Assignment Maria Alexandra Stanciu, student number r0776676

Comparative genomics between Caenorhabditis elegans and Daphnia Pulex

1. Introduction

This assignment has the purpose to compare the genomics of two eukaryotes organisms: Caenorhabditis elegans and Daphnia pulex. To do so, I will try to identify gene pairs of orthologs and paralogs, a phylogenetic analysis between species tree and a pair of orthologs will be done and also identification of conservation sites will be conducted. C.elegans is a transparent nematode and is a well known organism as it is used as a model in the microbiology field. The genome of C. elegans has five or six chromosomes(hermaphrodites) and about 100Mb in size out of which 26% are introns and 47% are intergenic regions. The number of protein coding genes is estimated to be 20,470. Daphnia pulex is a crustacean that can be found in many regions around the world. D.pulex has 12 chromosomes and 200Mb genome size. It has a compact genome and it is estimated to have 30,907 protein coding genes. It is known from literature that the D. pulex genome contains a lot of paralogs genes due to gene duplications which diverge in time to neofunctionalization[1].

2. Methods

2.1. Orthologs and paralogs identification

The analysis started with downloading for both organism all the protein coding genes from UNIPROT[2] database and a BLASTP[3] was performed using these two fasta files as follows: a search for homologous sequences was performed with all sequences from C. elegans against D. pulex as the database, a search for homologous sequences was performed with all sequences from D.pulex against C.elegans as a database. The output of these two searches contained each protein coding gene from one species against all the other protein coding genes and a score given by BLASP. To identify the pairs of orthologs in the two species the method best bi-directional hits(BBH) was used which states that a pair of genes are orthologs if they are found to have the best hit in both output searches. It is known that many complications can appear by identifying orthologs and one of them if arising from gene duplication: if the same gene in the most common ancestor between the two species undergo gene duplication after the speciation event, BBH would identify only one of the duplicates as being the ortholog; the genes that were duplicated after the speciation event are named in-paralogs. To overcome this problem a BLASTP search was done for each species against itself as a database and if the best score of one predicted gene from this search was higher than the BBH, it was considered a paralog and the duplicated genes were considered co-orthologs with the protein coding gene form the other species found with BBH. The BBH and paralogs identification were done with a python script(supl.1). The output of this script is a file containing a python dictionary as follows:

{SpeciesC.elegans_ortholog1: [SpeciesD.pulex_ortholog1, [SpeciesC.elegans_ortholog1,SpeciesC.elegans_paralog1,..],[SpeciesD.pulex_ortholog1,SpeciesD.pulex_paralog1,..]] }, where SpeciesC.elegans_ortholog1 and SpeciesD.pulex_ortholog1 are the pair of orthologs and [SpeciesC.elegans_ortholog1,SpeciesC.elegans_paralog1] and [SpeciesD.pulex_ortholog1,SpeciesD.pulex_paralog1,..] are paralogs for each protein coding gene in each species.
Scripts and files used for this method can be found in archive Supl_1 .

2.2. Speciation/duplication on single gene identification

To analyse speciation/duplication events along evolution between the two species a pair of co-orthologs was selected in order to create a gene tree.The protein coding genes selected were: P02566(Myosin-4), P02567(Myosin-1) from C.elegans that were found to be paralogs and E9FZS8, E9FZS9 also identified as paralog pair of genes from D.pulex.The genes P02566, P02567 are co-orthologs with E9FZS8(Myosin heavy chain isoform 3), E9FZS9(Myosin heavy chain isoform 1). A BLASTP was performed for P02566 and E9FZS8 to search for homologs genes in other organisms for these 2 sequences. A selection of 25 different species was performed and the paralogs for each species were added to the search output. Next step was to create a multiple sequence alignment with CLUSTALW2[4] and a phylogenetic tree was created using CLUSTALW2 with bootstrap 1000 as supporting value. To identify the duplication/speciation events that occurred which resulted in the selected pair of co-orthologs genes a species tree was created as well. To create the species tree ribosomal RNA was used for each of the 25 species.The 18S rRNA was downloaded for each species from SILVA database[5] and CLUSTALW2 was used to create a multiple sequence alignment and a phylogenetic tree with bootstrap 1000 as supporting value. To visualize the trees created iTOL[6] was used.
Scripts and files used for this method can be found in archive Supl_2 .

2.3. Functional regions identification

To be able to identify functional regions in the protein coding genes I searched for conserved sites in the multiple alignment created by CLUSTALW2 for the gene tree because, usually the sites that are conserved throughout evolution play an important role in the function of the protein. To identify these sites I created a python script which has as an input the multiple sequence alignment and as output a file with a dictionary of form: {column_number: [score, [residues]]}. I used the method Simpson Diversity to calculate the score for each column and I kept only the columns that had the score less than 0.2 and did not contain any gaps. Simpson Diversity has the following formula : $D = 1 - \sum_{i=1}^{S} n_i(n_i - 1)/N(N-1)$, where D is the score, n_i is the number of residues of type i, S is the unique number of residues and N is the total number of aligned sequences.
Scripts and files used for this method can be found in archive Supl_3.

3. Results

3.1. Orthologs and paralogs identification

The number of protein coding genes found for C.elegans was 27,344 and the protein coding genes for D.pulex identified was 32,304. The number of orthologs found was 6,163. For C. elegans 2,862 genes from the orthologs identified were found to have at least one paralog and for D.pulex 2531 from the orthologs identified were found to have at least one paralog. Although less genes from D.pulex were found to have paralogs, the ones that were found have many paralogs.

3.2. Speciation/duplication on single gene identification

The phylogenetic trees created can be seen in Fig.1 and Fig.2 (better resolution Supl.2). By comparing the two trees we can identify the events which led to these genes being co-orthologs. In Fig.2 the number “1” in red represents a duplication event since the two paralogs of C. elegans (P02566, P02567) are separated by this event. The number “2” in red represents a speciation in which C.elegance gene P02567 is evolving in a separate clade from D.pulex from this common ancestor on. At node “3” (the red number in Fig.2) we can see a duplication event as gene E9FZS8 is evolving separately from E9FZS9 which is formed after another duplication event(not marked in red as the other nodes discussed).

At this point we have the set of paralogs from C.Elegans co-orthologs with D. pulex as identified by best bi-directional hits as well. As we can see in Fig. 2 although, the genes E9FZS8, E9FZS9 are not in-paralogs since they do not have the same common ancestor, but E9FZS9 is in-paralog with E9FZT0. For this analysis only these genes were selected to create the phylogenetic tree, but there were many genes in D.pulex which showed very high homology and this is in accordance with available literature saying that the D.pulex genome evolved by means of gene duplication as well.

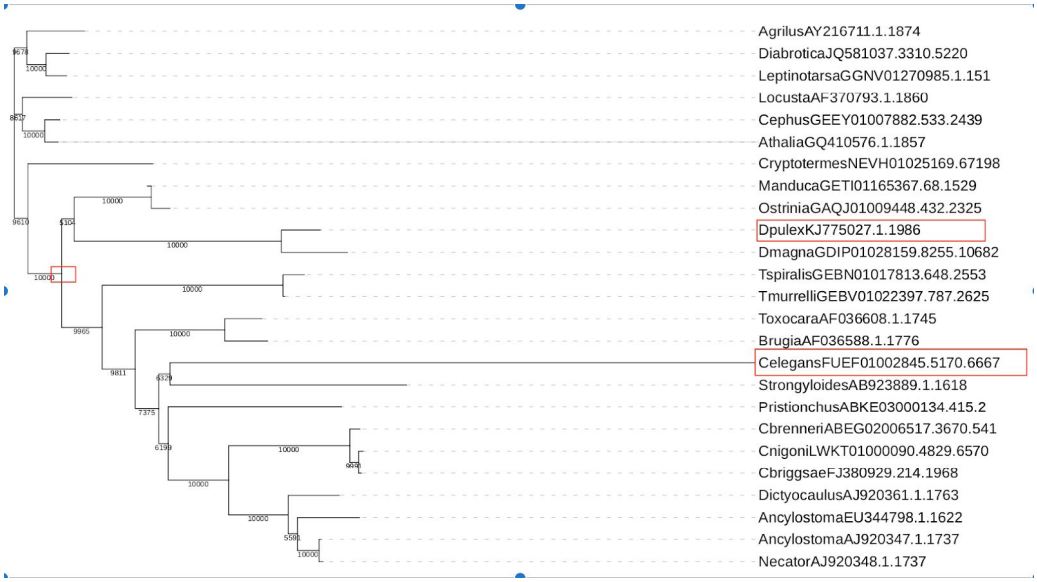

Fig.1 Species Tree. The ones in red color are the two species analysed.

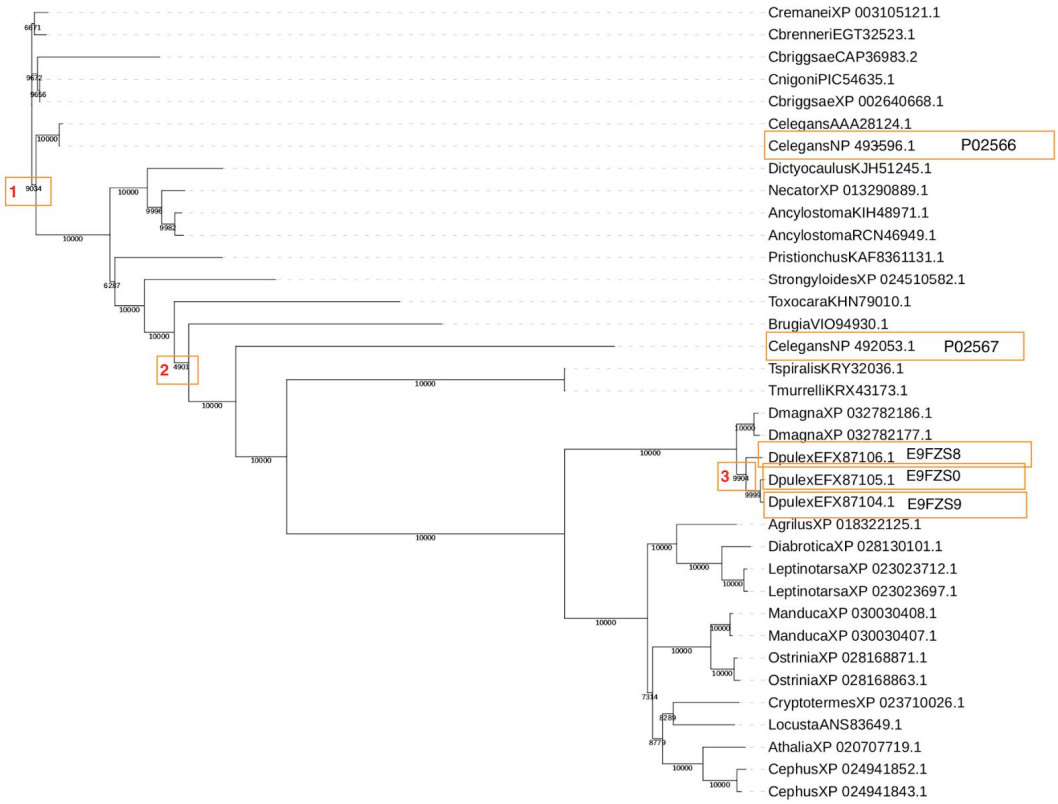

Fig.2 Gene tree. The ones in orange are the co-orthologs pair of genes chosen to be analysed.

3.3. Functional regions identification

By using a cut-off of 0.2 as maximum Simpson Diversity score there were found 915 residues to fall under this constraint out of 2071 columns in the alignment. By looking at the proportion of conserved sites from the total number of residues that are forming the proteins and considering the species selected for the alignment that is not very wide (nematodes, crustaceans, arthropoda) we could expect to have a high number of conserved residues. Since C. elegans is a very well studied organism, being a model organism we could investigate protein regions found to be important for C.elegans and infer their functions in other organisms, like D. pulex. By searching E9FZS8 in the UNIPROT database we can see that this protein is inferred from homology.

In C. elegans, the gene P02566 encodes for Myosin-4 and P02567 for Myosin-1, both proteins being part of a bigger complex named myosin heavy chain involved in muscle contraction. Also from NCBI we know that Myosin-4 protein also called Uncoordinated protein 54 interacts with chaperone unc-45 and ubiquitin-protein ligase ufd-2 to form a complex. We also see that a specific region of the protein is identified to be the motor domain(Myosin head) and this domain is part of a protein family which is highly conserved. For C. elegans this motor domain was identified to be in the region 88-775. By checking how many conserved sites we have in this region, we find 347 residues with no gaps and almost identical. We could statistically verify if

these conserved sites are significant and if yes we could infer that D.duplex protein E9FZS8 has a motor domain.
I also found from NCBI that C. elegans protein binds ATP at sites: 122-130 ,174-181, 231-241, 460-465 and if we check how many conserved sites are found between 122-130, there are 5 identical sites. Also between 231-241 positions there are 6 sites with scores less than 0.2. This looks promising in identifying ATP binding sites in the other homologs proteins identified.
This analysis is far from being complete, since these sites found on NCBI are specific for each organism and they should be counted differently. Also the region 231-241 from C.Elegans corresponds to a part of protein sequence "GESGAGKT" identified to play a role in ATP binding and if a search is done in the protein alignment file we can see that this region is perfectly conserved for all the proteins in the alignment, but one CAP23983.2 from C.briggsae species. This conserved region was not identified by the conservation rule applied here, so maybe better refinements should be done and statistical analysis should confirm the significance of these findings.

4. Discussion

Although this comparative genomic analysis focused mainly on identifying orthologs and what can be inferred from one gene by looking at its homologs, we can conclude that it is a good starting point to a complete analysis. Just by looking at the genomes at the two species C. elegans and D. pulex we can say that they differ on chromosome number and that D.pulex has at least 10 000 protein coding genes more than C. elegans and also a larger genome. There were around 6000 pairs of genes identified as orthologs suggesting the evolution of the two species, but the orthologs identification is a very complex task due to gene duplication, gene loss, gene transfer, domain shuffling and this is still an intensive studied domain[7]. The phylogenetic analysis for both species tree created in 18S rRNA and gene tree underlined the speciation and duplication events for one specific gene, Myosin-4 which is involved in muscle contraction. This analysis suggested many duplication events in D.pulex organisms and also an earlier duplication event for C.elegans before speciation between the two species happened. By refining even more the comparison, conserved sites pointed out the existence of a motor domain in D.pulex gene (E9FZS8) and also the conservation region for ATP binding suggesting that E9FZS8 gene found to be ortholog with Myosin-4 in C.elegans could have related functions. To be able to draw clear conclusions a more in depth analysis has to be performed, but these findings can be a good start.

Exercise 4
A.
I managed to find the promoter of C. elegans gene Myosin-4(unc-54, P02566) here
https://epd.epfl.ch/cgi-bin/get_doc?db=ceEpdNew&format=genome&entry=unc%2D54_1
annotated as unc-54_1. The workflow used to identify the promoters implies an experimental step and is described here:
https://epd.epfl.ch/epdnew/documents/Pf_epdnew_001_pipeline.php
The sequence is: ATGATGCGAGTGATGAGTGCCTGCTTCTTCCTCGGCAGGAGGATGTTGGCACCCGCAAGC
Although I tried to find a similar sequence around ortholog gene Myosin heavy chain isoform 1

in D.pulex I haven't managed(E9FZS8). This paper : https://www.genetics.org/content/204/2/593.long looks promising in discovering the gene promoter in D. pulex, but I haven't had enough time to investigate it further.

Note: all the supplementary files and the source code can be found on github: https://github.com/mariaalexandrastanciu/ComparativeGenomics

References:

[1] Colbourne, J. K. et al. The ecoresponsive genome of *Daphnia pulex*. *Science* **331**, 555–561 (2011).
doi: 10.1126/science.1197761

[2] The UniProt Consortium, UniProt: a worldwide hub of protein knowledge, *Nucleic Acids Research*, Volume 47, Issue D1, 08 January 2019, Pages D506–D515, *https://doi.org/10.1093/nar/gky1049*

[3] NCBI Resource Coordinators, Database resources of the National Center for Biotechnology Information, *Nucleic Acids Research*, Volume 44, Issue D1, 4 January 2016, Pages D7–D19, *https://doi.org/10.1093/nar/gkv1290*

[4] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–4680.
doi: 10.1093/nar/22.22.4673

[5] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl. Acids Res. 41 (D1): D590-D596.*

[6] Interactive Tree Of Life (iTOL) v4: recent updates and new developments
Nucleic Acids Research, Volume 47, Issue W1, 02 July 2019, Pages W256–W259, *https://doi.org/10.1093/nar/gkz239*

[7] Tekaia F. Inferring Orthologs: Open Questions and Perspectives. Genomics Insights. January 2016.
doi: 10.4137/GEI.S37925