Pol Xavier Pérez Rodríguez
Roger Bosch Unamuno

Abstract

In this report we will see the analysis and research of a protein sequence with an unknown gene, function, and specie. We are interested in inferring the specie, and doing a functional annotation, including the Go terms, a description, and the conserved Pfam domains, among others.

By doing a blast search we obtained a list of similar proteins (homologous) and we took the best reciprocal hit (*Thecamonas trahens*) and we searched for information about it in different databases. It corresponds to an ATP synthase subunit delta which has the function, together with other proteins of this family, of transporting protons across a membrane to generate an electrochemical gradient that powers ATP synthesis.

Performing a multiple sequence alignment and with the homologous sequences, we were able to obtain information about how the protein has evolved by looking at the gene tree of our protein and the species tree.

We finally concluded that this protein is highly conserved between species in terms of gene order and no gene losses, horizontal gene transfer, or duplications. As the species tree and gene tree are the same, all the sequences are orthologs between them. So the evolutionary history tells us that the function of this protein plays an important role for the organisms that have it.

Pol Xavier Pérez Rodríguez
Roger Bosch Unamuno

Introduction

In this report we study a de novo protein sequence and we give a broad view of its function and how it has evolved.

Organism	**Protein**	**Per. identity**	**Query Cover**
Thecamonas trahens	XP_013762613 (ATP synthase subunit delta)	65.97%	81%
Trichoplax adhaerens	XP_002114008.1 (ATP-synt_DE_N domain)	46.15%	51%
Monopterus albus	XP_020467879.1 (ATP synthase subunit delta)	38.95%	53%
Anabas testudineus	XP_026198866.1 (ATP synthase subunit delta)	41.30%	51%
Kyptolebias marmoratus	XP_017262604 (ATP synthase subunit delta, mithochondrial)	43.48%	51%
Helobdella robusta	XP_009031646.1 (ATP-synt_DE_N domain)	43.33%	50%

In the figure above is shown the most similar sequences to ours, the first one having an identity of the 65.97% and it is the best hit. We suppose that the function and family of our protein are going to be the same as its best homologs, this implies that our sequence belongs to a part from an ATP synthase. This family of proteins carries out the transport of protons across a membrane to generate an electrochemical gradient that powers ATP synthesis. Basically, the molecular function of our protein is able the synthesis of ATP from ADP and phosphate by the transfer of protons from one side of a membrane to the other driven by the proton-motive force according to the reaction ADP + H2O + phosphate + H+(in) -> ATP + H+(out) (this reaction corresponds to hydrolysis). This membrane can be the plasma membrane or the mitochondrial inner membrane. Our protein corresponds to the Delta/Epsilon chain (beta-sandwich domain) which is a part of the head unit of the ATP synthase. The subunit is called epsilon in bacteria and delta in mitochondria. ATP synthases are composed of two linked complexes: the F1 ATPase complex is the catalytic core and is composed of 5 subunits and our chain is here, while the F0 ATPase complex is a membrane-embedded proton channel that is composed of at least 3 subunits.

Using *Trichoplax adhaerens* genome we searched for protein XP_002114008 and we found that this protein is codified by a gene called TRIADDRAFT_64074 located between positions 2,206,326..2,207,876 (length: 1,551 nt, and gene position: 790). This gene is not surrounded by similar genes that have the same function.

Pol Xavier Pérez Rodríguez
Roger Bosch Unamuno

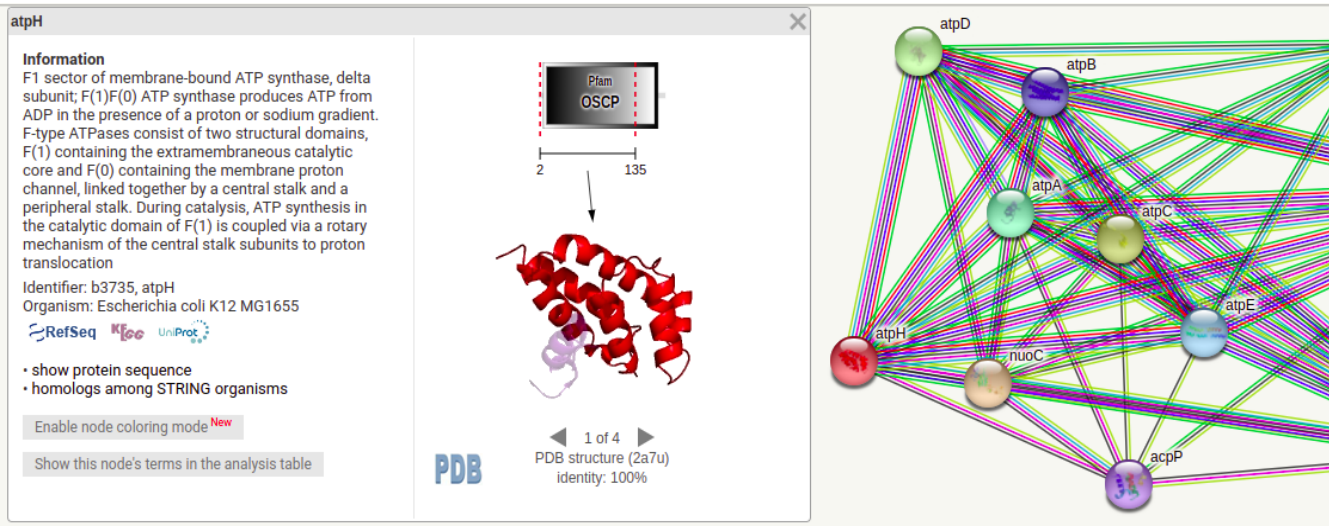
Materials and Methods

Information about functional annotation was obtained by doing a blast search and taking the best reciprocal hit (*Thecamonas trahens*) and searching for information about it in different databases. The description was done using Uniprot, Pfam, Genome Data Viewer (with *Trichoplax adhaerens*, because *Thecamonas trahens* was not available and this is the second-best reciprocal hit of our sequences), and using the go therms associated with the protein family that our protein belongs to which explains the biological, molecular, and cellular functions were obtained with Interpro.

To obtain information about how the protein has evolved we first need to make a multiple sequence alignment with the homologous sequences from the species of our selection (*Thecamonas trahens, Trichoplax adhaerens, Anabas testudineus, Monopterus albus, Kryptolebias marmoratus*) from the blast search and our protein. To obtain better results we must also take a sequence that is more distantly related to our target (*Fonticula alba* "outgroup"). Once we put all the sequences together in a fasta file, using MAFFT we can obtain the multiple sequence alignment and a tree in Newick format. To visualize the tree we use Phylo.io.

Now that we are able to see the evolutionary relationships between our sequences, we need the species. The paralogy and orthology can be inferred with two different methods: species overlap and species reconciliation. The species overlap does not need the species tree, so as our tree is composed of different species, the conclusions observed would be that sequences are orthologs between each other. The species reconciliation needs the species tree so we are able to search for duplication and loss events, so this is the method that we decided to use. After reconstructing our gene/protein tree with new speciation/duplication/loss events, we are able to search for orthologs and paralogs. If the last common ancestor between two sequences is a duplication event they will be paralogs, and if the last common ancestor is a speciation event they will be considered orthologs.

To look for gene order conservation for this protein we will use "Protein association networks (String)" and we will search "ATP synthase subunit delta".



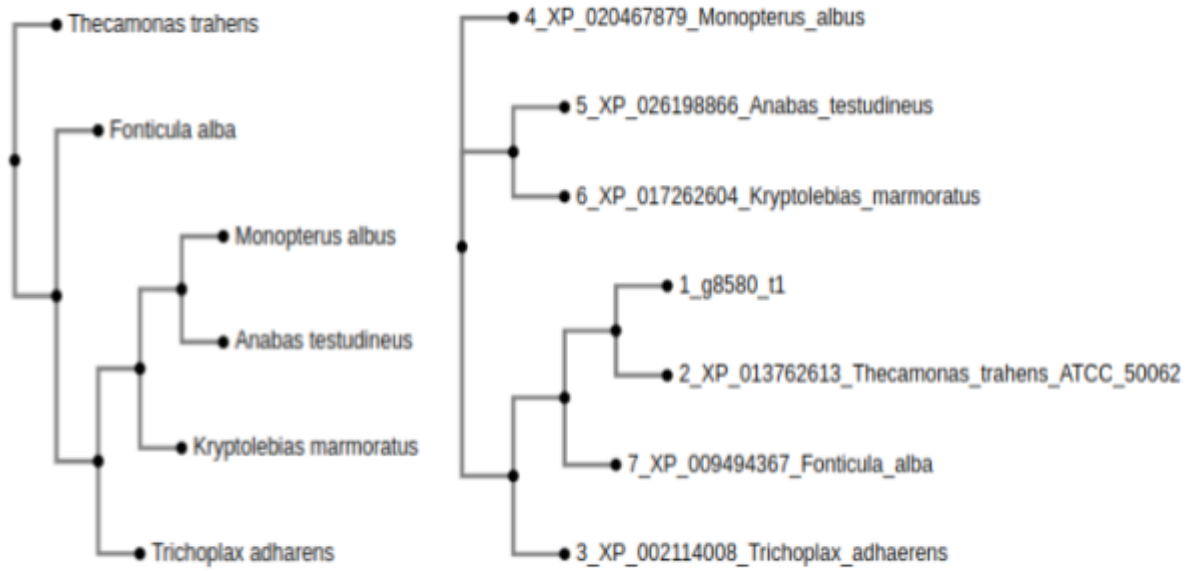
This network shows genes that are known to interact with our gene of interest. As we can see in the figure above all the genes that interact with our input (atpH) are the other subunits of "ATP synthase".
In settings, we will select experiments + co-ocurrence + neighborhood in basic settings sources. This will make remain the species that have similar interactions among the proteins shown in the network. Then we will switch to the neighborhood view and this will show us the gene conservation among the different organisms.

Results and Discussion

In *Archaea* and *Eukaryota* seem that there is no information available. As we can see in the figure above, in Bacteria the gene order is highly conserved between the species because the blocks of synteny are equally ordered between species.

The first thing that we tried to do to build the gene tree was to search for species trees that contained the species that we picked. We did not find any tree that contained those species but we found one that contained our specie, but when we were trying to get the sequences of the species that appeared in the species tree we only found *Thecamonas trahens* (our specie). As a final measure, we build the species tree of our sequences based on the Taxonomy found on the taxonomy browser.

Pol Xavier Pérez Rodríguez
Roger Bosch Unamuno



Species tree on the left side and gene tree on the right side. Comparing the species tree with the gene tree we did not find any differences, this means that they have the same number of sequences, no duplications, and same topology (identical trees).

Conclusion

Our proteins belongs to the ATP synthase subunit delta. ATP synthase carries out the transport of protons across a membrane to generate an electrochemical gradient that powers ATP synthesis. About its evolutionary history, with all the results shown before, we are able to conclude that the gene that codifies our protein is conserved in species (is not lost or duplicated) and it is very stable in terms of gene order. This implies that the role that plays this protein in the organisms in which is present is important even though the aminoacids belonging to that protein between the different species are not that much conserved.

References

https://www.sciencedirect.com/science/article/pii/S1434461010000246?via%3Dihub
https://academic.oup.com/mbe/article/29/4/1277/1196360?login=true
https://www.ebi.ac.uk/interpro/result/InterProScan/iprscan5-R20220613-152951-0637-21304867-p2m/
https://www.uniprot.org/uniprot/A0A0L0D914
https://pfam.xfam.org/family/PF02823.19#tabview=tab0
https://mafft.cbrc.jp/alignment/software/
http://orthology.phylomedb.org/search#content_title
https://string-db.org/cgi/geneneighbors?taskId=bxVdiEVzKLY0&sessionId=biv2oCcQgJaI