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

Laura McDonnell (ANDREW ID: lmcdonne)

## 1 Introduction

Protein pathways evolve through complex mechanisms that include gene duplication events and subsequent divergence in both structure and function. Whole genome duplication (WGD) events are particularly influential, as they replicate entire genomes, creating paralogous gene copies that have the potential to develop new functions or regulatory roles [Birchler and Yang, 2022]. This process not only leads to redundancy but can also spawn entirely new pathways by diversifying the interactions among proteins. As shown in Figure1, the fate of these duplicated genes varies: one copy may become nonfunctional—a process known as nonfunctionalization—while in other cases, one copy may acquire a novel function (neofunctionalization) or both copies may partition the original function between them (subfunctionalization). Understanding these different fates and their effects is fundamental to understanding how new biological functions and complex regulatory networks arise, making this topic of broad interest to computational biologists and evolutionary geneticists alike.

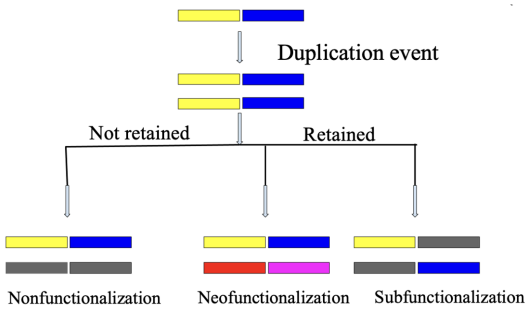


Figure 1: Possible fates of genes following a Whole Genome Duplication (WGD) event. The original gene functions are shown in blue and yellow. Post duplication, genes may have loss of function and become nonfunctional (shown in gray), OR can be retained, becoming neofunctional, i.e., develop new functions (shown in pink and red), or subfunctional, i.e., original gene functions are split between the new copies.

This project aims to develop a comprehensive computational workflow that integrates comparative phylogenetic analysis with structural modeling techniques to study protein pathway diversification. This integrative approach involves constructing and reconciling gene trees to differentiate between orthologous and paralogous relationships, followed by detailed sequence analysis and protein structure prediction to assess the functional consequences of divergence. By applying these methods to the Myostatin pathway in salmonids—a model system where WGD has led to notable evolutionary divergence—the workflow will be rigorously tested and validated. Myostatin, a key protein in the TGF-Beta signaling pathway, functions as a skeletal muscle growth inhibitor [Elkina et al., 2011]. The signaling pathway is shown in Figure2. Disruption of myostatin, such as through gene knockout experiments, results in pronounced muscle hypertrophy, underscoring its pivotal role in muscle regulation. In the salmonid lineage, multiple consecutive WGD events have provided a unique opportunity to examine the evolutionary outcomes of gene duplication in a naturally occurring context. Given the central regulatory role of myostatin, this pathway serves as an excellent model system to explore whether the mechanisms that maintain duplicated genes—via neofunctionalization or subfunctionalization—lead to pathway diversification and increased specificity of molecular interactions.

At the core of this investigation is the hypothesis that the mechanisms maintaining duplicated genes after WGD in the salmonid lineage—whether through nonfunctionalization, neofunctionalization, or subfunctionalization—have driven diversification within the myostatin pathway. Specifically, the aim to answer the question: Did the evolutionary processes following consecutive WGD events in salmonids lead to pathway diversification and increased specificity of molecular interactions in the myostatin pathway? By combining gene tree reconciliation, sequence analysis, and protein structure prediction, this project seeks to illuminate the intricate relationships between genomic events and subsequent structural adaptations. The insights gained from this study could not only enhance our understanding of protein pathway evolution but also inform strategies in drug design, synthetic biology, and evolutionary studies where predicting protein function is important.

To elucidate the impact of whole genome duplication (WGD) on gene duplication and divergence, a comparative analysis was conducted using species that represent a gradient of duplication events shown in Figure3. The mouse genome, which contains a single copy of the target gene, serves as a baseline for evolutionary conservation. In contrast, the danio rerio genome, harboring two copies, and the salmon genome, with four copies resulting from consecutive WGD events, provide a natural framework to examine the incremental effects of gene duplication. This selection enables a detailed investigation into how increasing WGD events contribute to functional specialization, pathway diversification, and structural divergence among gene copies.

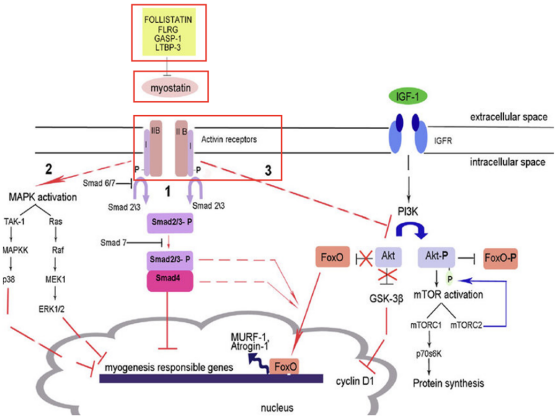


Figure 2: Myostatin is a protein that functions as a skeletal muscle growth inhibitor and is a part of the TGF-Beta Signaling Pathway. [Elkina et al., 2011] Key interacting partners are outlined in red.

Understanding the evolutionary fate of duplicated genes is essential for unraveling the complexities of functional innovation and adaptation. Previous studies have highlighted the importance of gene duplication as a driving force in evolution ([Ehrenreich, 2020] & [Magadum et al., 2013]), yet many questions remain regarding the link between structural alterations and functional outcomes in protein products. Thus, this project seeks to illuminate the intricate relationships between genomic events and subsequent structural adaptations. The insights gained from this study could not only enhance our understanding of protein pathway evolution but also inform strategies in fields such as drug design and synthetic biology. The evolution of protein pathways following WGD is of broad interest to computational biologists because it connects molecular evolution, structural biology, and systems biology. By understanding how duplicated genes diverge, insights can be gained into the emergence of novel biological functions, the adaptation to new environments, and even the origins of complex diseases. This topic has implications for drug design, synthetic biology, and evolutionary studies, making it a critical area of inquiry in computational genomics.

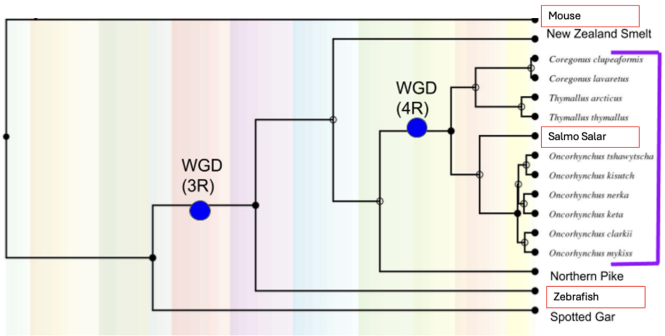


Figure 3: The species tree of interest, with two recent WGD's marked on the branches and salmonids marked by the purple bracket

## 2 Methods

The approach is divided into three key phases: sequence analysis, phylogenetic reconstruction, and structural modeling and interaction analysis. The workflow consists of the following steps: (1) homologous sequences of myostatin and its interacting partners across multiple species are identified using BLAST [Altschul et al., 1990] and (2) these findings are refined through multiple sequence alignment with MAFFT [Katoh et al., 2005]. Next, evolutionary relationships are reconstructed by (3) estimating maximum likelihood phylogenies with PhyML [Guindon et al., 2010], followed by (4) gene tree reconciliation using Notung [Chen et al., 2000] to distinguish between orthologous and paralogous relationships. Finally, (5) structural divergence is assessed by aligning protein structures with FATCAT [Ye and Godzik, 2004] and (6) simulating species-specific protein-protein interactions using Replica Dock [Harmalkar et al., 2022]. Together, these

methods provide a comprehensive framework to investigate whether gene duplication in the salmonid lineage has driven pathway diversification and enhanced interaction specificity.

### 2.1 Sequence Analysis

To reconstruct the evolutionary history of the myostatin pathway, the analysis begins with the identification of homologous sequences. TBLASTN is used to query known myostatin genes and their interacting partners against the genomes of mouse, zebrafish, and salmon. For each homolog candidate, the expectation value (E-value) $E(s)$ is defined and sequences satisfying $E(s) < E_{threshold}$ are retained, with $E_{threshold} = 1 \times 10^{-10}$. The set of high-confidence homologs is denoted by:

$$S = \{s \in \text{Candidates } | E(s) < E_{threshold}\}$$

Next, multiple sequence alignments for each gene family are performed using MAFFT. The resulting alignment is represented as a matrix $A$ where $A_{ij}$ is the residue at position $i$ in sequence $j$. This alignment is critical for subsequent phylogenetic inference.

### 2.2 Phylogenetic Reconstruction

For phylogenetic reconstruction, PhyML is used to estimate maximum likelihood (ML) trees. The likelihood $L$ of an evolutionary model given the alignment data $X$ is computed as:

$$L(\theta|X) = \prod_{i=1}^{n} P(X_i|\theta)$$

where $n$ is the number of sites in the alignment, $X_i$ denotes the data at site $i$, and $\theta$. represents the evolutionary model parameters (e.g., substitution rates, branch lengths). Model selection is performed by comparing likelihood scores using criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to determine the best-fit model.

Finally, gene trees are constructed and reconciled with Notung software, using the known species tree to distinguish between orthologous and paralogous relationships. This reconciliation process incorporates evolutionary events like gene duplication and loss, and it aids in annotating branches with events corresponding to nonfunctionalization, neofunctionalization, or subfunctionalization.

### 2.3 Structural Modeling and Interaction Analysis

The second phase of the approach focuses on the structural consequences of gene duplication within the myostatin pathway. For each protein of interest, structural models are either retrieved from public databases or predicted de novo. Then, structural alignments are performed using FATCAT, which computes the optimal superposition of protein structures. The degree of structural divergence is quantified using the root-mean-square deviation (RMSD), defined as:

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^{N} ||r_i - r_i'||^2}$$

where $N$ is the number of aligned atoms, and $r_i$ and $r_i'$ are the coordinates of atom $i$ in the two structures being compared.

Following structural alignment, species-specific docking is applied using Replica Dock. This step simulates protein-protein interactions between myostatin and its interacting partners. Docking scores and predicted binding energies serve as quantitative metrics for interaction specificity and strength. These metrics will be correlated with phylogenetic divergence and structural changes to evaluate whether gene duplication has led to increased pathway diversification and interaction specificity.

## 3 Implementation Details

Genomic and proteomic data for this project are obtained from NCBI. Query sequences corresponding to myostatin and its interacting partners are used to search the available proteomes via TBLASTN 2.12.0+. The resulting BLAST hits are filtered using an E-value threshold of $1 \times 10^{-10}$, and the high-confidence matches for each query in each genome are saved in FASTA format.Subsequently, multiple sequence alignments are

performed using MAFFT with the BLOSUM62 substitution matrix to ensure accurate alignment of the sequences. After alignment, the files are converted and saved in PHYLIP format, which serves as the input for downstream phylogenetic inference using PhyML. PhyML 3.0 is employed to estimate maximum likelihood trees, with the software automatically selecting the best-fit evolutionary model based on likelihood scores (using AIC and BIC criteria) and then employing the model with the highest likelihood for the final tree estimation. Gene tree reconciliation is performed using Notung 2.9.1.5 to incorporate the known species tree, enabling the distinction between orthologous and paralogous relationships and the annotation of events such as nonfunctionalization, neofunctionalization, or subfunctionalization.

In addition to the tools already described, FATCAT 2.0 and Replica Dock 2.0 are incorporated into the workflow for the structural analysis steps. FATCAT 2.0 is used for flexible protein structure alignment, allowing an accurate assessment of the degree of structural divergence between protein models. Replica Dock 2.0 is employed to perform species-specific protein-protein docking simulations, providing quantitative measures of interaction strength and specificity between myostatin and its interacting partners. Both tools were selected for their enhanced performance and improved interfaces for automated batch processing. Default settings are used unless otherwise specified, and any parameter adjustments are documented within the custom Python scripts to ensure full reproducibility.

The entire pipeline is automated through custom Python scripts that integrate each tool and manage data flow between steps, ensuring reproducibility. Below is a pseudocode snippet outlining the overall workflow:

```
for each query in myostatin_interactors:
    results = TBLASTN(query, target_genomes, evalue_threshold=1e-10)
    save_results_as_fasta(results)

for each gene_family_fasta in fasta_files:
    alignment = MAFFT(gene_family_fasta, matrix="BLOSUM62")
    save_alignment_as_phylip(alignment)

for each phylip_file in phylip_files:
    best_model = PhyML(phylip_file, model_selection="auto")
    ml_tree = PhyML(phylip_file, model=best_model)
    reconciled_tree = Notung(ml_tree, species_tree)

for each protein in reconciled_tree:
    structure = retrieve_or_predict_structure(protein)
    aligned_structures = FATCAT(structure, reference_structure)
    rmsd = calculate_RMSD(aligned_structures)
    docking_results = ReplicaDock(protein, interacting_partner)
    log_results(rmsd, docking_results)
```

All software versions (PhyML 3.0, Notung 2.9.1.5) and parameter settings are documented within the scripts to facilitate full reproducibility by others in the field.

# References


[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410.

[Birchler and Yang, 2022] Birchler, J. A. and Yang, H. (2022). The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *The Plant Cell*, 34(7):2466–2474.

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

[Ehrenreich, 2020] Ehrenreich, I. M. (2020). Evolution after genome duplication. *Science*, 368(6498):1424–1425.

[Elkina et al., 2011] Elkina, Y., von Haehling, S., Anker, S. D., and Springer, J. (2011). The role of myostatin in muscle wasting: an overview. *J Cachexia Sarcopenia Muscle*, 2(3):143–151.

[Guindon et al., 2010] Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phyml 3.0. *Systematic Biology*, 59(3):307–321.

[Harmalkar et al., 2022] Harmalkar, A., Mahajan, S. P., and Gray, J. J. (2022). Induced fit with replica exchange improves protein complex structure prediction. *PLOS Computational Biology*, 18(6):1–21.

[Katoh et al., 2005] Katoh, K., Kuma, T., Toh, H., and Miyata, T. (2005). Mafft version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, 33:511–518.

[Magadum et al., 2013] Magadum, S., Banerjee, U., Murugan, P., et al. (2013). Gene duplication as a major force in evolution. *J Genet*, 92:155–161.

[Ye and Godzik, 2004] Ye, Y. and Godzik, A. (2004). Fatcat: A web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Research*, 32:W582–W585.