1. Introduction

In this study, Prosthecochloris aestuarii and Rubrobacter xylanophilus were subjected to comparative genomic analysis. Orthologs were found between protein sequences from two different species using the Best Bidirectional Hits (BBH) method, and a phylogenetic tree was created by aligning the homologs of one example of orthologs. Speciation and Duplication events were marked on the tree, which was compared with a species tree based on rRNA resources. Multiple Sequences Alignment (MSA) was applied in the end to identify conserved regions in protein sequences and potential promotor regions in DNA sequences.

2. Methods

2.1. Compute orthologs using Best Bidirectional Hits

All protein sequences of Prosthecochloris aestuarii (RefSeq: GCF_000020625.1) and Rubrobacter xylanophilus (RefSeq: GCF_000014185.1) were downloaded from NCBI database (https://www.ncbi.nlm.nih.gov/data-hub/genome/) as FASTA files. Two FASTA files were used to build a database of Prosthecochloris aestuarii and a database of Rubrobacter xylanophilus, then protein sequences from one species were used to run protein blast with the database of another species' protein sequences and the database of its own protein sequences. Four output files were generated from protein blast (output_Pa.txt, output_Pa_self.txt, output_Rx.txt, ouput_Rx_self.txt), and a python script (BBH.py) was created to generate BBH for two species using these output files, and it also indicated all in-paralogs.

2.2. Check orthology with a phylogenetic tree

One protein sequence of orthologs from co-orthology was picked to run protein blast in NCBI with standard databse and retrieve homologs to that protein sequence, then these homologs and the BBH were used to do MSA and generate the phylogenetic tree using clastalw2. When building the tree in clastalw2, the function of excluding positions with gaps and correcting for multiple substitutions was turned on, and the bootstrap value was set as 1000. The tree file was visualized in iTol (http://itol.embl.de) including branch lengths and bootstrap values, and speciation events were marked as red squares and duplication events were marked as green squares. In addition, rRNA sequences of all homologs species were downloaded from silva rRNA database (https://www.arb-silva.de/), and a species tree was built using these rRNA sequences. The species tree was also visualized in iTol to discuss the differences from the phylogenetic tree of homologs.

2.3. Identify functional regions

The resulting file of MSA was used to run a python script (Conservation.py), which uses an algorithm to calculate the variability to measure conservation and results in a list of variabilities to every position in the sequence (Variability.txt). The list of variability was exploited to generate a variability plot in R, and the value of variability is equal to 1 means all sequences share the same amino acid in that position.

$$variability = \frac{\mathrm{N} * \mathrm{k}}{\mathrm{n}}$$

N: The number of sequences in the alignment

k: The number of different amino acids at a given position n: The frequency of the most common amino acid at that position

2.4. Identify promotor regions

In order to identify promotor regions, the DNA sequences to specific proteins (orthologs) were downloaded from the NCBI and moved to one FASTA file. Clastalw2 was used to do the MSA analysis of these DNA sequences, and PVS (http://imed.med.ucm.es/PVS/) was used to see the variability plot.

3. Results

3.1. Orthologs

In two FASTA files (R.x_protein.faa and P.a_protein.faa) for species, 3205 protein-coding genes in the genome of Rubrobacter xylanophilus, and 2302 protein-coding genes in the genome of Rubrobacter xylanophilu. After running protein blast and BBH.py, 804 pairs of BBH were found between the two species, and 184 pairs of in-paralogs were identified. Because both in-paralogs are orthologous to the BBH, a total of 1792 orthologs were found between the two species.

3.2. Phylogenetic trees

WP_012504624.1 was picked as the example to run protein blast, this is a S41 family peptidase of Prosthecochloris aestuarii, its in-paralog is WP_012505586.1 and the BBH of it is WP_011563862.1 that is a S41 family peptidase of Rubrobacter xylanophilus. After protein blast, 72 homologs sequences from 26 different species were downloaded in one FASTA file (homologs.txt), and rRNA sequences of these species were also downloaded in one FASTA file (rRNA_species.txt). The MSA analysis was done for each FASTA file (guidetree_homolos, guidtree_rRNA, align_homologs and align_rRNA), and A phylogenetic tree (Supplementary Figure 1 and Supplementary Figure 3) and a species tree (Supplementary Figure 2) were built to study the speciation events and duplication events.

3.3. Conserved regions

A variability plot (Figure 1) was generated using the resulting file from MSA analysis. The variability is lower, the sequence region is more likely to be a conserved region. In the variability list (Variability.txt), positions 141T, 170I, 195E, 223I, 252K, 284G, 286Q, 289S, 290V, 291K, 293A, 303V, 308T, 337T, 347T, 348I, 351S, 352E, 355A, 362D, 367V, 372Q, 374L, 376R, 397I,398Q,400Q,403G have the lowest variability 1 (start with 0), and from the plot, sequence region from 330 to 410 has relatively low variability.

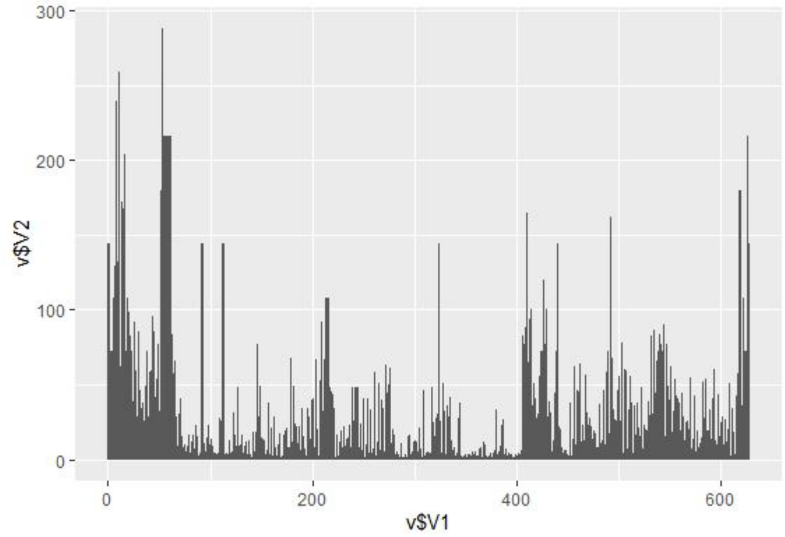


Figure 1: Variability plot of 72 homologs. X-axis represents each position in the sequence, Y-axis represent the variability value of each position.

3.4. Promotor regions

Three protein sequences were used to study the orthology, they are WP_012504624.1, its in-paralog WP_012505586.1 and the BBH WP_011563862.1 Both WP_012504624.1 and WP_012505586.1 are orthologous to WP_011563862.1, so a FASTA file containing three DNA sequences related to these three proteins were generated (orthologs.txt). Then the FASTA file was used to do the MSA analysis and generate an output file (align_orthologs). The alignment file was uploaded to PVS to create the variability plot (Figure 2). In the plot, positions from 0 to 70 have a region with relatively low variability.

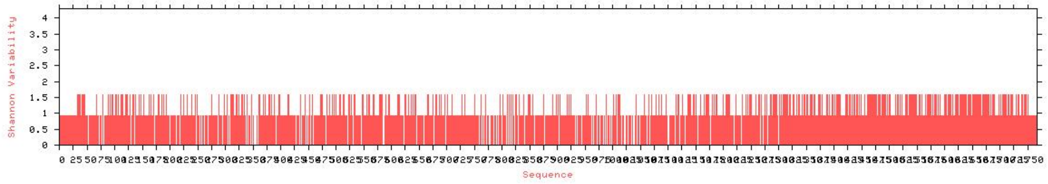


Figure 2: Variability plot of 3 orthologs. X-axis represents each position in the sequence, Y-axis represent the variability value of each position.

4. Discussion

804 pairs of BBH, 184 pairs of in-paralogs and a total of 1792 orthologs were found between proteomes of Prosthecochloris aestuarii (3205 protein-coding genes) and Rubrobacter xylanophilus (2302 protein-coding genes), which means a relatively large percent of protein-coding genes are orthologous between two species and some of them are even co-orthologous, this discovery indicates that these protein-coding genes are evolved from a common ancestral gene and retained similar functions in two species.

The species tree indicates that there exist at least five or six speciation events between these two species from the common ancestral species and the branch length is relatively long between the two species, which means the orthologous protein-coding genes that are still retained in the two species should have essential functions. In the phylogenetic tree with 72 different homologous proteins, the orthologous proteins (BBH: WP 011563862.1 and WP 012505586.1) stay very close and the in-paralogous protein (WP 012504624.1) is nearby. There is only one speciation event before BBH, which indicates these orthologous proteins share a similar function, and this function should be essential for these two species, so this function has never changed a lot during evolution.

If proteins have similar functions, they should have similar sequences. For conserved regions among these homologs, they are likely to have some functions like the Post-translational modification (PTM) in the interaction, and these sites are more likely to be close to each other.

In the end of this study, I was trying to find the promotor regions in DNA sequences of these orthologs. One study (Jin et al., 2006) showed that promotor regions are more likely to be conserved in orthologs, and most sites of the region at the beginning of these DNA sequences from 0 to 70 have variability of 1, so this region is the potential promotor region for these orthologs.

5. References

Jin, V. X., Singer, G. A. C., Agosto-Pérez, F. J., Liyanarachchi, S., & Davuluri, R. V. (2006). Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. *BMC Bioinformatics*, *7*(1), 114–114. https://doi.org/10.1186/1471-2105-7-114

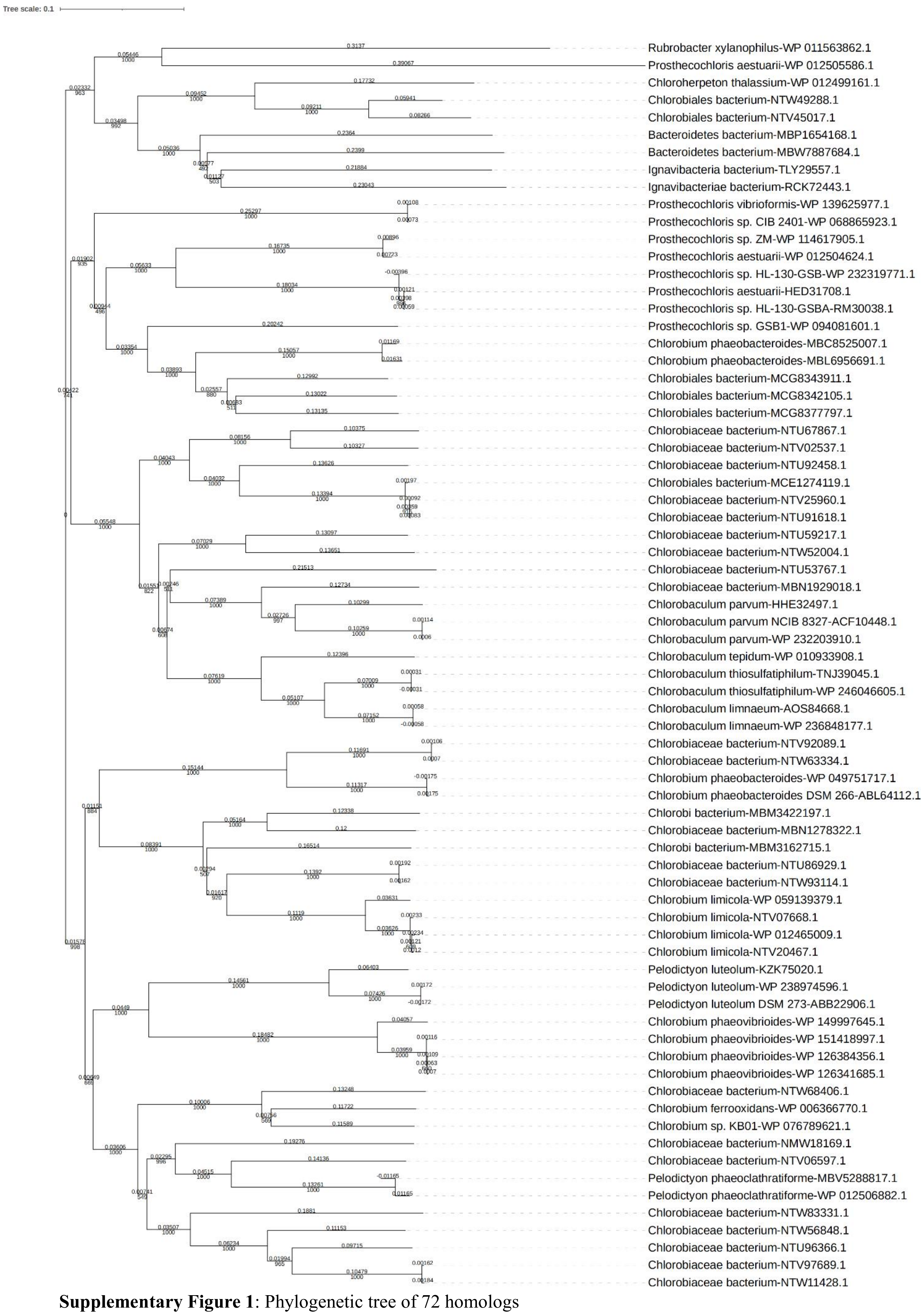
6. Appendix

BBH.py:

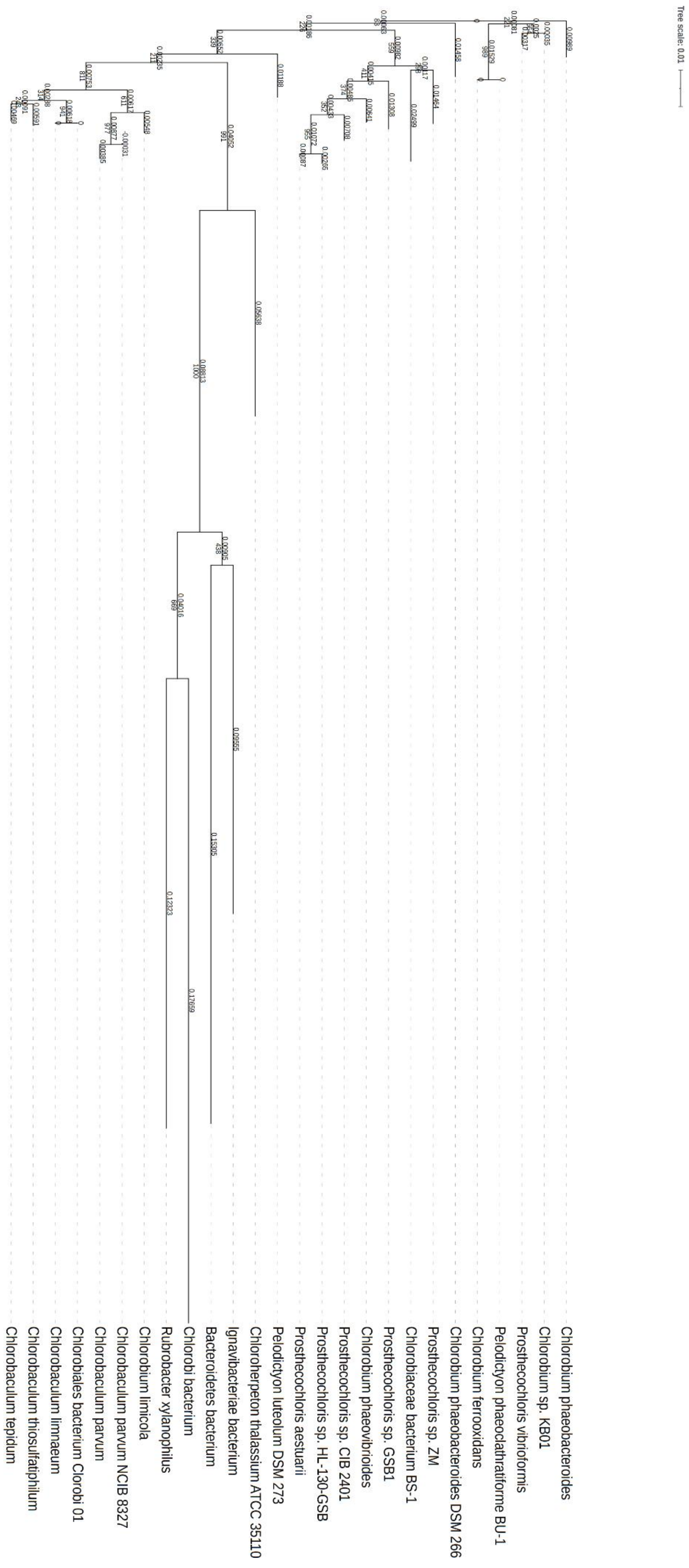
```
import pandas as pd
import numpy as np

BBH = {}
inparalogs = {}

# Create arrays for four different protein BLAST results
# output_Pa.txt: protein BLAST of protein sequences of Prosthecochloris aestuarii, database is protein
# sequences of Rubrobacter xylanophilus
# output_Rx.txt: protein BLAST of protein sequences of Rubrobacter xylanophilus, database is protein
# sequences of Prosthecochloris aestuarii
# output_Pa_self.txt: protein BLAST of protein sequences of Prosthecochloris aestuarii, database is protein
#sequences of Prosthecochloris aestuarii
# output_Rx_self.txt: protein BLAST of protein sequences of Rubrobacter xylanophilus, database is protein
# sequences of Rubrobacter xylanophilus
# In these four arrays, 1st column is the query name, 2nd column is the target name, 3rd column is the
# coverage rate, 4th column is the evalue
pa_df = pd.read_csv('D:/blast-BLAST_VERSION+/bin/output_Pa.txt', sep="\t", header=None)
pa_protein_hits = np.array(pa_df.iloc[:, [0, 1, 2, 10]])
print(len(pa_protein_hits))

rx_df = pd.read_csv('D:/blast-BLAST_VERSION+/bin/output_Rx.txt', sep="\t", header=None)
rx_protein_hits = np.array(rx_df.iloc[:, [0, 1, 2, 10]])
print(len(rx_protein_hits))

pa_self_df = pd.read_csv('D:/blast-BLAST_VERSION+/bin/output_Pa_self.txt', sep="\t", header=None)
pa_self_protein_hits = np.array(pa_self_df.iloc[:, [0, 1, 2, 10]])
print(len(pa_self_protein_hits))

rx_self_df = pd.read_csv('D:/blast-BLAST_VERSION+/bin/output_Rx_self.txt', sep="\t", header=None)
rx_self_protein_hits = np.array(rx_self_df.iloc[:, [0, 1, 2, 10]])
print(len(rx_self_protein_hits))

# For each protein sequence that finds the best hit in output_Pa.txt, check if each sequence has the
# in-paralogs
for i in range(len(pa_protein_hits)):
    protein = pa_protein_hits[i, 0]
    BH = pa_protein_hits[i, 1]
    evalue = pa_protein_hits[i, 3]

    # For each protein sequence that finds the best hit in output_Rx.txt, check if that sequence is the BH and
    # if the BH of this sequence is the protein
    for m in range(len(rx_protein_hits)):
        if rx_protein_hits[m, 0] == BH and rx_protein_hits[m, 1] == protein:
            BBH[protein] = BH
         # Now we have find one pair of BBH, and the next is to check if there exists inparalogs
            # First check if rx_protein has an inparalogs
            rxprotein = rx_protein_hits[m, 0]
            rxevalue = rx_protein_hits[m, 3]
            for n in range(len(rx_self_protein_hits)):
                # If two sequences from one species are more similar to each other than any sequence in
              # the other specie, this sequence has inparalogs
                # , and both sequences are orthologous to the BBH
                if rx_self_protein_hits[n, 0] == rxprotein and rx_self_protein_hits[n, 2] != 100 and
rx_self_protein_hits[n, 3] < rxevalue:
                    rxinparalogs = rx_self_protein_hits[n, 1]
                    inparalogs[rxprotein] = rxinparalogs

            # Then check if pa_protein has inparalogs
            for j in range(len(pa_self_protein_hits)):
                # If two sequences from one species are more similar to each other than any sequence in
                   # the other specie, this sequence has an inparalogs
                # , and both sequences are orthologous to the BBH
                if pa_self_protein_hits[j, 0] == protein and pa_self_protein_hits[j, 2] != 100 and
pa_self_protein_hits[j, 3] < evalue:
                    painparalogs = pa_self_protein_hits[j, 1]
                    inparalogs[protein] = painparalogs

print("The number of BBH pairs:", len(BBH))
print("The number of inparalogs:", len(inparalogs))
print("The number of orthologs:", len(BBH) * 2 + len(inparalogs))
```

Conservation.py:

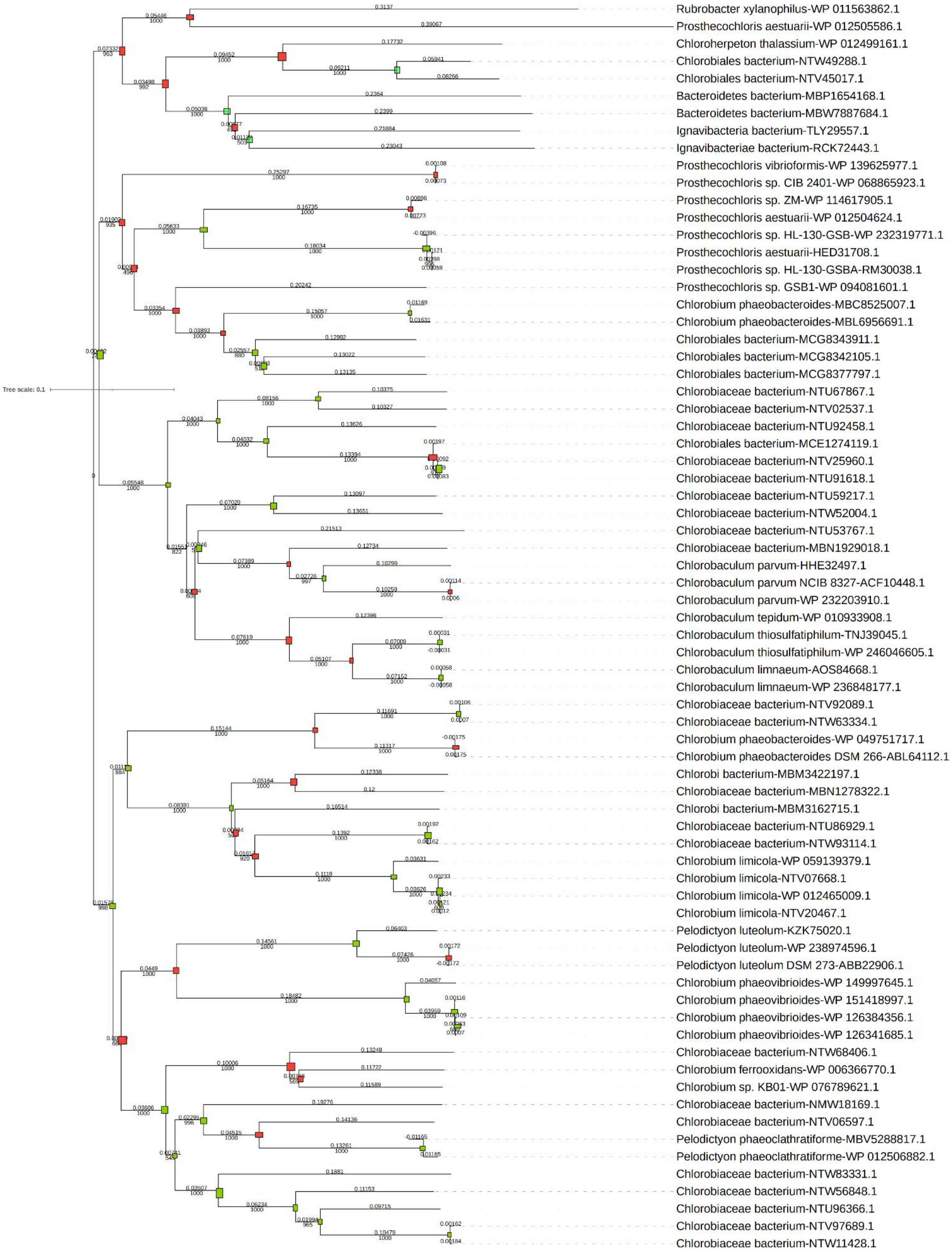
```
import pandas as pd
import numpy as np

# Obtain the multiple sequence alignment, and turn to array
MSA_df = pd.read_csv('C:/Users/LNH/Desktop/Study/Comparative/Comparative Genomics/Assignment/align_homologs', sep="\t", header=None, skiprows=1)
MSA = np.array(MSA_df)

# First of all, we need to delete all useless information before sequence, which means to
# delete 36 positions in each line
# then, we need to remove all lines that do not contain sequence
remove = []
for i in range(len(MSA)):
    MSA[i, 0] = MSA[i, 0][36:]
    if '.' in MSA[i, 0] or ':' in MSA[i, 0] or '*' in MSA[i, 0]:
        remove.append(i)
i = 0
for j in range(len(remove)):
   MSA = np.delete(MSA, remove[j] - i, axis=0)
   i = i + 1

# In this research, we have 72 different protein sequences, MSA divides them into 13 regions,
# the last region contains 29 sites, and each of the rest contians 50 sites
proteins = 72
region_sites = len(MSA[1, 0])
last_regoin_sites = len(MSA[-1, 0])
regions = len(MSA)//proteins
columns = (regions - 1) * region_sites + last_regoin_sites
# We create a empty list with 936 rows and 50 columns
seq = np.empty((len(MSA), region_sites), dtype=str)

# For convenience, we add "1" to fill in the last region, so each region has 50 sites
for i in range(len(MSA) - proteins, len(MSA)):
    for j in range(region_sites - last_regoin_sites):
        MSA[i, 0] = MSA[i, 0] + "1"

# We move every site from MSA to the list 'seq', so each element in the 'seq' represents one
# site on the sequence
for i in range(len(MSA)):
    for j in range(region_sites):
        seq[i, j] = list(MSA[i, 0])[j]

# We reshape the seq, now we have 13 regions like MSA result, each region has 72 rows
# (protein) and 50 columns(sites)
seq.shape = (regions, proteins, region_sites)
print(seq)

# Now we calculate varibality for each position
# Be careful! The position contain "1" should be ignored!
# We put all varibalities into one list 'Vs'
Vs = []
for i in range(regions):
    for j in range(region_sites):
        if seq[i][0][j] != "1":
            eachSite = []
            for m in range(proteins):
                eachSite.append(seq[i][m][j])
            n = 1
            for e in eachSite:
                if e != '-' and eachSite.count(e) > n:
                    max = e
                    n = eachSite.count(e)
            k = len(set(eachSite))
            N = proteins
            Variability = k * N / n
            Vs.append(Variability)

# Print all sites that has variability equal to 1, which means that amino acid conserved in all
# protwins
for i in range(len(Vs)):
    if Vs[i] == 1:
        print(i)
        print(seq[i//region_sites][0][i%regions])


# Create csv file for variablity, and we can plot the values in R
Vs_df = pd.DataFrame(data=Vs)
Vs_df.to_csv('C:/Users/LNH/Desktop/Study/Comparative/Comparative
Genomics/Assignment/Variability.txt', encoding = 'gbk', header=None)
```


Tree scale: 0.1
0.3137
Rubrobacter xylanophilus-WP 011563862.1
0.05446
1000
0.39067
Prosthecochloris aestuarii-WP 012505586.1
0.17732
Chloroherpeton thalassium-WP 012499161.1
0.02332
963
0.09452
1000
0.05941
Chlorobiales bacterium-NTW49288.1
0.09211
1000
0.08266
Chlorobiales bacterium-NTV45017.1
0.03498
992
0.2364
Bacteroidetes bacterium-MBP1654168.1
0.05036
1000
0.2399
Bacteroidetes bacterium-MBW7887684.1
0.00577
492
0.21884
Ignavibacteria bacterium-TLY29557.1
0.01127
503
0.23043
Ignavibacteriae bacterium-RCK72443.1
0.00108
Prosthecochloris vibrioformis-WP 139625977.1
0.25297
1000
0.00073
Prosthecochloris sp. CIB 2401-WP 068865923.1
0.00896
Prosthecochloris sp. ZM-WP 114617905.1
0.16735
1000
0.01902
935
0.00723
Prosthecochloris aestuarii-WP 012504624.1
0.05633
1000
-0.00396
Prosthecochloris sp. HL-130-GSB-WP 232319771.1
0.18034
1000
0.00121
Prosthecochloris aestuarii-HED31708.1
0.00298
998
0.00944
496
0.00059
Prosthecochloris sp. HL-130-GSBA-RM30038.1
0.20242
Prosthecochloris sp. GSB1-WP 094081601.1
0.01169
Chlorobium phaeobacteroides-MBC8525007.1
0.03354
1000
0.15057
1000
0.01631
Chlorobium phaeobacteroides-MBL6956691.1
0.03893
1000
0.12992
Chlorobiales bacterium-MCG8343911.1
0.00422
741
0.02557
880
0.13022
Chlorobiales bacterium-MCG8342105.1
0.00633
511
0.13135
Chlorobiales bacterium-MCG8377797.1
0.10375
Chlorobiaceae bacterium-NTU67867.1
0.08156
1000
0.10327
Chlorobiaceae bacterium-NTV02537.1
0.04043
1000
0.13626
Chlorobiaceae bacterium-NTU92458.1
0.04032
1000
0.00197
Chlorobiales bacterium-MCE1274119.1
0.13394
1000
0.00092
Chlorobiaceae bacterium-NTV25960.1
0.00259
0.00083
0
Chlorobiaceae bacterium-NTU91618.1
0.05548
1000
0.13097
Chlorobiaceae bacterium-NTU59217.1
0.07029
1000
0.13651
Chlorobiaceae bacterium-NTW52004.1
0.21513
Chlorobiaceae bacterium-NTU53767.1
0.01553
822
0.00246
511
0.12734
Chlorobiaceae bacterium-MBN1929018.1
0.07389
1000
0.10299
Chlorobaculum parvum-HHE32497.1
0.02726
997
0.00114
Chlorobaculum parvum NCIB 8327-ACF10448.1
0.00674
608
0.10259
1000
0.0006
Chlorobaculum parvum-WP 232203910.1
0.12396
Chlorobaculum tepidum-WP 010933908.1
0.00031
Chlorobaculum thiosulfatiphilum-TNJ39045.1
0.07619
1000
0.07009
1000
-0.00031
Chlorobaculum thiosulfatiphilum-WP 246046605.1
0.05107
1000
0.00058
Chlorobaculum limnaeum-AOS84668.1
0.07152
1000
-0.00058
Chlorobaculum limnaeum-WP 236848177.1
0.00106
Chlorobiaceae bacterium-NTV92089.1
0.11691
1000
0.0007
Chlorobiaceae bacterium-NTW63334.1
0.15144
1000
-0.00175
Chlorobium phaeobacteroides-WP 049751717.1
0.11317
1000
0.00175
Chlorobium phaeobacteroides DSM 266-ABL64112.1
0.01151
884
0.12338
Chlorobi bacterium-MBM3422197.1
0.05164
1000
0.12
Chlorobiaceae bacterium-MBN1278322.1
0.08391
1000
0.16514
Chlorobi bacterium-MBM3162715.1
0.00192
Chlorobiaceae bacterium-NTU86929.1
0.00294
507
0.1392
1000
0.00162
Chlorobiaceae bacterium-NTW93114.1
0.01617
920
0.03631
Chlorobium limicola-WP 059139379.1
0.1119
1000
0.00233
Chlorobium limicola-NTV07668.1
0.03626
1000
0.00234
Chlorobium limicola-WP 012465009.1
0.01578
998
0.00121
0.0012
Chlorobium limicola-NTV20467.1
0.06403
Pelodictyon luteolum-KZK75020.1
0.14561
1000
0.00172
Pelodictyon luteolum-WP 238974596.1
0.07426
1000
-0.00172
Pelodictyon luteolum DSM 273-ABB22906.1
0.0449
1000
0.04057
Chlorobium phaeovibrioides-WP 149997645.1
0.18482
1000
0.00116
Chlorobium phaeovibrioides-WP 151418997.1
0.03959
1000
0.00109
Chlorobium phaeovibrioides-WP 126384356.1
0.00063
0.0007
Chlorobium phaeovibrioides-WP 126341685.1
0.00649
669
0.13248
Chlorobiaceae bacterium-NTW68406.1
0.10006
1000
0.11722
Chlorobium ferrooxidans-WP 006366770.1
0.00756
569
0.11589
Chlorobium sp. KB01-WP 076789621.1
0.19276
Chlorobiaceae bacterium-NMW18169.1
0.03606
1000
0.02295
996
0.14136
Chlorobiaceae bacterium-NTV06597.1
0.04515
1000
-0.01165
Pelodictyon phaeoclathratiforme-MBV5288817.1
0.13261
1000
0.01165
Pelodictyon phaeoclathratiforme-WP 012506882.1
0.00741
549
0.1881
Chlorobiaceae bacterium-NTW83331.1
0.03507
1000
0.11153
Chlorobiaceae bacterium-NTW56848.1
0.06234
1000
0.09715
Chlorobiaceae bacterium-NTU96366.1
0.01994
965
0.00162
Chlorobiaceae bacterium-NTV97689.1
0.10479
1000
0.00184
Chlorobiaceae bacterium-NTW11428.1

Supplementary Figure 1: Phylogenetic tree of 72 homologs


Tree scale: 0.01
Chlorobium phaeobacteroides
Chlorobium sp. KB01
Prosthecochloris vibrioformis
Pelodictyon phaeoclathratiforme BU-1
Chlorobium ferrooxidans
Chlorobium phaeobacteroides DSM 266
Prosthecochloris sp. ZM
Chlorobiaceae bacterium BS-1
Prosthecochloris sp. GSB1
Chlorobium phaeovibrioides
Prosthecochloris sp. CIB 2401
Prosthecochloris sp. HL-130-GSB
Prosthecochloris aestuarii
Pelodictyon luteolum DSM 273
Chloroherpeton thalassium ATCC 35110
Ignavibacteriae bacterium
Bacteroidetes bacterium
Chlorobi bacterium
Rubrobacter xylanophilus
Chlorobium limicola
Chlorobaculum parvum NCIB 8327
Chlorobaculum parvum
Chlorobiales bacterium Clorobi 01
Chlorobaculum limnaeum
Chlorobaculum thiosulfatiphilum
Chlorobaculum tepidum

Supplementary Figure 2: Species tree of 26 species


Tree scale: 0.1
Rubrobacter xylanophilus-WP 011563862.1
Prosthecochloris aestuarii-WP 012505586.1
Chloroherpeton thalassium-WP 012499161.1
Chlorobiales bacterium-NTW49288.1
Chlorobiales bacterium-NTV45017.1
Bacteroidetes bacterium-MBP1654168.1
Bacteroidetes bacterium-MBW7887684.1
Ignavibacteria bacterium-TLY29557.1
Ignavibacteriae bacterium-RCK72443.1
Prosthecochloris vibrioformis-WP 139625977.1
Prosthecochloris sp. CIB 2401-WP 068865923.1
Prosthecochloris sp. ZM-WP 114617905.1
Prosthecochloris aestuarii-WP 012504624.1
Prosthecochloris sp. HL-130-GSB-WP 232319771.1
Prosthecochloris aestuarii-HED31708.1
Prosthecochloris sp. HL-130-GSBA-RM30038.1
Prosthecochloris sp. GSB1-WP 094081601.1
Chlorobium phaeobacteroides-MBC8525007.1
Chlorobium phaeobacteroides-MBL6956691.1
Chlorobiales bacterium-MCG8343911.1
Chlorobiales bacterium-MCG8342105.1
Chlorobiales bacterium-MCG8377797.1
Chlorobiaceae bacterium-NTU67867.1
Chlorobiaceae bacterium-NTV02537.1
Chlorobiaceae bacterium-NTU92458.1
Chlorobiales bacterium-MCE1274119.1
Chlorobiaceae bacterium-NTV25960.1
Chlorobiaceae bacterium-NTU91618.1
Chlorobiaceae bacterium-NTU59217.1
Chlorobiaceae bacterium-NTW52004.1
Chlorobiaceae bacterium-NTU53767.1
Chlorobiaceae bacterium-MBN1929018.1
Chlorobaculum parvum-HHE32497.1
Chlorobaculum parvum NCIB 8327-ACF10448.1
Chlorobaculum parvum-WP 232203910.1
Chlorobaculum tepidum-WP 010933908.1
Chlorobaculum thiosulfatiphilum-TNJ39045.1
Chlorobaculum thiosulfatiphilum-WP 246046605.1
Chlorobaculum limnaeum-AOS84668.1
Chlorobaculum limnaeum-WP 236848177.1
Chlorobiaceae bacterium-NTV92089.1
Chlorobiaceae bacterium-NTW63334.1
Chlorobium phaeobacteroides-WP 049751717.1
Chlorobium phaeobacteroides DSM 266-ABL64112.1
Chlorobi bacterium-MBM3422197.1
Chlorobiaceae bacterium-MBN1278322.1
Chlorobi bacterium-MBM3162715.1
Chlorobiaceae bacterium-NTU86929.1
Chlorobiaceae bacterium-NTW93114.1
Chlorobium limicola-WP 059139379.1
Chlorobium limicola-NTV07668.1
Chlorobium limicola-WP 012465009.1
Chlorobium limicola-NTV20467.1
Pelodictyon luteolum-KZK75020.1
Pelodictyon luteolum-WP 238974596.1
Pelodictyon luteolum DSM 273-ABB22906.1
Chlorobium phaeovibrioides-WP 149997645.1
Chlorobium phaeovibrioides-WP 151418997.1
Chlorobium phaeovibrioides-WP 126384356.1
Chlorobium phaeovibrioides-WP 126341685.1
Chlorobiaceae bacterium-NTW68406.1
Chlorobium ferrooxidans-WP 006366770.1
Chlorobium sp. KB01-WP 076789621.1
Chlorobiaceae bacterium-NMW18169.1
Chlorobiaceae bacterium-NTV06597.1
Pelodictyon phaeoclathratiforme-MBV5288817.1
Pelodictyon phaeoclathratiforme-WP 012506882.1
Chlorobiaceae bacterium-NTW83331.1
Chlorobiaceae bacterium-NTW56848.1
Chlorobiaceae bacterium-NTU96366.1
Chlorobiaceae bacterium-NTV97689.1
Chlorobiaceae bacterium-NTW11428.1

Supplementary Figure 3: Phylogenetic tree of 72 homologs with speciation events and duplications events. Red squares represent speciation events, and green squares represent duplication events.