REPORT: COMPARATIVE GENOMICS. GRADED ASSIGNMENT.

Aída Ortiz García, r0825188.

1. INTRODUCTION

The aim of this report is to carry out a comparative genomic analysis of two bacterial species: *Thermotoga maritima* and *Thermus thermophilus HB27*. The analysis will go through finding their Best Bidirectional Hits (BBH) in terms of protein sequence and identifying orthologs, followed by building both a phylogenetic tree with homologs of these two species and a species tree, and finishing by finding conserved regions of the sequences.

2. METHODS

BBH of both species was carried out using Python. The algorithm to find the BBH uses the orthologs retrieved for both species using BLASTP: files thermophilus_vs_thermotoga.tsv and thermotoga_vs_thermophilus.tsv.

In their respective BLAST alignments files, a S score and a E-value are indicated for each of the orthologs of each species. The S score measure the similarity of the query to the sequence shown, while the E-value measures the reliability of the S score. Thus, the E-values tells you about the probability due to chance, that there is another alignment with a similarity greater than the given S score.

The algorithm created is based on finding the BBH as being the bidirectional matches that have the highest S score for each query sequence of each species. Example:

```
Query= sp|O50146|LYSY_THET2 [LysW]-L-2-aminoadipate 6-phosphate reductase
OS=Thermus thermophilus (strain ATCC BAA-163 / DSM 7039 / HB27)
OX=262724 GN=lysY PE=1 SV=2

Length=344
                                                                    Score     E
Sequences producing significant alignments:                        (Bits)  Value

Q9X2A2 N-acetyl-gamma-glutamyl-phosphate reductase OS=Thermotoga ...  189     2e-58
G4FGG9 N-acetyl-gamma-glutamyl-phosphate reductase OS=Thermotoga ...  189     2e-58
R4P1P0 Cobalt-zinc-cadmium resistance protein CzcD OS=Thermotoga ...  31.2    0.078
Q9WZX9 ZT_dimer domain-containing protein OS=Thermotoga maritima ...  31.2    0.078
Q9WZC1 tRNA-2-methylthio-N(6)-dimethylallyladenosine synthase OS=...  27.3    1.4
G4FDH1 tRNA-2-methylthio-N(6)-dimethylallyladenosine synthase OS=...  27.3    1.4
R4P0T8 Uncharacterized protein OS=Thermotoga maritima (strain ATC...  25.8    1.4
Q9X0Z5 Uncharacterized protein OS=Thermotoga maritima (strain ATC...  25.8    1.4
Q9WYE5 Oligopeptide ABC transporter, periplasmic oligopeptide-bin...  26.9    2.0
G4FHM3 Xylose-regulated ABC transporter, substrate-binding compon...  26.9    2.0
Q9WYN1 Sensor histidine kinase OS=Thermotoga maritima (strain ATC...  25.0    8.4
R4NYD6 Two-component system histidine kinase OS=Thermotoga mariti...  25.0    8.4
Q9WZT7 Threonylcarbamoyladenosine tRNA methylthiotransferase MtaB...  25.0    9.2
G4FCZ4 MiaB family protein, possibly involved in tRNA or rRNA mod...  25.0    9.2
Q9X0A4 ATPase_2 domain-containing protein OS=Thermotoga maritima ...  24.6    9.4
G4FEY6 MotifATP/GTP-binding site motif A (P-loop) G-protein coupl...  24.6    9.5
```

Thus, it first finds the best matches regarding highest S score for query sequence of each species (= best hits), and then from those matches it looks at the matches between them (= best bidirectional hits = BBH).

The protein sequences used were downloaded from Uniprot, where it was easier to select for sequences specific to these species than in NCBI. A total of 2294 and 3934 sequences were retrieved for *Thermotoga maritima* (file: uniprot-thermotoga+maritima.fasta) and *Thermus thermophilus HB27* (file: uniprot-Thermus+thermophilus+HB27.fasta) respectively.

As stated before, after the protein sequences for each species were downloaded, BLASTP was computed to identify orthologs of each species. For the BLASTP process to be computed it was needed to create a BLAST database of each of the species first.

The phylogenetic tree for the species was built choosing a specific co-orthology example: the UvrABC system protein A. Homologs of this protein for 25 different species were downloaded, including *Thermotoga maritima* and *Thermus thermophilus HB27*. The homologs were found by carrying out BLASP with the target sequence.

Once the homologs for that protein in those 25 species was retrieved, a Multiple Sequence Alignment was computed for all of them (file: mult_alignm_25sp), which was used to build a boostrap phylogenetic tree using Clustalw2. Gaps were removed, and multiple substitution correction and 1000 number of boostraps were chosen. The tree was visualized in iTool.

The species tree was built by using the 16S rRNA sequences of each of the 25 species and performing the same steps as for the phylogenetic tree. The 16S rRNA sequences were downloaded from Silva: “Silva: high quality ribosomal RNA databases” (file: 16srna_25species.fasta).

Finally, functional regions were found by looking at conserved regions among the aligned sequences. Conservation was measured by computing the Wu-Kabat variability on each region:

$$Variability = \frac{N * k}{n}$$

where N is the total number of sequences, k is the number of different aa in that position, and n is the frequency of the most common aa at that position.

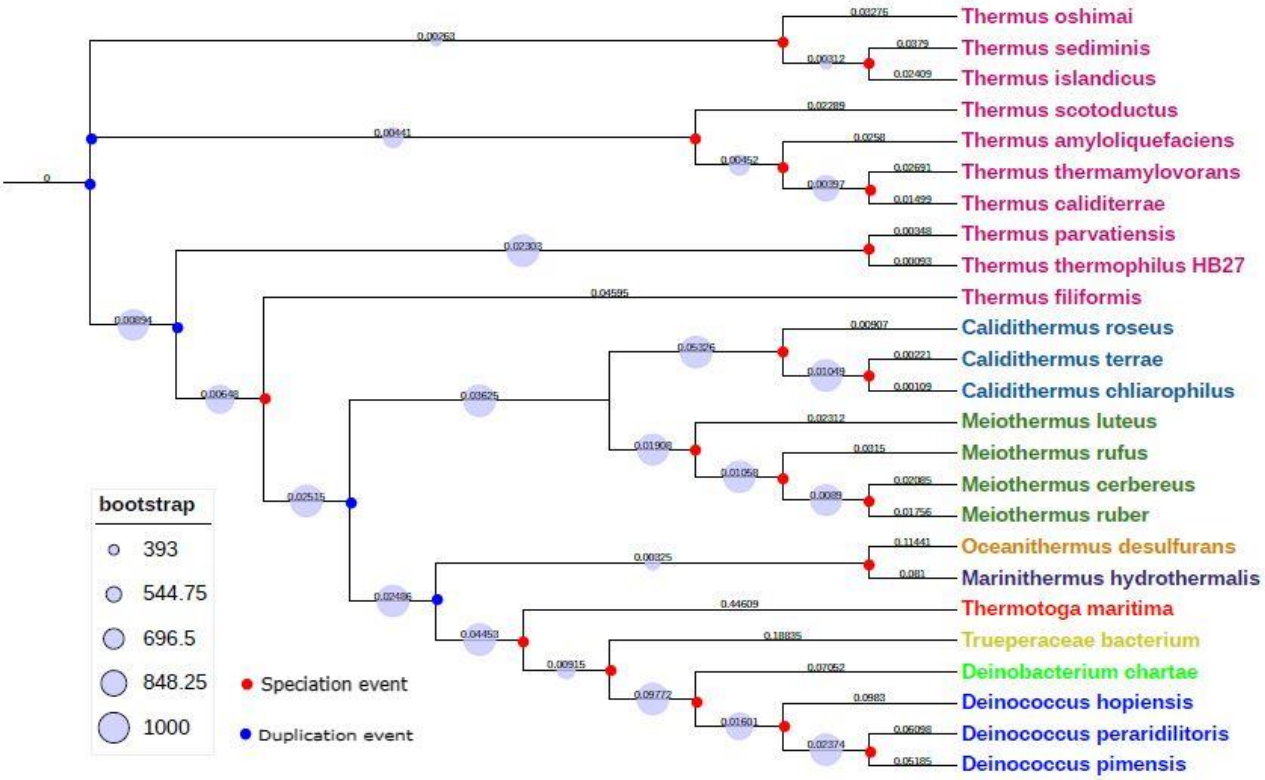
When variability in a specific region is equal to 1.0, the region is said to be conserved, as it means that there only exits one type of character in that position among all the sequences (no variation = conservation).

3. RESULTS

A total of 2294 and 3934 protein coding genes were retrieved for *Thermotoga maritima* and *Thermus thermophilus HB27* respectively. The BBH algorithm retrieved 3 BBH with **the highest S score per query sequence** between both species, which are the *DNA gyrase subunit A, UvrABC system protein A*, and *Carbamoyl-phosphate synthase large chain* proteins.

From the Python script*, Thermotoga maritima* has been found to have a total of 79 ortholog sequences in *Thermus thermophilus* (file: matches_thermotoga_thermophilus.txt), while *Thermus thermophilus* has 180 ortholog sequences in *Thermothoga maritima* (file: matches_thermophilus_vs_thermotoga.txt).

The following tree represents the phylogenetic tree for the UvrABC system protein A of 25 homologs, built with the boostrap and branch length values included:



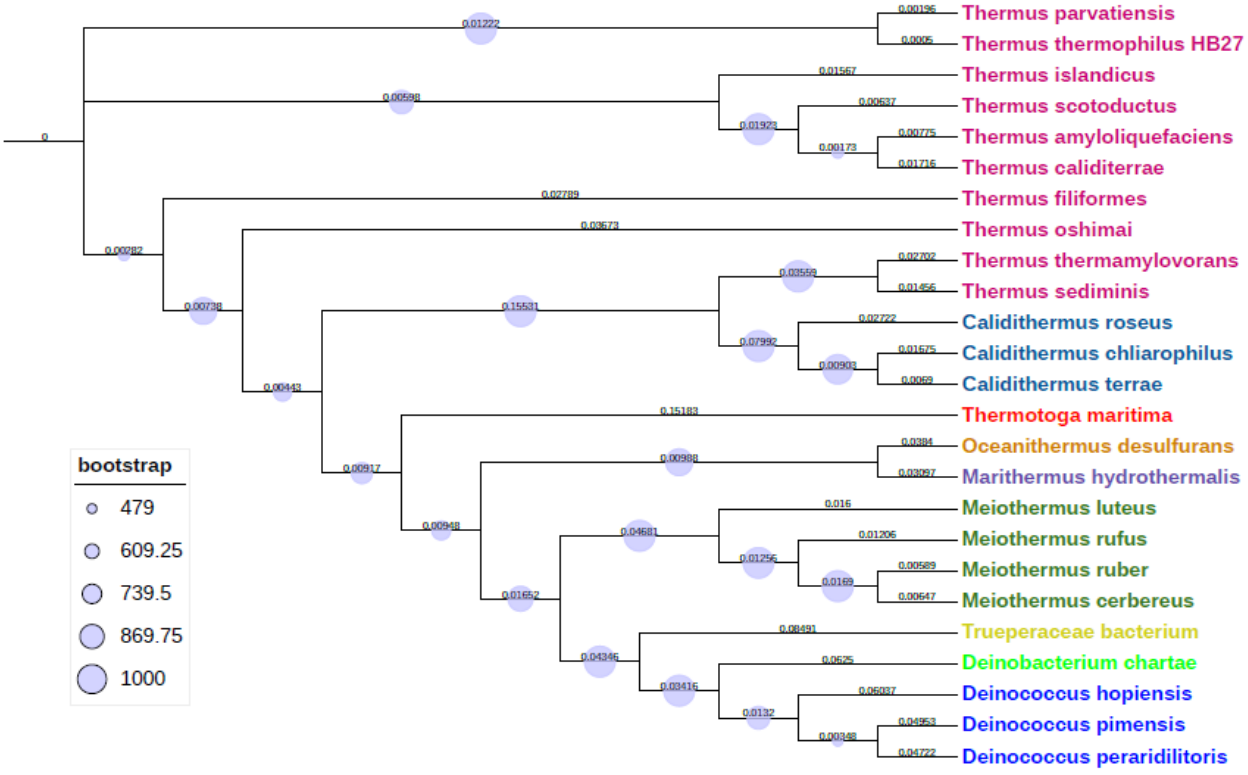
It is important to mention that Calidithermus genus has been recently replaced in the Meiothermus genus, meaning that the Calidithermus and Meiothermus genus are synonym names according to the literature. Thus, the 25 species would belong to 8 different genus: Themus, Cladithermus/Meiothermus, Oceanithermus, Marinithermus, Thermotoga, Trueperaceae, Deinobacterium, and Deinococcus.

Regarding speciation and duplication events occurring in the genes along time, the speciation ones are marked in red and in blue color for the duplication events in the phylogenetic tree above.

Speciation events give rise to orthologs, descending from the same ancestor gene but ending being orthologs. On the contrary, duplication events occur when both genes also share a common ancestor gene but the separation event has occurred within the same species, giving rise to paralog genes. Thus, the main difference is that duplication events do not give rise to genes from different species, but to genes within the same species.

As I have chosen just homolog sequences from different species, all of them are ortholog and not paralogs, although including some paralog sequences could have been a good idea to get a more complete analysis. Thus, all the selected sequences are **homologs**, but specifically **orthologs**.

Moreover, the result from taking the 16rRNA for each of the species is the following species tree:



It can be observed that both the phylogenetic and species tree topologies do not look exactly the same, although they are similar in general terms.

Regarding the functional regions, conserved positions are found to be **spread** in the alignment and not just concentrated in one region, being a total of 412 totally conserved positions, which represents a 25.29% of the alignment (file: conserved_regions.txt). Thus, a 25.29% of the sequence is seemed to be **totally** conserved (no variation along all 25 sequences from the different species (file: variability_values.txt).

4. DISCUSSION

Thermotoga maritima seem to have more known protein coding sequences than *Thermus thermophilus HB27*, as well as orthologs.

The literature states that in prokaryotes, orthology form BBH and, conversely, BBH can serve as a strong indication of gene orthology. Thus most of the BBH can be found to be orthologs.

However, this does not necessarily mean that that almost all orthologs are BBH. Thus, the observation that BBH is a predictor of orthology with a high precision rate but says nothing about its recall rate.

The BBH found by the algorithm are bases on the highest S score **per query sequence (=best hit per query sequence)**. However, it is also important to look at the E-values. The literature states that an E-value of E<e-5 tells that the alignment is highly unique. Therefore, if we would like to be sure that the alignment is not due to an error, the BBH would need to be under that threshold. However, in our case none of the BBH has a E-value under that threshold.

As it was stated before, orthology form BBH but the literature also states that this BBH method miss many pairs. Therefore, it shouldn't be taken as the rule.

The phylogenetic tree of the 25 homologs built for the UvrABC system protein A indicates that the gene first suffered a couple of duplication events within the same species, but then speciation events occurred giving rise to different genes of different species. The genes that codify for that protein within a certain genus seem to be more genetically similar than to the rest of the species from another genus. However, some of the species' genes from the Thermus genus seem to be more similar to the genes of the species in the Calidithermus genus. Moreover, the genes from the Oceanithermus and Marinithermus genus are more similar between them, which could indicate that the speciation events occurred produced mutations that could be correlated to the water conditions in which these bacteria live.

For the species tree, we can infer that Calidithermus genus is more closely related to the Thermus genus than to the Meiothermus genus. This does not relate to the current literature, which says that Meiothermus and Calidithermus genus are synonyms. The building of the species tree was repeated with sequences retrieved from different resources, but still retrieved the same result. However, the phylogenetic tree does group the genes from both genus together. Thus, a possible explanation is that another criteria was used by the literature to join both genus in one.

Regarding the functional analysis, conserved regions are considered to be good candidates of functional regions but literature states that they may not be always functional. Our conserved positions of the sequences involved a around a 25% of the total. The literature states this is a low percentage of conservation. Thus, this protein gene must have been suffered several mutations allowing it to be more specific to each species features, and therefore might not be a protein involved in basal pathways, but in allowing each species to improve their fitness in their specific environments. The spreading of the conserved regions along each sequence could be supporting this explanation.

5. REFERENCES

- Biocomp.unibo.it. 2022. [online] Available at: <http://www.biocomp.unibo.it/casadio/LMBIOTEC/evalue>
- Cai, X., Hu, H. and Li, X., 2009. A new measurement of sequence conservation. *BMC Genomics*, 10(1).
- Dalquen, D. and Dessimoz, C., 2013. Bidirectional Best Hits Miss Many Orthologs in Duplication-Rich Clades such as Plants and Animals. *Genome Biology and Evolution*, 5(10), pp.1800-1806.
- Glover, N., Sheppard, S. and Dessimoz, C., 2021. Homoeolog Inference Methods Requiring Bidirectional Best Hits or Synteny Miss Many Pairs. *Genome Biology and Evolution*, 13(6).
- Takuno, S. and Gaut, B., 2013. Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proceedings of the National Academy of Sciences*, 110(5), pp.1797-1802.
- Uniprot.org. 2022. *carB - Carbamoyl-phosphate synthase large chain - carB gene & protein*. [online] Available at: <https://www.uniprot.org/uniprot/P00968>
- Wolf YI, Koonin EV. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol Evol.* 2012;4(12):1286–1294.