ENHANCING E. COLI GENOMIC ANALYSIS WITH RETRIEVAL-AUGMENTED GENERATION

Chugh Kritika
School of Natural Sciences and Mathematics
University of Texas at Dallas
Richardson, TX 75080, USA
kritikachugh2@gmail.com

ABSTRACT

This study presents a framework that leverages retrieval-augmented generation (RAG) to enhance the interpretation and analysis of complex bioinformatics data in Escherichia coli (E. coli) genomics. By integrating bioinformatics tools including pairwise alignment, NCBI annotation, multiple sequence alignment (MSA) with large language models (LLMs) such as GPT-3-mini, Gemini 2.0 Advanced Flash Thinking Experimental model, and Grok 3, our approach combines real-time data retrieval with dynamic natural language generation. This integration enables the conversion of raw computational output into coherent and accessible narratives, facilitating a deeper understanding of genomic organization and gene function. The RAG framework augments LLM capabilities by retrieving the latest domain-specific knowledge, which is then used to refine and contextualize the insights generated. Through custom prompt engineering, our system synthesizes diverse datasets to highlight key aspects of genomic variation, conserved synteny, and annotation consistency across multiple E. coli strains. In general, our work demonstrates that integrating RAG with traditional bioinformatics methods offers a powerful, scalable solution to transform complex genomic datasets into actionable biological insights, paving the way for more efficient and accurate genomic analysis in microbial research.

1 INTRODUCTION

Escherichia coli (E.coli) is a model organism that has been extensively studied due to its diverse genetic background and its role as both a commensal and a pathogen. Genomic analyses in E. coli are essential for understanding bacterial evolution, virulence, antibiotic resistance, and environmental adaptabilityYang et al. (2019). Bioinformatics tools such as pairwise sequence alignment, NCBI-based gene annotation, and multiple sequence alignments (MSA) provide valuable insight into genomic structure and function. Previous studies have shown that traditional bioinformatics tools such as BLASTAltschul et al. (1990) and NCBI annotation Coordinators (2015) are essential for understanding genomic structure, although they often require expert interpretation to decipher complex outputs. However, interpreting the complex outputs of these methods can be time-consuming and may require additional manual curation to uncover subtle genomic variations and evolutionary patterns. Recent advances in artificial intelligence, particularly the emergence of large language models (LLMs), offer promising avenues for streamlining the interpretation process. LLMs such as GPT o3-mini ope (2023), Gemini 2.0 Flash Experimental Thinking model Anil & et al. (2023), and Grok 3 gro (2025) have demonstrated the ability to process and synthesize large volumes of textual and numerical data, generating coherent and human-readable narratives. In this study, we integrate LLMs into our bioinformatics workflow to automatically convert raw outputs from pairwise alignments, NCBI annotations, and MSAs into concise summaries that elucidate key genomic features in *E. coli*.

2 METHODS

The goal is to use LLM to leverage the genomic context of a gene by integrating a structured prompt and a synteny table (showing the conserved gene order). Using the RAG framework, the system retrieves relevant genomic data and allows LLMs to generate coherent explanations of why a gene appears as a pseudogene or is misannotated in certain *E. coli* strains. The focus is on contextualizing the gene's presence, function, and surrounding genomic environment rather than directly interpreting BLAST and MSA results. This approach enhances annotation accuracy and facilitates better genomic insights.

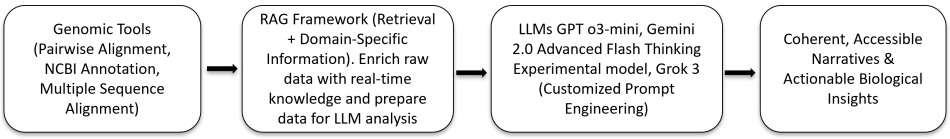

Figure 1: Workflow: Integrated Genomic Analysis Workflow Using RAG and LLMs

2.1 DATA PREPARATION FOR ANNOTATION ANALYSIS

To evaluate the potential of Large Language Models (LLMs) in genomic interpretation, we applied our framework to nearly 20 genes and selected sra, a regulatory RNA/pseudogene candidateKeseler & et al. (2011), for in-depth analysis. Sra was chosen due to its inconsistent annotations across E. coli strains, making it an ideal test case for assessing how LLMs handle genomic context and misannotation. We analyzed sra using synteny analysis, BLAST, NCBI annotations, and MSA to examine gene order, mutations, and pseudogene classification. The information from general feature format file (gff), NCBI and synteny table served as the foundation for LLM-driven reasoning under our Retrieval-Augmented Generation (RAG) framework, enabling structured genomic insights of the gene sra.

a. GFF file (ECOR50 strain): Accession: NZQOYJ01000074.1, Gene: bdm (ID: 04644, Product: Protein bdm); Accession: NZQOYJ01000074.1, Gene: maeA (ID:04645, Product: NAD-dependent malic enzyme)

b. NCBI annotation for sra: Coordinates: 13028–13165, Gene: sra, locustag=BEA19RS25490, Pseudo, Product: Ribosome-associated protein

c. Synteny table: Our synteny analysis shows that sra is a soft core gene in W3110, UTI89, MG1655, and CFT073, positioned between the core genes maeA and bdm. However, ECOR50 lacks sra entirely. The table contains the tuple with gene name, strain ID, strand orientation, and core status, where "core" genes are present in all E. coli strains, "soft core" genes are missing in one or two strains of *E.coli*.

Strain	Upstream gene info	Soft core gene	Downstream gene info
ECOR50	('bdm', '04644', '+', 'core')	Sra is missing	('maeA', '04645', '+', 'core')
W3110	('maeA', '3099', '-', 'core')	('sra', '3101', '-', 'soft core')	('bdm', '3103', '-', 'core')
UTI89	('maeA', '3243', '-', 'core')	('sra', '3245', '-', 'soft core')	('bdm', '3247', '-', 'core')
MG1655	('maeA', '3192', '-', 'core')	('sra', '3194', '-', 'soft core')	('bdm', '3196', '-', 'core')
CFT073	('maeA', '3543', '-', 'core')	('sra', '3545', '-', 'soft core')	('bdm', '3547', '-', 'core')

Table 1: Synteny (conserved order of genes) for sra gene

2.2 PROMPTING LLMS

We harnessed the power of LLMs by designing a structured prompt to guide their analysis. For instance, in the case of sra gene, the prompt instructed the models to analyze the annotation of the sra

gene in E. coli using synteny data, functional annotations (NCBI, GFF), BLAST, and MSA results. Explain how the genomic context of sra, including flanking core genes, supports or challenges its annotation.

2.3 RETRIEVING FUNCTIONAL ANNOTATIONS TO AUGMENT GENERATION

The Pairwise Alignment (BLAST) and MSA results provide detailed genomic data on the sra gene across multiple E. coli strains. The Pairwise Alignment shows sequence similarities and differences between the sra gene of ECOR50 and a reference strain, highlighting mutations, insertions, and deletions. The MSA, on the other hand, compares the sra gene across several E. coli strains, identifying conserved regions and variations in sequence that might impact the gene's function.

The LLM processes both the BLAST results and the MSA to assess the gene's structure, functional relevance, and variations across different strains. From the BLAST data, the LLM identifies critical mutations, such as premature stop codons, that may lead to the gene's misannotation or pseudogene status in certain strains. The MSA data provides further insights into conserved regions, helping the LLM explain why certain strains might exhibit functional discrepancies in sra. We designed specific prompts to analyze alignment results effectively, as shown in the results section.

```
Query: CFT073 Query ID: lcl|Query_1246744 Length: 138
>NZ_QOYJ01000074.1
Sequence ID: Query_1247023 Length: 16669
Range 1: 13028 to 13165

Score:236 bits(261), Expect:7e-63,
Identities:135/138(98%),  Gaps:0/138(0%), Strand: Plus/Plus

Query  1      ATGAAATCGAACCGTCAGGCACGTCATATTCTTGGACTGGACCATAAAATTTCTAACCAG  60
              ||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||||
Sbjct  13028  ATGAAATCGAACCGTCAGGCACGTCATATTCTTTGACTGGACCATAAAATTTCTAACCAG  13087

Query  61     CGCAAAATAGTTACCGAAGGTGACAAATCCAGCGTGGTAAATAATCCCACCGGCAGAAAA  120
              |||||||||||||||||||||||||||||||||||||||||||| || ||||||||||||
Sbjct  13088  CGCAAAATAGTTACCGAAGGTGACAAATCCAGCGTGGTAAATAACCCAACCGGCAGAAAA  13147

Query  121    CGCCCCGCTGAAAAGTAA  138
              ||||||||||||||||||
Sbjct  13148  CGCCCCGCTGAAAAGTAA  13165
```

Figure 2: Pairwise alignment result of sra gene against ECOR50 strain of *E.coli*

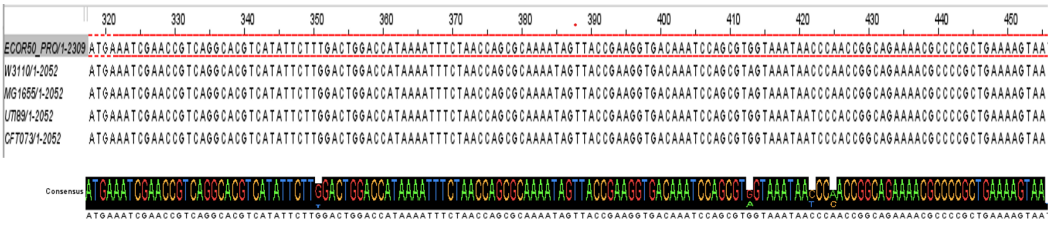

Figure 3: Showing only MSA of the sra region

3 RESULTS

To analyze the annotation of the sra gene in E. coli, we utilized all three LLMs with RAG to process functional annotation data, synteny information, and alignment results. We present the response of GPT o3-mini in the main text, while the responses of the other two are included in the appendix for comparison.

3.1 GENERAL PROMPT

Analyze the annotation of the sra gene in E. coli using synteny data and functional annotations (NCBI, GFF). Explain how the genomic context of sra, including flanking core genes, supports or challenges its annotation.

GPT o3-mini response: The synteny data and the conserved positioning of sra between two core genes (maeA and bdm) across multiple E. coli strains argue that the gene is part of a stable genomic neighborhood, hinting at an ancestral functional role related to ribosome association. However, the NCBI annotation labeling it as a pseudogene combined with its absence in the ECOR50 strain—raises questions about whether sra remains functional or has been inactivated in some lineages.

3.2 PROMPT FOR PAIRWISE ALIGNMENT

Given the BLAST alignment results, the sequence in strain CFT073 has GGA at position 13030-13033, while in ECOR50, it has changed to TGA. Does this introduce a premature stop codon, potentially rendering the sra gene nonfunctional? Consider synteny analysis, functional annotation, and conserved genes (bdm, maeA) when reasoning.

GPT o3-mini response: The pairwise alignment confirms that TGA at positions 13030–13033 in ECOR50 corresponds to a premature stop codon when compared to GGA in strains like CFT073. This mutation would truncate any putative sra product, strongly supporting the annotation of sra as a pseudogene in ECOR50. Synteny analysis and the presence of sra as a "soft core" gene in other strains reinforce that while sra was likely once functional in the ancestral E. coli lineage, it is no longer intact in ECOR50..

3.3 PROMPT FOR MULTIPLE SEQUENCE ALIGNMENT

In the MSA image, check if ECOR50 has a premature stop codon (TGA), while other strains have GGA (Gly). Show a table of codon differences and explain whether this mutation makes sra a pseudogene.

GPT o3-mini response: The multiple sequence alignment (MSA) clearly shows that while most E. coli strains (W3110, UTI89, MG1655, CFT073) have a conserved GGA codon (coding for glycine) at the critical position, ECOR50 exhibits a TGA codon. Since TGA functions as a stop codon in the bacterial genetic code, its presence in ECOR50 leads to premature termination of translation. This truncation likely results in a nonfunctional protein product, thereby supporting the classification of sra as a pseudogene in this strain.

Strain	Codon in MSA	Amino Acid	Interpretation
ECOR50	TGA	Stop (premature)	Introduces an early termination
W3110	GGA	Glycine	Intact coding sequence
UTI89	GGA	Glycine	Intact coding sequence
MG1655	GGA	Glycine	Intact coding sequence
CFT073	GGA	Glycine	Intact coding sequence

Figure 4: GPT o3-mini generated table of codon differences in MSA

4 CONCLUSION

The integration of retrieval-augmented generation (RAG) with large language models (LLMs) has significantly enhanced the interpretation of complex E. coli genomic data by transforming raw computational outputs into coherent and accessible narratives. By leveraging functional annotation, synteny analysis, and sequence alignments, this approach enables rapid and comprehensive identification of sequence variations and annotation discrepancies. We have applied this framework to approximately 20 genes and are expanding its scope to a broader set of genomic targets. While LLM

performance depends on high-quality prompt engineering and requires manual verification to ensure accuracy, our results highlight the potential of this method to streamline bioinformatics workflows and improve the efficiency and scalability of microbial genomic analysis.

REFERENCES

Gpt o3-mini large language model. 2023. URL https://openai.com/index/openai-o3-mini/.

x.ai. 2025. URL https://x.ai/grok.

S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

Rohan Anil and Sebastian Borgeaud et al. Gemini: A family of highly capable multimodal models. 2023. URL https://arxiv.org/abs/2312.11805.

NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic acids research*, 43(Database issue:D6–D17, 2015.

Keseler and Ingrid M et al. Ecocyc: a comprehensive database of escherichia coli biology. *Nucleic acids research*, 39 (Database issue:D583–90, 2011.

Zhi-Kai Yang, Hao Luo, Yanming Zhang, Baijing Wang, and Feng Gao. Pan-genomic analysis provides novel insights into the association of e.coli with human host and its minimal genome. *Bioinformatics (Oxford, England)*, 35,12:1987–1991, 2019.

A APPENDIX

The responses of the other two LLMs are provided in this section for comparison.

A.1 GENERAL PROMPT

Analyze the annotation of the sra gene in E. coli using synteny data and functional annotations (NCBI, GFF). Explain how the genomic context of sra, including flanking core genes, supports or challenges its annotation.

Gemini 2.0 flash thinking experimental model: The synteny data, combined with the NCBI pseudogene annotation, strongly suggests that the sra gene has been lost or significantly altered in ECOR50. The conserved genomic context in other E. coli strains highlights the typical presence and likely function of sra within this region. Further analysis using BLAST and MSA would provide additional evidence to support these conclusions.

Grok 3: The sra gene in E. coli is annotated as a "ribosome-associated protein" but labeled a pseudo gene in some contexts. Synteny shows sra consistently between core genes maeA and bdm in most strains (W3110, UTI89, MG1655, CFT073), supporting a conserved, non-essential role. However, it's absent in ECOR50, with maeA and bdm adjacent, suggesting a deletion or loss of function. This supports sra's legitimacy where present but challenges it in ECOR50. BLAST and MSA could confirm its status, but the data indicates a dispensable, strain-specific gene with potential ribosomal function.

A.2 PROMPT FOR PAIRWISE ALIGNMENT

Given the BLAST alignment results, the sequence in strain CFT073 has GGA at position 13030-13033, while in ECOR50, it has changed to TGA. Does this introduce a premature stop codon, potentially rendering the sra gene nonfunctional? Consider synteny analysis, functional annotation, and conserved genes (bdm, maeA) when reasoning.

Gemini 2.0 flash thinking experimental model: The change from GGA to TGA in the sra gene of ECOR50 introduces a premature stop codon. This strongly suggests that the sra gene has become a

pseudogene in this strain. The synteny analysis and functional annotation support this conclusion, indicating that sra is typically functional in other E. coli strains but has been inactivated in ECOR50.

Grok 3: Yes, the GGA→TGA change at positions 13030–13033 in ECOR50 likely introduces a premature stop codon, rendering the sra gene nonfunctional in this strain. This supports its absence in ECOR50's synteny and aligns with its pseudo gene annotation. The conserved flanking genes (maeA and bdm) indicate sra's dispensability, while its presence and potential functionality in other strains (like CFT073) suggest a strain-specific role, possibly in ribosome association, that's not universally critical. To confirm, analyzing the full CDS, reading frame, and additional mutations via MSA or full-genome sequencing would be ideal.

A.3 PROMPT FOR MULTIPLE SEQUENCE ALIGNMENT

In the MSA image, check if ECOR50 has a premature stop codon (TGA), while other strains have GGA (Gly). Show a table of codon differences and explain whether this mutation makes sra a pseudogene.

Gemini 2.0 flash thinking experimental model: The MSA image confirms the presence of a premature stop codon (TGA) in the sra gene of ECOR50, while other strains have a GGA codon. This mutation likely renders the sra gene a pseudogene in ECOR50 due to the production of a truncated, nonfunctional protein. This conclusion is supported by the synteny analysis and the functional annotation of sra.

ECOR50 Codon	Other Strains Codon	Amino Acid Change
TGA	GGA	Stop codon

Figure 5: Gemini generated table of codon differences in MSA

Grok 3: The TGA codon at positions 13030–13033 in ECOR50 introduces a premature stop codon, rendering the sra gene nonfunctional and supporting its classification as a pseudogene in this strain. In contrast, the GGA codon (glycine) in W3110, MG1655, UTI89, and CFT073 suggests a potentially functional sra gene in those strains. This mutation explains sra's absence in ECOR50's synteny and aligns with its pseudo gene annotation, highlighting strain-specific evolutionary divergence.

Strain	Codon (13030–13033)	Amino Acid	Functional Implication
ECOR50	TGA	Stop	Premature termination, likely nonfunctional
W3110	GGA	Glycine	Functional (if in frame)
MG1655	GGA	Glycine	Functional (if in frame)
UTI89	GGA	Glycine	Functional (if in frame)
CFT073	GGA	Glycine	Functional (if in frame)

Figure 6: Grok3 generated table of codon differences in MSA