A Shuffle-Based Statistical Approach for Robust Pseudogene Annotation

Pedro M. Barcelos[1] *, **Marcos Catanho**[2] †, **Antônio B. de Miranda**[2]
Edward H. Haeusler[1] ‡, **Sergio Lifschitz**[1] §

[1]Departamento de Informática
Pontifícia Universidade Católica do Rio de Janeiro (PUC-RIO)
Rio de Janeiro – Brazil

[2]Laboratório de Genética Molecular de Microrganismos – Instituto Oswaldo Cruz
Fundação Oswaldo Cruz – Rio de Janeiro – Brazil

{pbarcelos,hermann,sergio}@inf.puc-rio.br

{mcatanho,amiranda}@fiocruz.br

Abstract. *The accurate annotation of pseudogenes is a significant challenge in genomics, as their decaying sequences often fall into a "twilight zone" of similarity that confounds automated methods. This paper describes a robust, homology-based methodology designed to overcome this issue. The core of the approach is a shuffle-based statistical evaluation used to establish a custom, empirically-derived significance threshold. This allows for the confident discrimination of true, biologically significant sequence remnants from stochastic background noise, providing a reliable framework for annotating pseudogenes and unannotated coding sequences in large-scale genomic projects.*

1. Introduction

The annotation of genomes is a foundational step in modern biological research, yet it is often incomplete. Pseudogenes, deactivated, remnant copies of functional genes, are particularly difficult to identify accurately. As they are theoretically free from selective pressure, they accumulate mutations that can obscure their similarity to their functional counterparts [Rost 1999]. Standard annotation pipelines, which rely on predefined gene structure models, often misidentify or completely overlook these molecular fossils. The propagation of such errors can have a compounding effect on downstream comparative and functional genomic analyses. This paper builds upon a previously established proof of concept, further developing and applying the methodology to a well-characterized prokaryotic model to demonstrate its broad utility

*Pedro Marçal Barcelos, Sergio Lifschitz and Edward H. Haeusler are partially funded by CAPES (Brazilian Ministry of Education Funding Agency)

†The authors acknowledge the National Laboratory for Scientific Computing (LNCC/MCTI, Brazil) for providing HPC resources of the SDumont supercomputer, which have contributed to the research results reported within this paper. URL: http://sdumont.lncc.br

‡Edward H. Haeusler is partially funded by FAPERJ grant APQ1 E-26/210.258/2019.249292 and CNPq grant 309287/2023-5

§Sergio Lifschitz is partially funded by the National Council for Scientific and Technological Development (CNPq), grant 306525/2021-6

Current computational methods for pseudogene annotation fall largely into two categories: feature-based and homology-based [Zheng et al. 2007]. The first search for intrinsic aberrations in functional codons that could disrupt protein processing. The second approach is the most common, identifying the pseudogenes by their similarity to known functional proteins. However, the major limitation in both cases is the arbitrary similarity thresholds (e.g., generic E-values). This leads to a critical problem: the criteria used to define a significant hit are often not uniform across different studies, resulting in inconsistent and frequently incomplete annotations [Xiao et al. 2016]. This lack of a standardized, statistically-grounded foundation for significance motivates the need for a more objective approach.

To address this, we developed a genome-wide comparative methodology that does not rely on pre-existing gene models. Instead, it employs a customized statistical filter to confidently identify genomic sequence regions that exhibit significant sequence similarity to (computationally predicted or experimentally verified) functional protein sequences. This approach allows for the systematic and automated discovery of both preserved and degraded protein-coding sequences genome-widely.

This paper is organized as follows: Section 2 details the methodology, outlining the multi-stage pipeline from statistical baseline establishment to the annotation of protein-coding sequence remnants, with particular emphasis on creating a robust null model and the empirically derived significance threshold. Section 3 describes the strategy for applying this methodology to Escherichia coli str. K-12 substr. MG1655, including the null model analysis, the genome-wide search, and the statistical validation steps. Finally, Section 4 presents the findings, summarizing our shuffle-based statistical approach's advantages and broad applicability for robust pseudogene and protein-coding gene annotation

2. The Annotation Methodology

The methodology, already established as a proof of concept [Abrahim et al. 2022], is structured as a multi-stage pipeline that begins with the establishment of a statistical baseline for sequence similarity and concludes with the annotation of putatively functional and non-functional remnants of protein-coding sequences.

The core of this methodology is the creation of a robust null model to empirically define the boundary between biologically meaningful similarity and random background noise. To achieve this, we first generated an artificial dataset of 10,000 protein sequences by shuffling a template sequence that matched the average amino acid composition and length of a large reference set of functional proteins. These shuffled sequences retain the statistical properties of real proteins but lack biological meaning. Next, we performed an all-versus-all BLAST comparison, aligning the functional proteins against the artificial dataset to define a "noise profile".

From this analysis, we established a multi-part significance threshold derived directly from the data, rather than relying on arbitrary cutoffs . An alignment was deemed significant only if it met three simultaneous criteria: (1) an E-value $\leq 1 \times 10^{-6}$, a value set to be more stringent than the best E-value observed in any random alignment (1.7×10^{-6}); (2) a raw alignment score > 120, a threshold determined because the maximum score against any artificial sequence was 119; and (3) a sequence identity $> 20\%$, a conserva-

tive baseline chosen because it represents the expected chance identity between random protein sequences while remaining below the zone where distant evolutionary relationships can still be detected. This empirical approach ensures that only alignments that significantly exceed the random noise profile are considered for further analysis as shown in Figure 1.

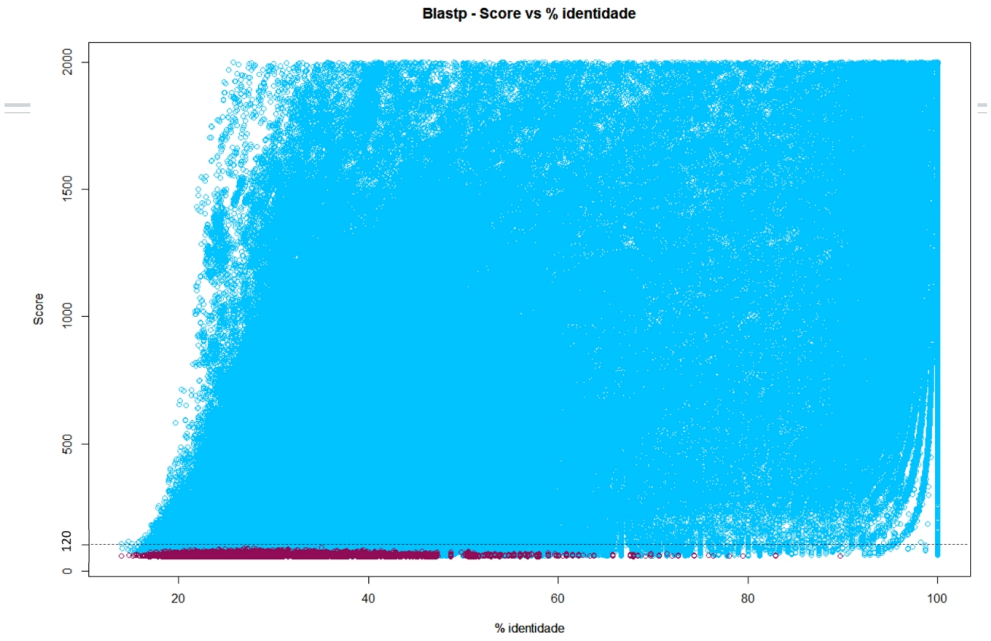

Figure 1. Functional vs artificial scatter plot. Scatter plot displaying the score and percent identity of pairwise alignments obtained comparing (i) functional proteins among each other (blue dots) and comparing these proteins with artificially created sequences with the average size and composition displayed by the group of functional sequences (red dots).

As a proof of concept, this methodology was successfully applied in a collaborative study examining the genomes of three human parasites: Leishmania major, Trypanosoma brucei, and Trypanosoma cruzi [Abrahim et al. 2022]. In that work, systematic searches uncovered thousands of pseudogenes, "molecular corpses" of once-living genes damaged by random mutations, and hundreds of cryptic protein-coding regions (CDS) not identified in the original annotation processes. These newly found elements presented distinct characteristics in each trypanosomatid regarding their mutation profile, abundance, and genomic density.

The results demonstrated that scanning genomes with functional proteins as proxies, relying on a custom threshold to distinguish significant similarities, and reassembling remnant sequences from their debris constitute a suitable and robust strategy to improve the accuracy and completeness of genomic annotations.

3. Our Proposed Approach

To further explore and apply this methodology to a well-characterized prokaryotic system, the same pipeline is being executed on the reference genome of *Escherichia coli* str. K-12 substr. MG1655[Blattner et al. 1997] [1]. The analysis of the null model for the functional

[1]NCBI Assembly Accession: GCF_000005845.2

protein dataset of UniProtKB/Swiss-Prot 2025 from UniProtKB [2] revealed that the best alignment between a real protein and a random sequence yielded a maximum score of 100 and the best E-value of 2.98×10^{-6}.

The analysis of the null model for the functional protein dataset of UniProtKB/Swiss-Prot 2025 from UniProtKB, [3] [uni 2025] [4] revealed that the best alignment between a real protein and a random sequence yielded a maximum score of 100 and the best E-value of 2.98×10^{-6}. This data-driven result allows us to confidently set a significance threshold well above the mark for random noise, for instance $E - value \leq 1 \times 10^{-6}$, $score > 100$, $identity > 20\%$.

The second step consists of a genome-wide search using the `tblastn` [Camacho et al. 2009] algorithm of the program `blast-2.15.0+` [5] . This search compares the complete Swiss-Prot proteome against the *E. coli* genome, using a permissive E-value threshold. The Expect-value (E-value) is a statistical measure representing the number of alignments with a given score or better that one would expect to find purely by chance when searching a database of a particular size. It is calculated based on the alignment score, the query, the database size and statistical parameters derived from the scoring matrix used. Therefore, a lower E-value indicates a more statistically significant match, suggesting the observed alignment is highly unlikely to occur randomly.

The goal of this phase is not to find final results, but rather to capture any DNA region that has even the slightest similarity to a known protein, ensuring that no potential pseudogene is missed. This initial stage generates millions (3 909 811) of low-confidence hits. The E-value distribution of these hits is shown in Figure 2, where a massive peak of over 2.5 million non-significant hits demonstrates the necessity of a stringent filtering before proceeding to the sequence similarity and statistical validation step.

The third step is the validation of each of these hits. For each candidate alignment from tblastn, a rigorous realignment is performed with ssearch, implementing the Smith-Waterman dynamic programming algorithm [Smith et al. 1981, Pearson 2000]. The statistical significance of each alignment is then assessed through a shuffling process with 500 iterations. Only alignments whose score exceeds the significance threshold established by the null model are retained. This process filters the millions of initial hits, resulting in a final, small, high-confidence dataset. The preliminary results of this analysis in E. coli will serve to refine the pipeline and demonstrate its universal applicability across different domains.

We acknowledge that this high-sensitivity approach, which generates millions of initial candidates before a computationally intensive validation step, presents a scalability challenge, particularly for larger eukaryotic genomes. While feasible for prokaryotic genomes on standard computing clusters, applying this method to gigabase-scale genomes may require significant computational resources. Future implementations could incorpo-

[2] Available at: `https://ftp.ebi.ac.uk/pub/databases/uniprot/current_release/`

[3] published on Wed Apr 23 2025 (573,230 sequence entries)

[4] Available at: `https://ftp.ebi.ac.uk/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz`

[5] Available at: `https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ncbi-blast-2.16.0+-x64-linux.tar.gz`

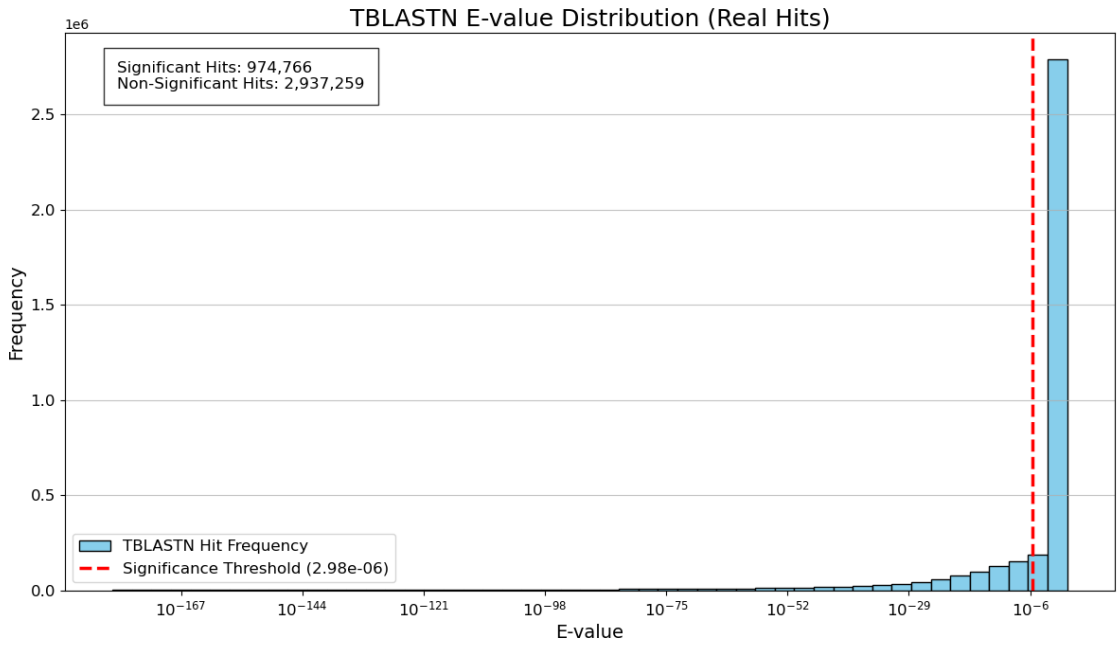

Figure 2. E-value Distribution of High-Sensitivity Search Hits. The histogram displays the frequency distribution of E-values from the initial tblastn search against the *E.coli* genome. The x-axis is on a logarithmic scale to visualize the wide range of values. The red dashed line indicates the significance threshold (E-value = 2.98e-06) determined from the null model analysis.

rate strategies such as pre-filtering based on k-mer indexes or leveraging hardware acceleration to mitigate this bottleneck, ensuring its applicability across diverse genomic scales.

4. Conclusions

The methodology provides a robust and scalable solution for annotating protein-coding genes and pseudogenes in genomes. By establishing an empirically-derived statistical threshold based on a custom null model, this approach effectively minimizes the inclusion of false positives that arise from spurious but still statistically significant (low E-value) alignments. This technique is not dependent on specific gene-finding algorithms and can be readily applied to any sequenced genome, enhancing the accuracy and completeness of genomic annotations.

The successful application of this methodology and the ongoing work in *E.coli*, underscore its versatility and effectiveness. Our preliminary results show that scanning genomes with functional proteins as proxies and a customized statistical threshold provides a powerful strategy for identifying well-preserved and highly degraded protein-coding sequences. This allows for a more comprehensive understanding of genome architecture and evolution, moving beyond reliance on pre-existing, and often incomplete, gene models. The precise identification of pseudogenes, in particular, offers valuable insight into gene decay pathways and the evolutionary history of gene families.

Future work will focus on two key areas: refinement and scalability. We will refine the statistical model by exploring how factors such as genomic GC content and

varying amino acid compositions influence the null model, allowing for the dynamic adjustment of significance thresholds in diverse genomic contexts. This could involve machine learning approaches to automate the optimization of these parameters. To address scalability, we will investigate computational optimizations to reduce the burden of the realignment step, such as parallelized implementations and algorithmic shortcuts for filtering low-confidence hits more efficiently. We also plan to integrate our pipeline with other annotation tools, creating a comprehensive system for de novo genome annotation. Finally, a critical next step is to investigate the functional implications of the newly identified pseudogenes, particularly their potential regulatory roles and contributions to genome plasticity.

References

(2025). Uniprot: the universal protein knowledgebase in 2025. *Nucleic acids research*, 53(D1):D609–D617.

Abrahim, M., Machado, E., Alvarez-Valín, F., de Miranda, A. B., and Catanho, M. (2022). Uncovering pseudogenes and intergenic protein-coding sequences in tritryps' genomes. *Genome Biology and Evolution*, 14(10):evac142.

Blattner, F. R., Plunkett III, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., et al. (1997). The complete genome sequence of escherichia coli k-12. *science*, 277(5331):1453–1462.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). Blast+: architecture and applications. *BMC bioinformatics*, 10(1):421.

Pearson, W. R. (2000). Flexible sequence similarity searching with the fasta3 program package. In *Bioinformatics methods and protocols*, pages 185–219. Springer.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein engineering*, 12(2):85–94.

Smith, T. F., Waterman, M. S., et al. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197.

Xiao, J., Sekhwal, M. K., Li, P., Ragupathy, R., Cloutier, S., Wang, X., and You, F. M. (2016). Pseudogenes and their genome-wide prediction in plants. *International Journal of Molecular Sciences*, 17(12):1991.

Zheng, D., Frankish, A., Baertsch, R., Kapranov, P., Reymond, A., Choo, S. W., Lu, Y., Denoeud, F., Antonarakis, S. E., Snyder, M., et al. (2007). Pseudogenes in the encode regions: consensus annotation, analysis of transcription, and evolution. *Genome research*, 17(6):839–851.