A human-planarian interologs network to annotate planarian transcriptomes: PlanNET

Computational Genomics Lab · Universitat de Barcelona · I→B=U→B

Castillo-Lara, Sergio; Abril, Josep F

SUMMARY

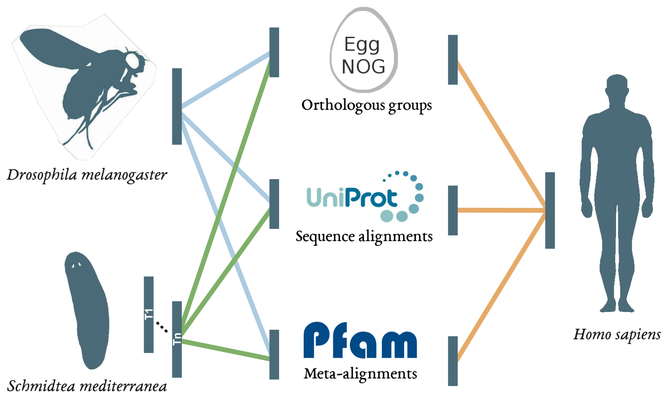
Regeneration and tissue renewal are essential processes of adult animals that must be tightly controlled, since its dysfunction leads to common illness as degenerative diseases or cancer. Planarians are emerging as a model organism to study regeneration in animals mainly due to the presence of a population of adult pluripotent stem cells, called neoblasts, a cell type able to produce all of the cellular lineages that conform these worms. However, the little available data of protein-protein interactions hinders the advances in understanding the mechanisms underlying its regenerating capabilities.

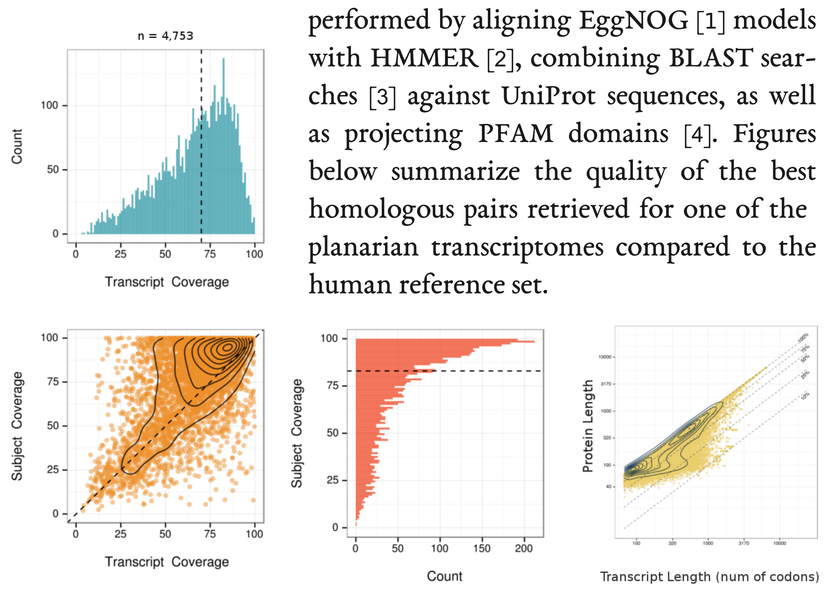
We have developed a protocol to predict protein-protein interactions using sequence homology data and a reference Human interactome. This methodology was applied on ten Schmidtea mediterranea transcriptomic sequence datasets and it has been automated to integrate future transcriptomic data for this species. We projected each network into a graph-based database manager, as interactions data can be queried much efficiently on such type of databases. On top of that we have deployed a web application, called PlanNET, to explore the multiplicity of networks and the associated sequence annotations.

PlanNET is available at `https://compgen.bio.ub.edu/PlanNET`

HOMOLOGY SEARCH

For each of the available *S. mediterranea* transcriptomes, we looked for the best homologous protein in Human. This search was then repeated to find *Drosophila melanogaster* transcripts-Human proteins pairs. These sequence associations were used to predict possible protein-protein interactions.

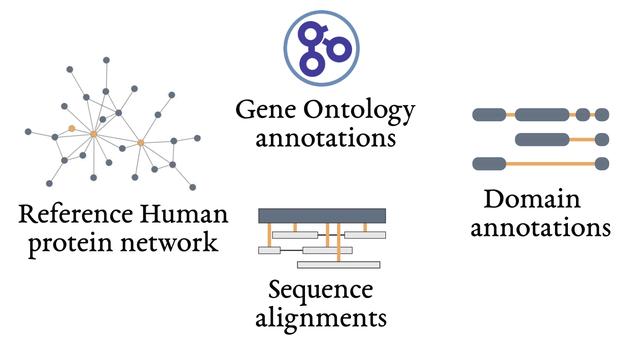




Functional annotation of all transcripts was performed by aligning EggNOG [1] models with HMMER [2], combining BLAST searches [3] against UniProt sequences, as well as projecting PFAM domains [4]. Figures below summarize the quality of the best homologous pairs retrieved for one of the planarian transcriptomes compared to the human reference set.

PREDICTION

FEATURE SELECTION

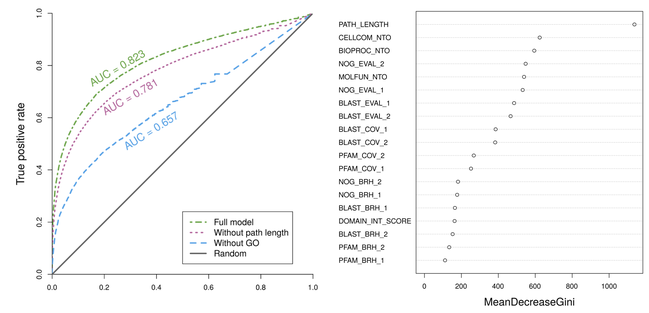


- *Sequence similarity* to human homologs
- *Shortest path* in human network between homologs
- *GO similarity* [5] between homologs
- Number of *interacting domains* on sequence

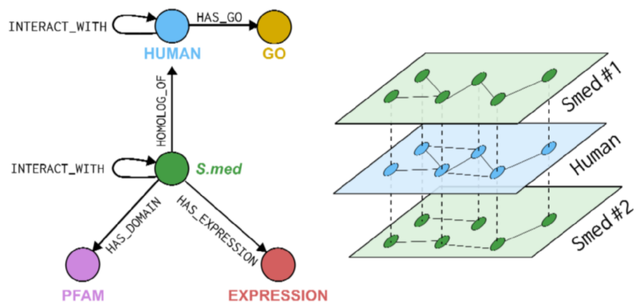
⋮

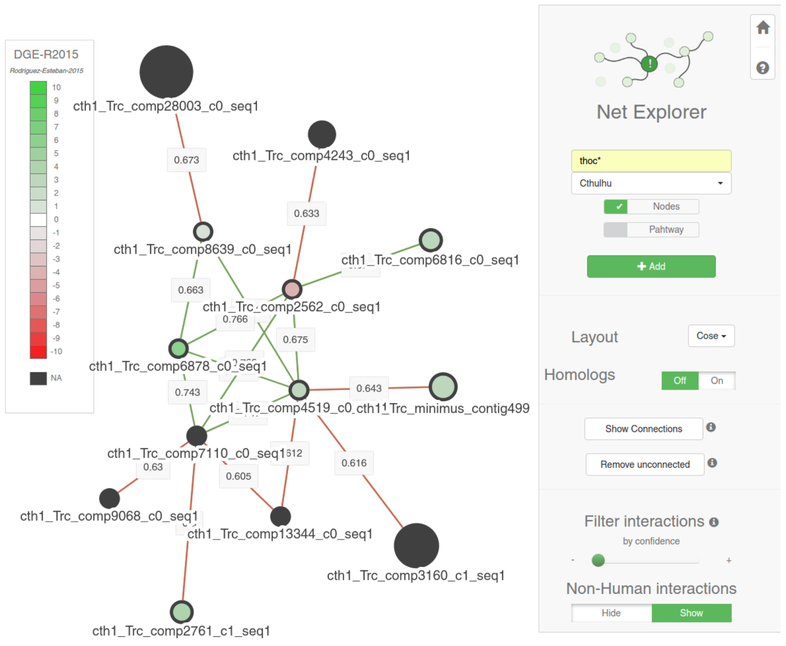
RANDOM FOREST

After collecting the features for each possible pair of transcripts, a Random Forest Classifier [6] was trained by using the *Drosophila melanogaster* data. The model performance was assessed by using Out of Bag estimates of precision, recall and Area under the ROC curve. On the bottom left figure we have the ROC curve of the full model, and of several models recomputed by removing the most important features (bottom right).



PLANNET WEB APPLICATION





DATABASE

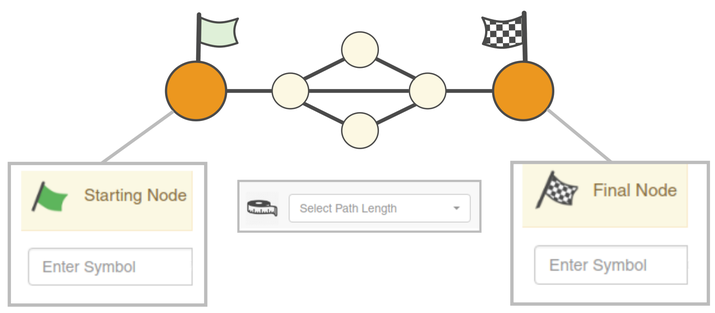
In order to store each one of the 10 planarian transcriptomes projected over the Human interactome we used the database manager Neo4j. This graph-based manager allows us to model these projections in a way that facilitates very complex queries to be performed in short time. Then, all transcriptomes end up connected through the Human reference interactome layer.

NET EXPLORER

This application allows users to explore the predicted protein-protein networks. Users can compare specific pathways across the different transcriptomes or compare them to the human reference network. Moreover, net explorer allows researchers to map gene expression data from different RNA-seq experiments on the nodes for an integrated visualization.

PATHWAY FINDER

Instead of looking for particular proteins and all their interactions like in net explorer, pathway finder allows researchers to look for specific pathways by selecting a starting node, a final node and a specific path length. Thanks to the flexibility of the database manager Neo4j, these queries can be performed quickly.



created with neo4j · django · python

REFERENCES

[1] Huerta-Cepas, J. et al., NAR, 44: D286–D293, 2016.

[2] Eddy, S, Bioinformatics, 14: 755–763, 1998.

[3] Camacho C., et al., BMC Bioinformatics 10:421, 2008.

[4] Punta, M. et al., NAR, 40: D290–D301, 2011.

[5] Mistry, M. et al., BMC Bioinformatics 9: 327, 2008.

[6] Andy Liaw et al., R News, 2(3): 18-22, 2002.

ACKNOWLEDGEMENTS

This work was supported by:
- Spanish Ministry of Economy (BFU2014-56055P).
- Generalitat de Catalunya (2014SGR687).
- Predoctoral fellowship by AGAUR (FI-FDR, 2017FI_B_00191).

RECOMB-CG 2017 | Sergio Castillo-Lara s.cast.lara@gmail.com *and* Josep F. Abril jabril@ub.edu | https://compgen.bio.ub.edu