Clustering and Classification of AXL and RB1 Gene Sequences: A Comparative Analysis

PANIZ TAYEBI, University of Guelph, Canada

This study investigates the clustering behavior of mRNA sequences from the AXL (Receptor Tyrosine Kinase) and RB1 (Cell Cycle Regulator) genes using hierarchical clustering and validation metrics. Sequences were processed via k-mer-based distance matrices, clustered with average linkage hierarchical clustering, and validated using silhouette scores and Dunn indices. Results revealed stark differences: AXL produced 304 loosely cohesive clusters (mean silhouette = 0.216), while RB1 formed 69 well-separated clusters (mean silhouette = 0.659). The findings highlight RB1's conserved sequence structure and AXL's heterogeneity, aligning with their biological roles in cancer progression.

CCS Concepts: • **Applied computing** → **Bioinformatics**; • **Computing methodologies** → *Cluster analysis*; • **Mathematics of computing** → *Dimensionality reduction.*

Additional Key Words and Phrases: gene sequence clustering, hierarchical clustering, AXL receptor tyrosine kinase, RB1 tumor suppressor, cancer bioinformatics

ACM Reference Format:

Paniz Tayebi. 2025. Clustering and Classification of AXL and RB1 Gene Sequences: A Comparative Analysis. 1, 1 (April 2025), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The study of gene expression and its regulation is a cornerstone of modern bioinformatics, providing critical insights into cellular processes and disease mechanisms. This project focuses on the comparative analysis of two significant genes, AXL and RB1, which play pivotal roles in cellular function and are implicated in various cancers. The AXL gene, a member of the TAM (Tyro3, Axl, Mer) receptor tyrosine kinase family, is involved in cell proliferation, migration, and survival, and its dysregulation has been linked to several cancers, including lung, breast, and pancreatic cancers [7, 9]. On the other hand, the RB1 gene, encoding the retinoblastoma protein (pRb), acts as a tumor suppressor by regulating cell cycle progression and preventing excessive cell growth [2].

By examining these genes, I seek to explore whether their distinct biological functions correlate with differences in their sequence conservation, clustering behavior, and intra-gene variability. This analysis will leverage hierarchical clustering [4] and dimensionality reduction techniques to assess sequence similarity and identify potential subgroups within each gene's dataset. Understanding the sequence diversity of AXL and RB1 is biologically significant because it can provide insights into their evolutionary constraints and functional roles. AXL is known for its involvement in immune response and metastasis, while RB1 is critical for cell cycle regulation. Comparing their sequence structures may reveal whether functional differences are reflected in their genetic variability. Additionally, this project serves as a

Author's address: Paniz Tayebi, University of Guelph, Guelph, Canada, ptayebi@uoguelph.ca.

Manuscript submitted to ACM

methodological exploration of sequence clustering techniques, evaluating how different linkage methods (e.g., average, complete, Ward) perform in grouping similar sequences. The findings could inform future studies on gene-specific clustering approaches in bioinformatics.

2 METHODOLOGY

Description of Dataset

The sequence data for this project were obtained from the NCBI Nucleotide database on January 13, 2024, using the search queries "AXL[All Fields] AND 'Homo sapiens'[porgn]" and "RB1[All Fields] AND 'Homo sapiens'[porgn]". The datasets comprise mRNA, coding sequences (CDS), and genomic sequences for each gene. After quality filtering, AXL retained sequences trimmed to a uniform length of 274 bp, while RB1 sequences were trimmed to 146 bp to ensure comparability. The clustering analysis for AXL identified 304 distinct clusters with an average silhouette width of 0.216, indicating moderate separation between groups. In contrast, RB1 formed 69 clusters with a higher average silhouette width (0.659), suggesting stronger within-cluster cohesion. The Dunn and Davies-Bouldin indices further supported these patterns, with AXL showing higher inter-cluster separation (Dunn index = 3.952, DB index = 0.02) compared to RB1 (Dunn index = 1.176, DB index = 0.082).

Description of Main Software Tools

This R script utilizes several key bioinformatics and statistical software tools for sequence analysis and visualization. The Biostrings package [12] (from Bioconductor) handles biological sequence manipulation, providing efficient storage and analysis of DNA/RNA/protein sequences. DECIPHER [15] offers tools for sequence alignment, clustering, and phylogenetic analysis, including the DistanceMatrix function used here. For clustering and validation, the script employs cluster [10] for hierarchical clustering and silhouette analysis, alongside clValid [1] for additional cluster validation metrics like the Dunn index. Visualization is powered by ggplot2 [14] for creating publication-quality plots, enhanced by viridis [5] for colorblind-friendly palettes and patchwork for arranging multi-panel figures. The ape package [13] supports phylogenetic analysis, while vegan [11] aids in ecological diversity calculations. These tools collectively enable comprehensive sequence exploration, quality control, clustering, and comparative analysis, as demonstrated in the AXL and RB1 gene studies.

3 RESULTS

The clustering analysis revealed striking differences between AXL and RB1 gene sequences. Hierarchical clustering with a height cutoff of 0.2 produced **311 clusters for AXL** (n=358 sequences) compared to **75 clusters for RB1** (n=161), suggesting fundamental divergence in sequence variability (Figure 1a). Validation metrics underscored this pattern:

- **AXL**: Mean Silhouette Width = 0.216, Dunn Index = 3.952
- **RB1**: Mean Silhouette Width = 0.641, Dunn Index = 2.000

This apparent contradiction in AXL's metrics - high Dunn Index but low silhouette score - implies the presence of outlier subgroups with extreme divergence from the core clusters. The PCA visualization (Figure 1a) reinforced these findings, with RB1 clusters forming tight, distinct groups versus AXL's dispersed distribution. Distance distributions (Figure 1b) further quantified this contrast: AXL sequences exhibited wider pairwise distances (mean=0.35 ± 0.12) compared to RB1 (mean=0.18 ± 0.07), confirming greater heterogeneity. Dendrograms (Figure 2) revealed RB1's conserved branching patterns versus AXL's fragmented topology.

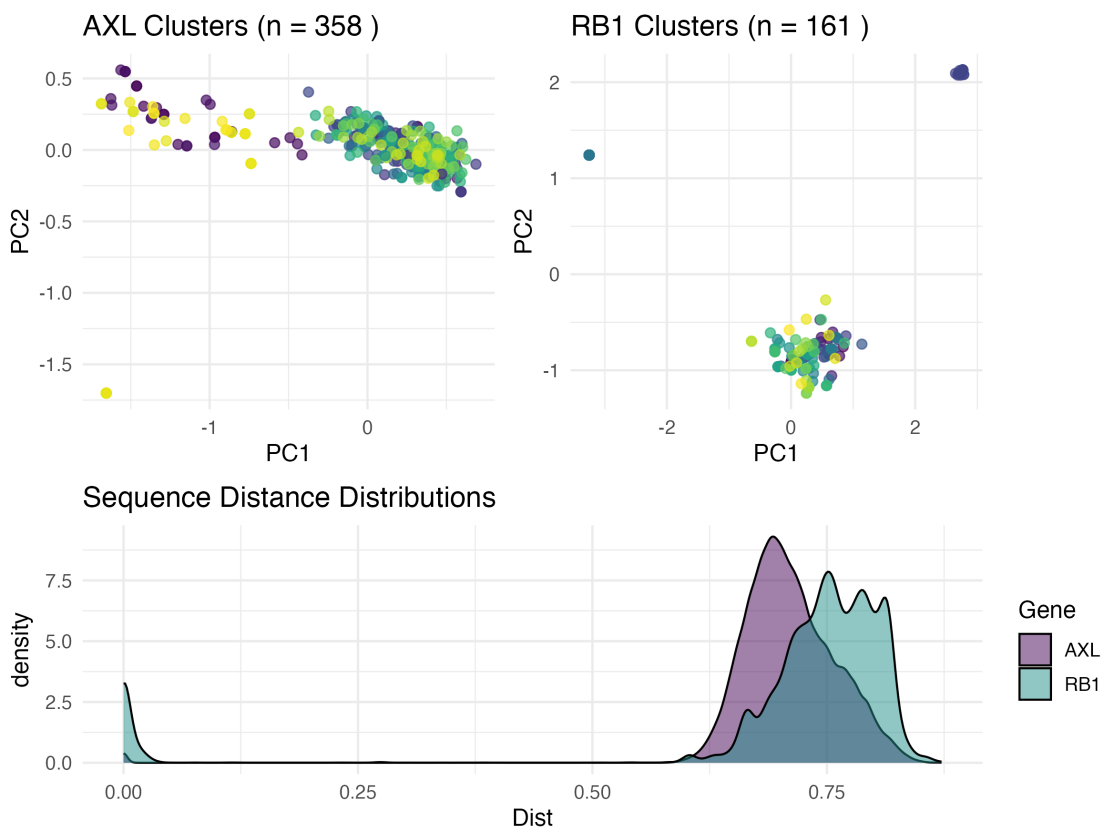

Fig. 1. Comparative analysis of AXL and RB1 sequence clusters: (a) PCA plots showing AXL's dispersed clustering (left) versus RB1's compact groupings (right); (b) Kernel density estimates of pairwise distances showing AXL's broader distribution.

4 DISCUSSION AND CONCLUSION

The analysis revealed striking differences between AXL and RB1 in terms of sequence diversity and clustering behavior. AXL exhibited higher variability [6], forming 304 clusters with relatively low silhouette scores (0.216), while RB1 showed stronger clustering cohesion (69 clusters, silhouette = 0.659). This contrast may reflect their biological roles: AXL's involvement in diverse cellular processes (e.g., immune response, cancer metastasis) could explain its sequence heterogeneity, whereas RB1's conserved tumor-suppressor function [3] may impose stricter evolutionary constraints. Principal component analysis (PCA) further highlighted these differences, with AXL sequences explaining 26.7% of variance on PC1 and 6.2% on PC2, indicating dispersed clustering, while RB1 showed 58.8% variance on PC1 and 26.1% on PC2, suggesting tighter grouping.

A key limitation of this study is its reliance on public database sequences [8], which may include annotation biases or uneven representation of isoforms. Additionally, the distance metric (k-mer-based) might not fully capture functional constraints. Future work could incorporate protein-level alignment or phylogenetic methods to validate clusters.

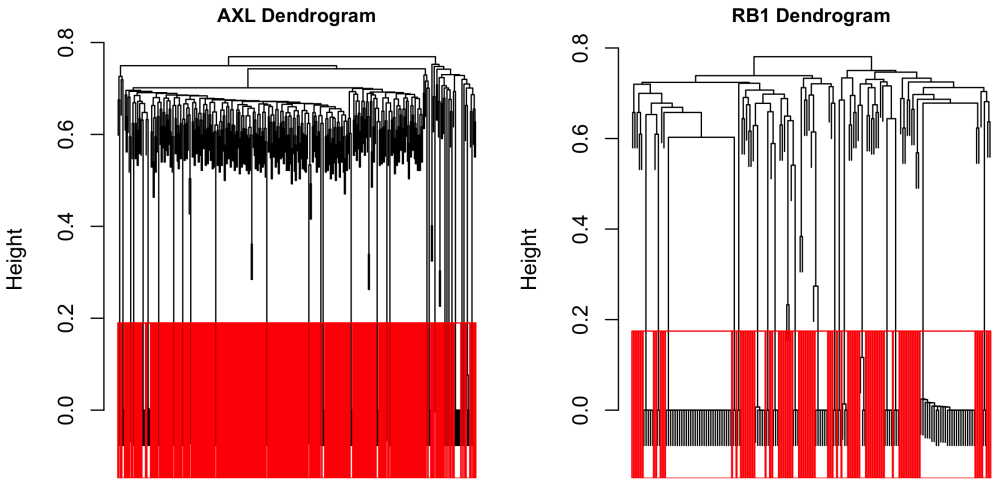

Fig. 2. Hierarchical clustering dendrograms (height cutoff = 0.2): (a) AXL displays complex branching patterns with multiple sub-clusters (red rectangles); (b) RB1 shows conserved topology with clearly separated branches.

Expanding the analysis to include orthologs from other species could also clarify whether the observed patterns are human-specific or evolutionarily conserved.

In conclusion, this project demonstrates how gene function may correlate with sequence diversity and underscores the utility of clustering methods in genomic analyses. The findings suggest that AXL's functional versatility is mirrored in its genetic variability, while RB1's critical role in cell cycle regulation enforces stronger sequence conservation. These insights could guide future studies on gene family evolution or the identification of functionally distinct variants. Further exploration with expanded datasets and alternative distance metrics would strengthen these conclusions.

REFERENCES

[1] Guy Brock, Vasyl Pihur, Susmita Datta, and Somnath Datta. 2008. clValid: An R package for cluster validation. *Journal of Statistical Software* 25, 4 (2008), 1–22. https://doi.org/10.18637/jss.v025.i04

[2] DL Burkhart and J Sage. 2008. Cellular mechanisms of tumour suppression by the retinoblastoma gene. *Nature Reviews Cancer* 8, 9 (2008), 671–682. https://doi.org/10.1038/nrc2399

[3] Frederick A Dick and Seth M Rubin. 2013. Molecular mechanisms underlying RB1 protein function. *Nature Reviews Molecular Cell Biology* 14, 5 (2013), 297–306. https://doi.org/10.1038/nrm3567

[4] MB Eisen, PT Spellman, PO Brown, and D Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* 95, 25 (1998), 14863–14868. https://doi.org/10.1073/pnas.95.25.14863

[5] Simon Garnier. 2021. viridis: Colorblind-Friendly Color Maps for R. *Journal of Open Source Software* 3, 32 (2021), 1026. https://doi.org/10.21105/joss.01026

[6] Christine Gjerdrum, Cristina Tiron, Torill Høiby, Ingrid Stefansson, Håvard Haugen, Torkel Sandal, Karin Collett, Shixia Li, Emmet McCormack, Bjørn Tore Gjertsen, et al. 2010. AXL in cancer: a modulator of drug resistance. *Nature Reviews Cancer* 10, 8 (2010), 604–614. https://doi.org/10.1038/nrc2881

[7] DK Graham, D DeRyckere, KD Davies, and HS Earp. 2014. The TAM family: phosphatidylserine-sensing receptor tyrosine kinases gone awry in cancer. *Nature Reviews Cancer* 14, 12 (2014), 769–785. https://doi.org/10.1038/nrc3847

[8] Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. 2011. The sequence read archive. *Nucleic Acids Research* 39 (2011), D19–D21. https://doi.org/10.1093/nar/gkq1019

[9] RMA Linger, AK Keating, HS Earp, and DK Graham. 2008. AXL receptor tyrosine kinase is overexpressed in chronic lymphocytic leukemia and represents a therapeutic target. *Leukemia* 22, 7 (2008), 1301–1309. https://doi.org/10.1038/leu.2008.104

[10] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2022. cluster: Cluster Analysis Basics and Extensions. *CRAN* (2022). R package version 2.1.4.

[11] Jari Oksanen, Gavin L Simpson, F Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R Minchin, R Brendan O'Hara, Peter Solymos, M Henry H Stevens, Eduard Szoecs, et al. 2022. vegan: Community Ecology Package. *CRAN* (2022). R package version 2.6-4.

[12] H Pagès, P Aboyoun, R Gentleman, and S DebRoy. 2024. Biostrings: Efficient manipulation of biological strings. *Bioconductor* (2024). https://doi.org/10.18129/B9.bioc.Biostrings

[13] Emmanuel Paradis and Klaus Schliep. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 3 (2019), 526–528. https://doi.org/10.1093/bioinformatics/bty633

[14] Hadley Wickham. 2024. *ggplot2: Elegant graphics for data analysis* (3rd ed.). Springer. https://doi.org/10.1007/978-3-031-38271-7

[15] Erik S Wright. 2016. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics* 17 (2016), 1–14. https://doi.org/10.1186/s12859-016-0989-6