Extremely Radiation-Resistant Bacterium Deinococcus radiodurans Viewed from the perspective of Comparative Genomics. Microbiol Mol. Biol. Rev. **65**, 44–79 (2001).
17. Bakermans, C., Tollaksen, S. L., Giometti, C.S., & Wilkerson, C. Proteomic analysis of Psychrobacter cryohalolentis K5 during growth at subzero temperatures. Extremophiles **11**, 343–354 (2007).
18. Tatusov, R. L., Koonin, E. V., & Lipman, D. J. A genomic perspective on protein families. Science **278**, 631–637 (1997).

19. Makarova, K.S., Wolf, Y. I., & Koonin, E. V. Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. *Life* **5**, 818 – 40 (2015).
20. Kumar, S., & Nussinov, R. How do thermophilic proteins deal with heat? Cell. Mol. Life Sci. **58**, 1216–1233 (2001).
21. Szilágyi,A., & Závodzky, P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Structure **8**, 493–504 (2000).
22. Marteinsson, V. T., *et al.* Physiological Responses to Stress Conditions and Barophilic Behavior of the Hyperthermophilic Vent Archaeon Pyrococcus abyssi. Appl. Environ. Microbiol. **63**, 1230–1236 (1997).
23. Schmid, A. K., Reiss, D. J., Pan, M., Koide, T., & Baliga, N. S. A Single transcription factor regulates evolutionarily diverse but functionally linked metabolic pathways in response to nutrient availability. Molecular Syst. Biol. **5**, 1–15 (2009).
24. Wolf, Y.I., Makarova, K.S., Yutin, N., & Koonin, E. V. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal of horizontal gene transfer. Biol Direct. **7**, 46, (2012).

Tables

Table 1-4 are available in the Supplementary Files section.

Figures

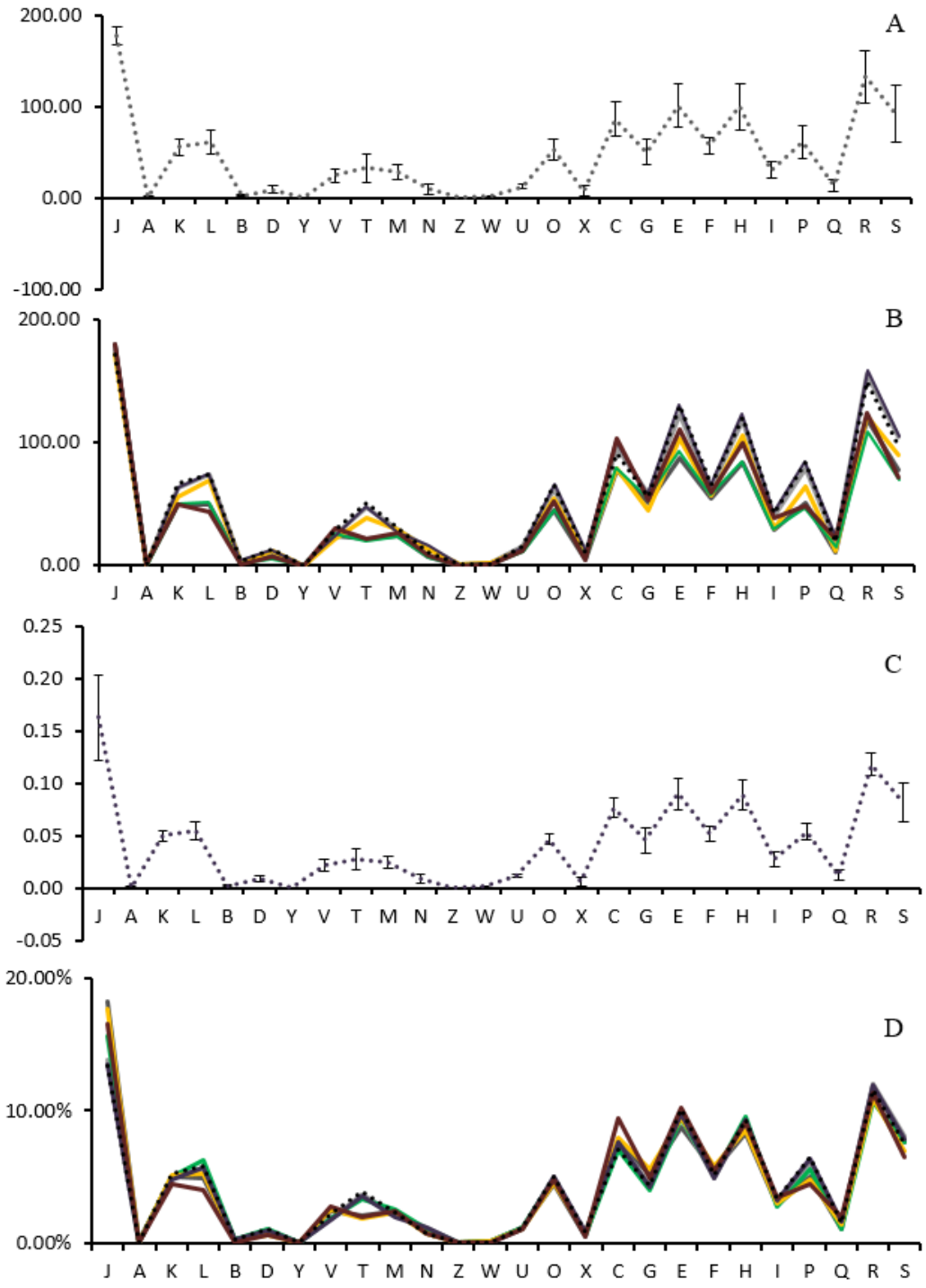

Figure 1

A. Mean quantities of COGs for all functional categories for all the archaea registered in the COG database, error bars represent the standard deviation. B. Mean quantities of COGs for each extremophile group represented in the archaea registered in the COG database. C. Mean percentages of COGs for all functional categories for all the archaea registered in the COG database, error bars represent the standard

deviation. D. Mean percentagess of COGs for each extremophile group represented in the archaea registered in the COG database.

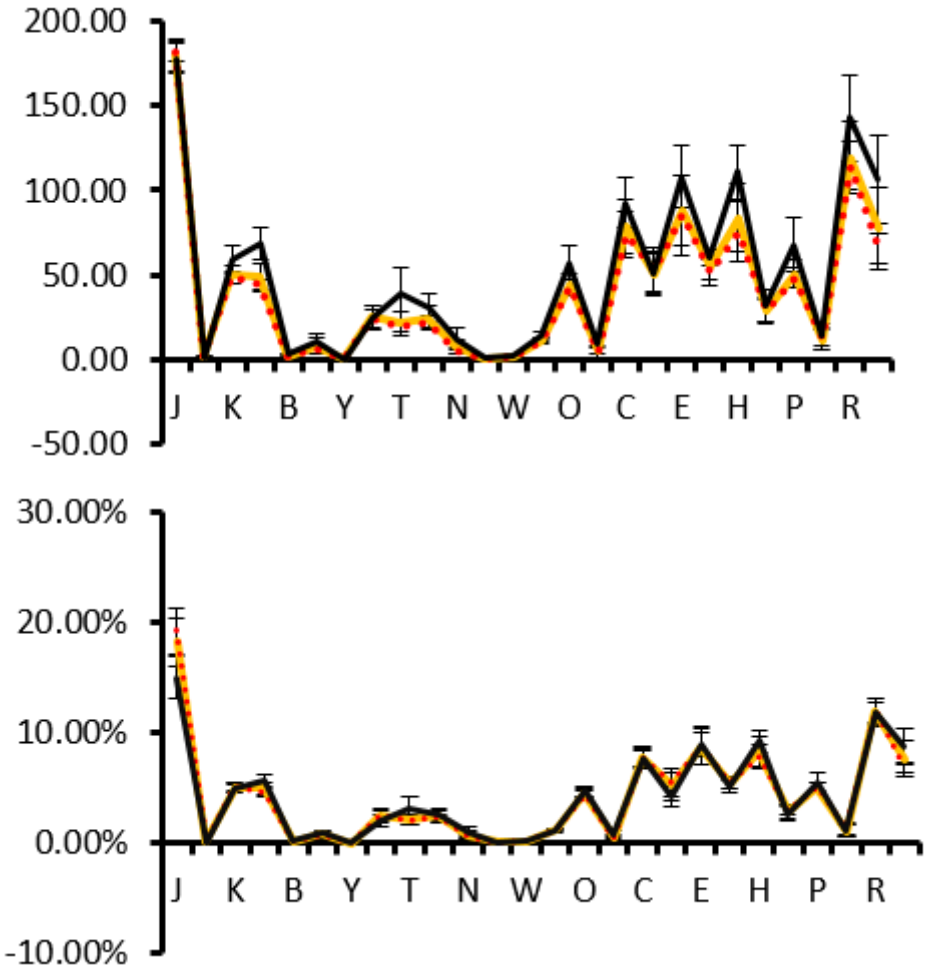

Figure 2

Comparison among Thermophiles (yellow), Crenarchaeaota (dotted red line) and Euryarchaeaota (black line). Up: Amount of COGs; Down: Percentages of COGs.

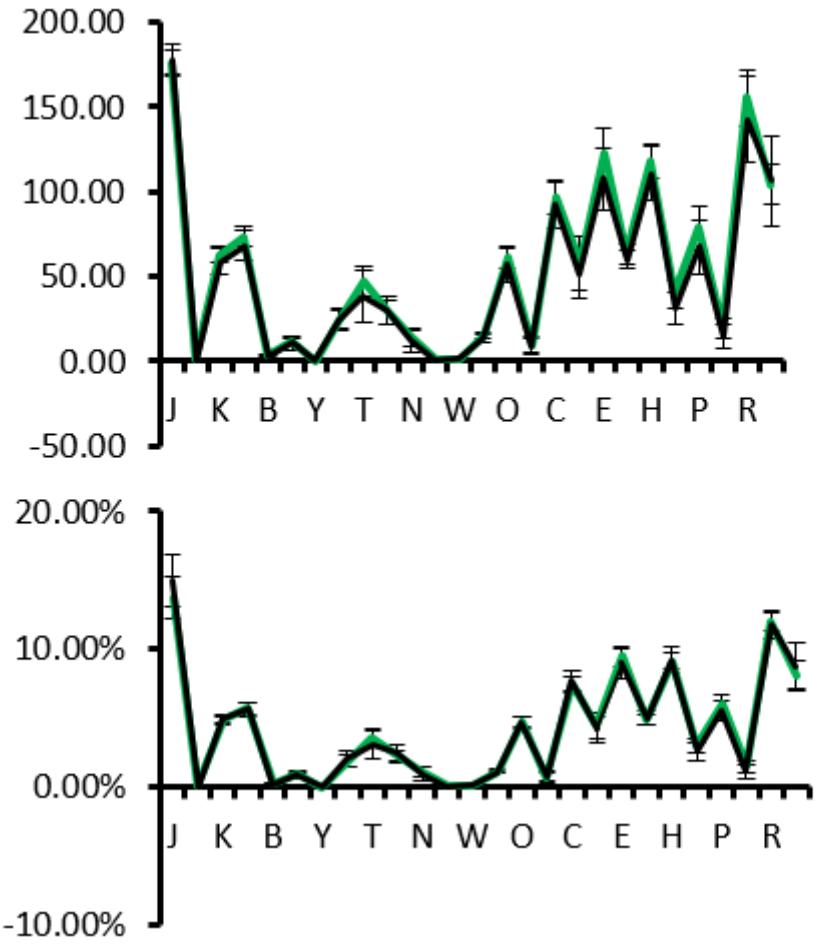

Figure 3

Comparison among Halophiles (green line), and Euryarchaeaota (black line). Up: Amount of COGs; Down: Percentages of COGs.

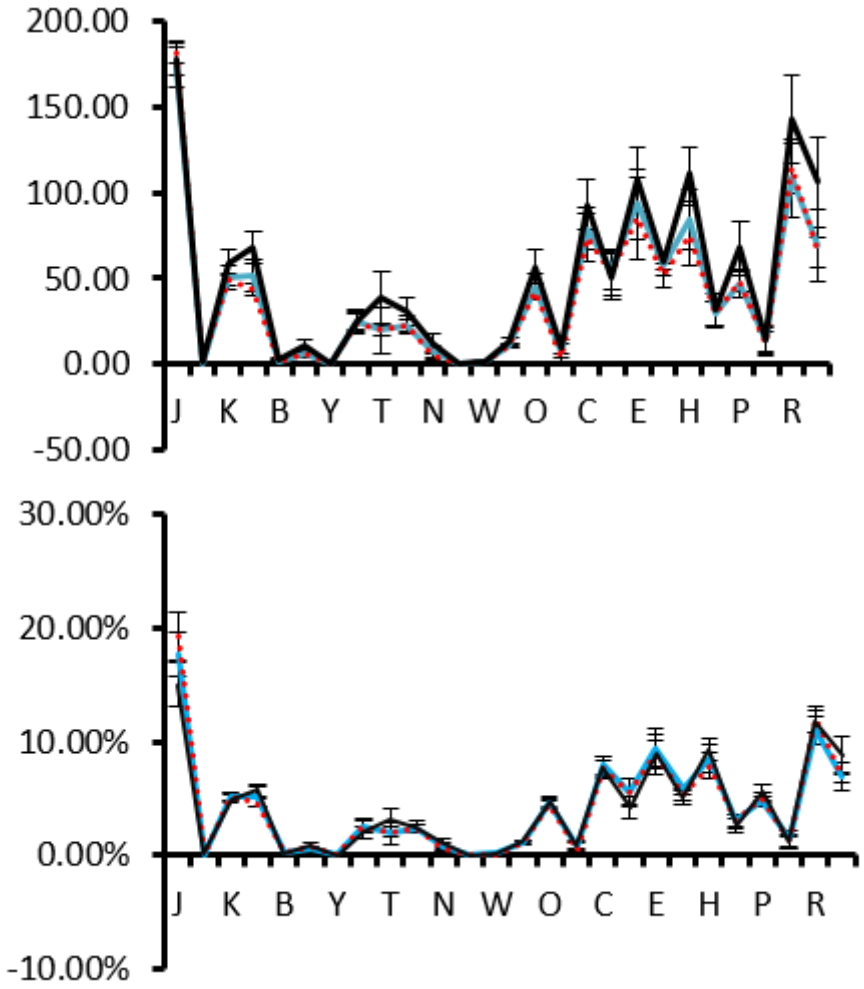

Figure 4

Comparison among Acidophiles (light blue), Crenarchaeaota (dotted red line) and Euryarchaeaota (black line). Up: Amount of COGs; Down: Percentages of COGs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Suppl1.docx
- Suppl2.docx
- Tables.docx