A Computational Pipeline for Investigating Protein Pathway Diversification Post-Genome Duplication

Laura McDonnell (ANDREW ID: lmcdonne)

1 Project Idea

In this project, I want to develop a computational workflow that integrates comparative phylogenetic analysis and structural modeling to investigate the evolution of protein pathways following whole genome duplication (WGD) events. By combining gene tree reconciliation, network analysis, and protein structural prediction, the workflow will allow me to examine the process of how duplicated genes diverge and acquire new functions. I will design the methodology to be broadly applicable, providing a more generalizable framework for studying evolutionary dynamics across different protein families. This approach will allow me to systematically analyze the impact of genomic duplications on pathway diversification, and possibly uncover novel insights into protein function and interaction evolution.

The project will require the use of publicly available genomic and proteomic datasets, along with using established methods discussed in class and in the literature while integrating new computational strategies. Initial validation of the workflow will be conducted using the Myostatin pathway in salmonids as a case study, but the end goal is to create a computational pipeline that can be applied to different systems. The analysis will involve constructing phylogenetic trees to analyze orthologous and paralogous relationships, followed by structural modeling to predict changes in protein binding interfaces that result from gene duplication. This approach is expected to improve understanding of the interplay between genomic events and functional evolution. And thus, offers a valuable tool for future evolutionary genomics research.

2 Software to Write

- **Data Preprocessing:** Scripts to download, clean, and format genomic and protein interaction datasets for downstream analysis
- **Phylogenetic Analysis Tool:** Custom software for aligning sequences, constructing gene trees, and performing reconciliation analyses to differentiate between orthologs and paralogs
- **Structural Modeling Interface:** A set of programs or scripts to interface with existing structural prediction tools (such as PyMOL, Rosetta) to assess changes in protein structure and binding sites

3 Papers to Read

- [2]
- [3]
- [1].

References

[1] K. Chen, D. Durand, and M. Farach-Colton. Notung: a program for dating gene duplications and optimizing gene family trees. *Bioinformatics*, 18(Suppl 1):S22–S29, 2000.

[2] M. Lynch and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, 2000.

[3] J. Zhang. Evolution by gene duplication: an update. *Trends in Ecology and Evolution*, 18(6):292–298, 2003.