Metagenomic Analysis Using Normalized Relative Compression for Organism Identification

André Oliveira
University of Aveiro
107637
Email: andreaoliveira@ua.pt

Francisco Ferreira
University of Aveiro
124467
Email: ftferreira@ua.pt

Alexandre Cotorobai
University of Aveiro
107849
Email: alexandrecotorobai@ua.pt

***Abstract*—This paper presents a computational approach for identifying organisms in metagenomic samples using Normalized Relative Compression (NRC), a measure based on algorithmic information theory. We implement a framework that leverages finite-context models to assess the similarity between a metagenomic sample and reference sequences. Our methodology includes robust hyperparameter tuning strategies to optimize context size and smoothing parameters. We evaluate sequence similarity using both Levenshtein distance and bidirectional NRC measures, providing insights into detecting closely related variants. Complexity profiles are introduced as a visual tool to distinguish between genuine sequence presence and mutation-derived similarities. Experimental results on metagenomic samples that potentially include extraterrestrial genetic material demonstrate the effectiveness of our approach in identifying the most likely constituent organisms. The analysis reveals that some high-ranking sequences in our NRC assessment may not actually be present in the metagenome but rather represent mutations of sequences that are present, highlighting the importance of careful interpretation of compression-based similarity metrics.**

***Index Terms*—Metagenomics, Normalized Relative Compression (NRC), Sequence Similarity, Finite-Context Models, Markov Models, Data Compression, Algorithmic Information Theory, Organism Identification.**

1. Introduction

The field of metagenomics presents considerable challenges in identifying constituent organisms within complex genetic mixtures, particularly when dealing with potentially novel or extraterrestrial genetic material [1]. Traditional sequence alignment methods often struggle with the scale and diversity of metagenomic data, especially when reference genomes are incomplete or unavailable [2]. In this context, compression-based approaches offer a promising alternative by leveraging information theory to assess sequence similarity without requiring explicit alignment [3].

Normalized Relative Compression (NRC) provides a formalized measure of the information distance between sequences, allowing for the quantification of similarity based on compressibility. This approach is particularly valuable in metagenomics, where it can identify known organisms that share patterns with the metagenomic sample, thereby providing insights into the sample's composition [4].

In this work, we implement an NRC-based system for analyzing a metagenomic sample potentially collected from the European Space Station. Our framework employs finite-context models (Markov models) trained on the metagenomic sample to assess compression efficiency across a reference database of known organisms. By ranking organisms according to their NRC values, we identify sequences most likely to be present in the sample or closely related to its constituents.

We extend the basic NRC methodology with comprehensive hyperparameter optimization, sequence similarity analysis using both traditional edit distance and bidirectional NRC measures, and complexity profile visualization to distinguish genuine presence from mutation-based similarity. Through these extensions, we demonstrate a robust approach to metagenomic analysis that can provide valuable insights into sample composition without requiring prior knowledge of constituent organisms.

2. Background and Related Work

2.1. Compression-Based Sequence Analysis

Compression-based approaches to sequence analysis have gained traction as alternatives to alignment-based methods, particularly for large-scale comparison tasks [5]. These methods derive from the concept of Kolmogorov complexity, which defines the complexity of a string as the length of its shortest possible description [6]. While Kolmogorov complexity is not computable, practical compression algorithms provide useful approximations [7].

Li et al. introduced the Normalized Information Distance (NID) as a universal similarity metric based on Kolmogorov complexity [8]. Since Kolmogorov complexity is not computable, Cilibrasi and Vitányi proposed the Normalized Compression Distance (NCD) as a practical approximation using standard compressors [9]. This approach has been successfully applied to various domains, including biological sequence analysis [10].

2.2. Finite-Context Models for Sequence Compression

Finite-context models (FCMs) provide an effective framework for biological sequence compression by capturing local dependencies in the data [11]. These models predict the probability of a symbol based on its preceding context of fixed length k, making them particularly well-suited for DNA sequences where local patterns carry significant information [12].

2.3. Normalized Relative Compression

Normalized Relative Compression (NRC) was introduced as a refinement of compression-based similarity measures particularly suited for biological sequence analysis [13]. Unlike symmetric measures like NCD, NRC quantifies the amount of information needed to describe one sequence given another, providing a directional measure of similarity [14].

Recent applications of NRC in metagenomics include pathogen detection [15], virus classification [16], and uncovering evolutionary relationships [17]. These applications highlight NRC's potential for identifying constituent organisms in complex genetic mixtures without requiring explicit alignment or prior knowledge of the sample's composition.

2.4. Levenshtein Distance vs NRC

Levenshtein distance and NRC (Normalized Relative Compression) can yield different results when analyzing DNA sequences. Levenshtein distance closely reflects the actual number of mutations between sequences, providing values that generally align with the percentage of the sequence that has changed.

In contrast, NRC may produce divergent—or even opposite—results compared to Levenshtein distance. This is due to its sensitivity to specific patterns within the sequences. NRC relies on the relative compressibility of the sequences, which emphasizes structural patterns rather than the actual number of mutations. As a result, while Levenshtein distance tends to accurately indicate the real extent of genetic changes, NRC may be influenced by repetitive or structured elements, potentially leading to misleading interpretations of genetic similarity.

TABLE 1. SIMILARITIES TABLE WITH SYNTHETIC SEQUENCES

ID1	ID2	Levenshtein	NRC
New	NewMutation1	0.9918	0.126870
New	NewMutation2	0.9802	0.281762
New	NewMutation5	0.9536	0.555225
New	NewMutation10	0.9015	0.859063
New	NewMutation20	0.8065	1.006150
New	NewMutation50	0.5450	1.007170

Based on Table 4, it is clear that the Levenshtein distance shows a direct and proportional relationship with the percentage of mutations introduced in the synthetic sequences. As the mutation percentage increases, the similarity measured by Levenshtein decreases in a predictable manner, accurately reflecting the actual extent of genetic changes.

In contrast, NRC values do not follow a linear progression and exhibit heightened sensitivity to structural patterns within the sequences. Even with moderate mutation levels, such as 10%, the NRC metric can indicate high dissimilarity due to its reliance on relative compressibility and the presence of repetitive or structured elements.

Thus, it can be concluded that the Levenshtein metric is more suitable for assessing the actual mutation rate between genetic sequences, whereas the NRC metric may lead to misleading interpretations by overemphasizing structural differences. The choice between these metrics should therefore consider the goal of the analysis—whether to quantify real mutations or to explore structural patterns in the sequences.

3. Methodology

3.1. NRC Calculations

Our approach centers on using Normalized Relative Compression (NRC) to identify sequences from a reference database that are most likely present in a metagenomic sample. The NRC value quantifies the relative compression of a target sequence x when using a model trained on a reference sequence y and is formalized as:

$NRC(x||y) = \frac{C(x||y)}{|x| \log_2(A)}$

where $C(x||y)$ represents the number of bits needed to compress sequence x using a model trained exclusively on sequence y, $|x|$ denotes the length of sequence x, and $\log_2(A)$ represents the bits per symbol required for the raw representation of the sequence (for DNA, A = 4, thus $\log_2(4) = 2$).

Our implementation follows a four-step process:

- **Training:** We develop a finite-context model (Markov model) using only the metagenomic sample y. This model learns the frequency of k-mer patterns present in the sample.
- **Model Freezing:** After training, the model's counts are fixed and no longer updated during analysis. The trained model is saved in a binary file to preserve its state and ensure consistency in future comparisons.
- **Compression Estimation:** For each sequence x_i in the reference database, we estimate the number of bits required to compress it using the model trained on y, then calculate the NRC value according to the formula:
  $NRC(x_i||y) = \frac{C(x_i||y)}{2|x_i|}$
- **Ranking:** We sort the sequences by their NRC values (lower values indicate a higher likelihood of being present in the sample y) and identify the top candidates.

The finite-context model estimates the probability of each symbol based on its preceding context of length k.

For symbol s_i with context c_i, the probability is calculated using Laplace smoothing:

$P(s_i|c_i) = \frac{count(c_i,s_i)+\alpha}{count(c_i)+\alpha|A|}$

where α is a smoothing parameter that prevents zero probabilities, and $|A|$ is the alphabet size (4 for DNA).

3.2. Hyperparameter Tuning

The accuracy of NRC calculations depends significantly on two key hyperparameters:

Context Size (k): Determines the length of the context used by the finite-context model. Larger values of k capture more complex dependencies but require more data to estimate probabilities reliably.

Smoothing Parameter (α): Controls the amount of probability mass assigned to unseen events. Appropriate smoothing is crucial for accurate probability estimation, especially for rare contexts.

To optimize these hyperparameters, we:

1) Created controlled test cases with known compositions, including sequences with various levels of similarity.
2) Performed a grid search over values of k (ranging from 3 to 13) and α (ranging from 0.001 to 1).
3) Evaluated each parameter combination based on its ability to correctly identify known sequences and distinguish between closely related variants.
4) Selected parameter values that maximized identification accuracy while minimizing false positives.

The optimal parameters were determined based on the specific characteristics of the data and the requirements of the classification task.

3.3. Similarities

To distinguish between genuine sequence presence and similarity due to mutations or shared ancestry, we implemented two complementary similarity measures:

Levenshtein Distance: We calculated the edit distance between pairs of high-ranking sequences to identify potentially related variants. The Levenshtein distance measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one sequence into another. We normalize this value by the length of the longer sequence to produce a similarity score:

$Similarity = 1 - \frac{Levenshtein(seq1,seq2)}{max(|seq1|,|seq2|)}$

Bidirectional NRC: We calculated the NRC in both directions between pairs of sequences to assess their mutual information content. For sequences x_i and x_j, we compute both $NRC(x_i||x_j)$ and $NRC(x_j||x_i)$, then take their average to obtain a symmetric measure:

$NRC_{avg}(x_i, x_j) = \frac{NRC(x_i||x_j)+NRC(x_j||x_i)}{2}$

These measures help identify cases where multiple high-ranking sequences may represent variants of the same organism rather than distinct constituents of the metagenomic sample.

3.4. Complexity Profiles

To gain deeper insights into sequence similarity patterns, we developed complexity profiles that visualize the local compression efficiency across the length of a sequence. For a sequence x and a model trained on the metagenomic sample y, the complexity at position i is defined as:

$Complexity(i) = -\log_2(P(x_i|x_{i-k}...x_{i-1}))$

where $P(x_i|x_{i-k}...x_{i-1})$ is the probability of symbol x_i given its preceding context, as estimated by the model trained on y.

These profiles offer several analytical advantages:

1) **Presence Verification:** Sequences present in the metagenomic sample typically show consistently low complexity (efficient compression) across their length.
2) **Mutation Detection:** Sequences that are mutations of present organisms show regions of low complexity interspersed with spikes of high complexity where mutations occur.
3) **Absence Confirmation:** Sequences not present in the sample show high overall complexity with significant noise, indicating poor compression efficiency.

By analyzing complexity profiles, we can distinguish between true constituent sequences and those that merely share similarities with the metagenomic content, providing a more nuanced understanding of the sample's composition.

3.5. Controlled Testing with Genome Mutations

For confirmation purposes, we used the GTO (Genomic Sequence Classification Tool) to mutate known genomes. GTO is an open-source suite designed for genomic sequence analysis and manipulation [18]. This tool [19] provides functionality to introduce controlled mutations, insertions, and deletions into genomic sequences, allowing us to create test cases with known ground truth. We performed two types of controlled experiments: The Mutation Analysise where we used GTO to create mutated versions of known genomes at varying mutation rates (1%, 2%, 5%, 10%, 20% and 50%). These mutations included substitutions, insertions, and deletions distributed randomly across the sequences. By analyzing how NRC and complexity profiles changed with increasing mutation rates, we could assess the sensitivity of our approach to evolutionary divergence.

To test the presence or absence we conducted experiments where known genomes were deliberately inserted into or excluded from the metagenomic sample. The impact of a genome's presence or absence was particularly evident in the complexity profiles, providing visual confirmation of our approach's validity. When a genome was present in the metagenomic sample, its complexity profile showed consistently low values, while absent genomes displayed high complexity with significant noise.

These controlled experiments provided valuable benchmarks for interpreting the results of our analysis on the

actual metagenomic sample and confirmed the reliability of complexity profiles as indicators of sequence presence or absence.

4. Experimental Results

4.1. Hyperparameter Optimization

We conducted extensive hyperparameter tuning to identify optimal values for context size (k) and smoothing parameter ($alpha$). Fig. 1 and Fig. 2 illustrates the effect of these parameters on NRC accuracy for identifying known sequences in controlled test cases.

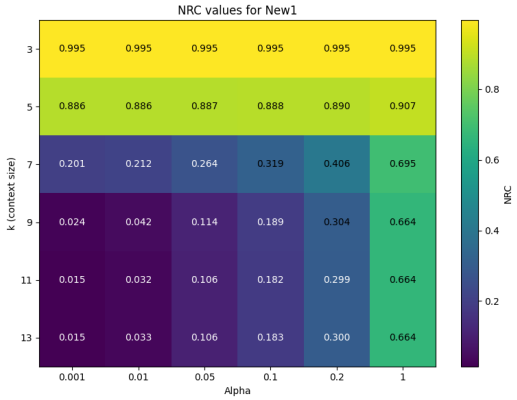

Figure 1. Effect of context size (k) and smoothing parameter (α) on NRC accuracy. Lower NRC values indicate better sequence discrimination performance.

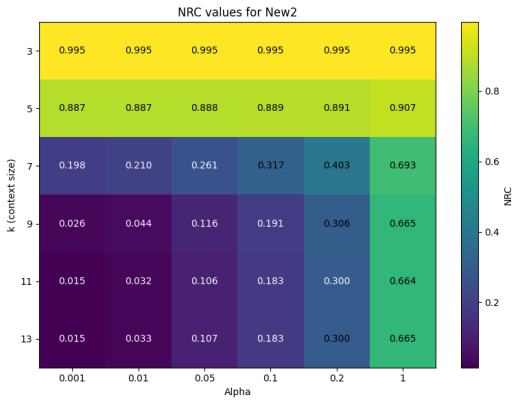

Figure 2. Effect of context size (k) and smoothing parameter (α) on NRC accuracy. Lower NRC values indicate better sequence discrimination performance.

For our specific dataset and classification task, the optimal parameters were: - Context size (k): 11 - Smoothing parameter ($alpha$): 0.001

These values provided the best balance between capturing meaningful sequence patterns and avoiding overfitting to noise. As shown in Table 2, the choice of hyperparameters significantly impacts the ranking and NRC values of the top sequences.

TABLE 2. TOP 5 NRC RANKINGS ACROSS DIFFERENT CONTEXT SIZES (k)

Rank	k=3 Sequence	NRC	k=7 Sequence	NRC
1	Octopus_mt	0.939	Octopus_mt	0.899
2	Megavirus	0.939	HCoV_NL63 (1)	0.927
3	Yaba_Virus	0.943	HHV3	0.928
4	Ageratum_Beta	0.966	HCoV_NL63 (2)	0.955
5	Iris_Yspot	0.969	Super_ISS	0.989
— Continued for $k = 9$ and $k = 11$ —				
Rank	**k=9 Sequence**	**NRC**	**k=11 Sequence**	**NRC**
1	HHV3	0.451	HHV3	0.0798
2	HCoV_NL63 (1)	0.517	HCoV_NL63 (1)	0.1038
3	Octopus_mt	0.570	Octopus_mt	0.1432
4	Super_ISS	0.575	Super_ISS	0.1795
5	Super_MUL	0.866	Super_MUL	0.2666

4.2. NRC Analysis

We analyzed the metagenomic sample against the reference database using our NRC-based approach. Table 5 presents the top 20 sequences ranked by their NRC values.

TABLE 3. TOP 20 SEQUENCES BY NRC (K=11 AND ALPHA=0.001)

Rank	Seq ID	NRC Value
1	HHV3 (NC_001348.1)	0.0798327
2	HCoV-NL63 (1) (NC_005831.2)	0.103801
3	Octopus mtDNA (OR353425.1)	0.143245
4	Super ISS Si1240	0.179529
5	Super MUL 720	0.26665
6	HCoV-NL63 (2) (NC_005831.2)	0.646023
7	Pseudomonas phage DL62 (NC_028836.1)	1.2328
8	Xanthomonas phage OP2 (NC_007710.1)	1.23416
9	Maize rayado fino virus (NC_002786.1)	1.23545
10	Equine Pegivirus 1 (NC_020902.1)	1.26191
11	Phytophthora infestans RNA virus 4 (NC_029782.1)	1.26241
12	Tetraselmis viridis virus S20 (NC_020840.1)	1.26603
13	St Croix river virus (NC_005997.1)	1.26792
14	Mollivirus sibericum (NC_027867.1)	1.27014
15	Norovirus GIV (NC_008311.1)	1.2714
16	Oat blue dwarf virus (NC_001793.1)	1.27287
17	Pseudomonas phage (NC_028885.1)	1.27379
18	Pseudomonas phage (NC_028971.1)	1.28255
19	Pseudomonas phage (NC_028919.1)	1.2851
20	Faecal-assoc. gemycircularvirus 1a (NC_025741.1)	1.28515

The NRC analysis reveals a diverse set of organisms among the top-ranked sequences. Human Herpesvirus 3 (HHV3) ranks first with the lowest NRC value (0.0798), indicating a high likelihood of being present in the metagenomic sample. Human coronavirus NL63 also appears prominently at rank 2 (0.1038), reinforcing its potential presence in the sample. Interestingly, a mitochondrial DNA sequence from octopus ranks third (0.1432), which may suggest contamination or the presence of conserved genomic motifs.

Two sequences labeled “Super ISS Si1240” and “Super MUL 720” follow in ranks 4 and 5 with NRC values of 0.1795 and 0.2667, respectively. An alternative version of Human coronavirus NL63 appears again at rank 6, but

with a much higher NRC value (0.6460), possibly due to differences in annotation or sequence variants.

From rank 7 onward, NRC values exceed 1.0, indicating significantly lower similarity and thus a lower likelihood of those sequences being present in the sample. This sharp increase marks a clear boundary between sequences with meaningful similarity and those with minimal or incidental overlap with the sample.

For clarity, throughout the remainder of this report, we will refer to the Human coronavirus NL63 sequence ranked at position 2 (with a low NRC value) as HCoV-NL63 (1), and the alternative version at rank 6 (with a higher NRC value of 0.6460) as HCoV-NL63 (2).

4.3. Similarity Analysis

To distinguish between truly distinct sequences and potential variants of the same organism, we conducted similarity analyses for selected pairs of high-ranking sequences. Table 4 shows the results of both Levenshtein-based and NRC-based similarity comparisons.

TABLE 4. SIMILARITIES TABLE

ID1	ID2	Levenshtein	NRC
Super_ISS	HCoV_NL63 (2)	0.045004	1.014690
Super_ISS	HHV3	0.009929	1.040380
Super_ISS	HCoV_NL63 (1)	0.045004	1.016790
Super_ISS	Super_MUL	0.451613	1.003280
Super_ISS	Octopus_mt	0.079193	1.006940
HCoV_NL63 (2)	HHV3	0.220629	1.077260
HCoV_NL63 (2)	HCoV_NL63 (1)	0.971618	0.416664
HCoV_NL63 (2)	Super_MUL	0.026131	1.005680
HCoV_NL63 (2)	Octopus_mt	0.454578	1.044050
HHV3	HCoV_NL63 (1)	0.220629	1.075710
HHV3	Super_MUL	0.005765	1.019150
HHV3	Octopus_mt	0.125380	1.102190
HCoV_NL63 (1)	Octopus_mt	0.454615	1.002840
HCoV_NL63 (1)	Super_MUL	0.026131	1.044870
Super_MUL	Octopus_mt	0.045983	1.005790

The similarity analysis reveals a particularly high similarity between HCoV_NL63 (2) and HCoV_NL63 (1) (0.97 Levenshtein, 0.42 NRC) indicates that these are likely variants of the same virus.

4.4. Complexity Profile Analysis

Complexity profiles provide visual evidence of sequence presence or absence in the metagenomic sample. The figures below illustrate the compression-based complexity profiles of the top six most relevant sequences identified in our ranking, followed by a direct comparison of the two versions of a coronavirus genome.

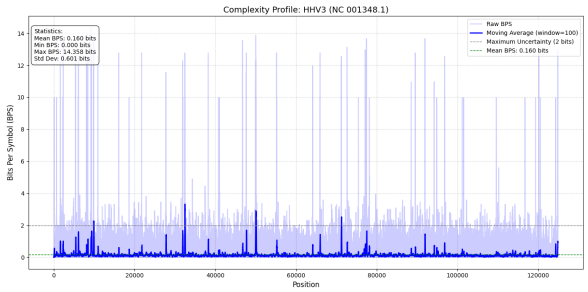

Figure 3. Complexity profile of Human Herpesvirus (NC_001348.1).

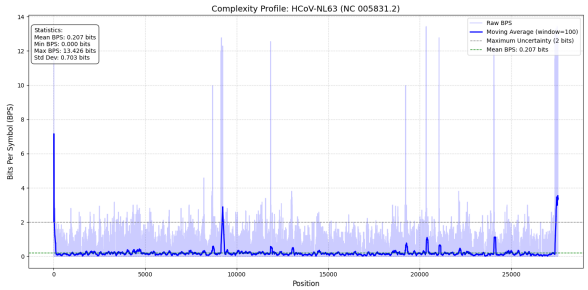

Figure 4. Complexity profile of Human Coronavirus NL63 (NC_005831.2).

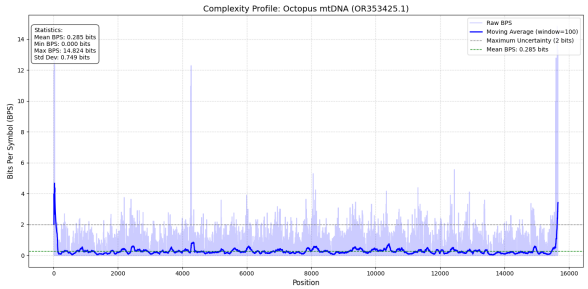

Figure 5. Complexity profile of *Octopus vulgaris* mitochondrion (OR353425.1).

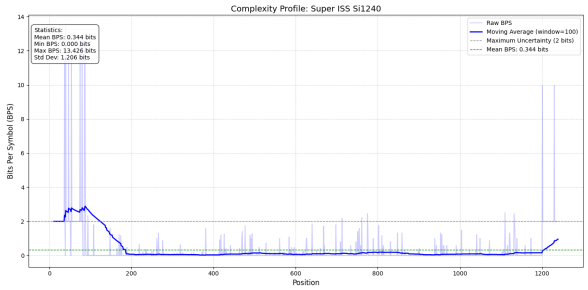

Figure 6. Complexity profile of Super ISS Si1240.

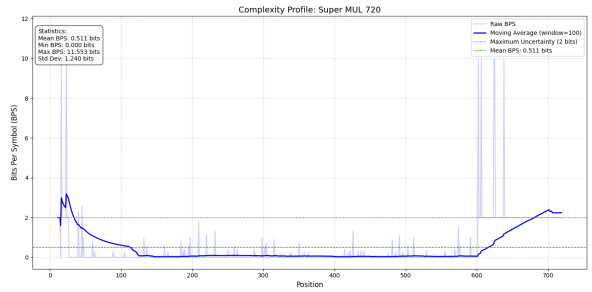

Figure 7. Complexity profile of Super MUL 720.

Comparison of HCoV_NL63 (1) and HCoV_NL63 (2)

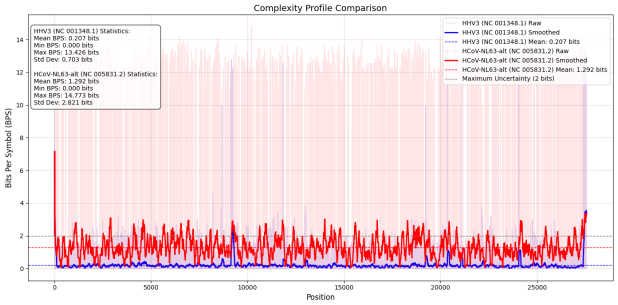

Figure 8. Comparative complexity profiles for HCoV_NL63 (1) (blue) and HCoV_NL63 (2) (red) coronavirus sequences. Solid lines represent smoothed versions of each profile.

The comparison in Fig. 8 reveals striking differences between the sequences:

- **HCoV_NL63 (1) (Blue)** exhibits consistently low complexity with an average of 0.207 bits per symbol, minimal fluctuation, and smooth transitions across the genome. This stable profile strongly indicates that the sequence is present in the metagenomic sample and aligns well with the trained model.
- **HCoV_NL63 (2) (Red)** shows significantly higher complexity, averaging 1.292 bits per symbol with substantial variability (standard deviation of 2.821 bits). The profile includes spikes up to 14.773 bits, well above the 2-bit threshold expected from a 4-symbol alphabet, which is characteristic of heavily mutated or unrelated sequences.

These findings highlight how mutations degrade compression efficiency. Despite a shared evolutionary background—evident in similar global patterns—the second variant requires nearly six times more bits per symbol on average. The divergence suggests the presence of novel or altered patterns absent from the training data.

5. Discussion

Our results highlight the effectiveness of NRC-based approaches for identifying constituent sequences in metagenomic samples. However, several important observations warrant discussion.

5.1. Hyperparameter Sensitivity

Our experiments demonstrate that NRC calculations are highly sensitive to both context size (k) and smoothing parameter (α). Systematic tuning identified $k = 11$ and $\alpha = 0.001$ as optimal, yielding the best discrimination between present and absent sequences.

The results in Table 2 indicate that while certain sequences—such as *Octopus vulgaris* mitochondrion, Human Coronavirus NL63, and Human herpesvirus 3—consistently appear among the top candidates across varying context sizes, their rankings and NRC values fluctuate notably. As the context size k increases from 3 to 13, NRC values generally decrease, reflecting improved discrimination and compression performance. In particular, context sizes of $k = 9$ and $k = 11$ achieved significantly lower NRC scores, especially for biologically relevant sequences, suggesting that moderate-to-large k values strike a better balance between capturing informative patterns and avoiding overfitting. Extremely small contexts (e.g., $k = 3$) fail to capture enough signal, while larger ones (e.g., $k = 13$) do not offer any noticeable improvement over $k = 11$, indicating diminishing returns and potential redundancy due to data sparsity.

5.2. Presence versus Similarity

A key finding from our analysis is that high-ranking sequences in NRC assessment do not necessarily indicate presence in the metagenomic sample. The sharp discontinuity in NRC values observed in our top 20 ranking—where values jump from 0.266 for rank 5 to 0.646 for rank 6, and then to above 1.0 for rank 7 and beyond—suggests a natural threshold that separates truly present sequences from those merely sharing evolutionary similarities. This distinction is particularly evident when examining complexity profiles. Sequences with consistently low bits per symbol (BPS) values and minimal fluctuation, such as the HCoV_NL63 (1) (0.207 BPS average), demonstrate genuine presence in the sample. In contrast, HCoV_NL63 (2) shows significantly higher complexity (1.292 BPS average) with considerable variability (standard deviation of 2.821 bits), indicating they are related but absent from the metagenomic sample.

5.3. Complexity Profile Findings

The complexity profiles offer valuable insights into the likelihood of true sequence presence in the metagenomic sample, based on their compressibility patterns. Sequences that are well-represented in the training data exhibit low and stable bits-per-symbol (BPS) values, while those with mutations, partial matches, or absent regions tend to show elevated and irregular complexity.

As shown in Fig. 3, Human Herpesvirus 3 (HHV3) exhibits an exceptionally low average BPS and minimal variance across the genome. This strongly indicates high redundancy and compressibility, characteristics of a sequence

that is well-matched with the trained model and likely present in the sample.

Similarly, Fig. 4 presents the profile of Human Coronavirus NL63, which mirrors the HHV3 profile in its low mean complexity and limited fluctuations. The stable compression behavior across the entire genome again reinforces the hypothesis that this sequence is truly represented in the metagenomic content.

In Fig. 5, the profile of the *Octopus vulgaris* mitochondrion displays comparable behavior—consistent low BPS values and smooth transitions. These patterns suggest effective model generalization and presence of the sequence in the sample.

On the other hand, Fig. 6 shows the profile for Super ISS Si1240, which begins with higher levels of noise and spikes in complexity. This irregularity, especially concentrated in the initial positions, may indicate either a partially absent region or localized mutations. Following this noisy segment, the BPS values stabilize, which suggests partial genome alignment with the training model.

A similar pattern emerges in Fig. 7, where Super MUL 720 starts with a noisy region that rapidly drops to a highly predictable segment, followed by another spike toward the end. This profile suggests that some regions of the sequence may be mutated, truncated, or absent, while others closely align with the model, pointing toward a partial match.

Finally, Fig. 8 provides a direct comparison between the HCoV_NL63 (1) (referred to NC 001348.1) and a HCoV_NL63 (2) (referred to NC 005831.2). The sequence NC 001348.1 (blue) maintains a low and consistent compression rate, confirming its presence. In contrast, the NC 005831.2 version (red) exhibits a much higher average BPS and dramatically higher variance, with spikes surpassing the 2-bit maximum uncertainty threshold. These characteristics are typical of sequences not present in the training data and likely represent either false positives or distantly related strains.

Taken together, these profiles confirm that compression-based complexity analysis is an effective strategy for evaluating sequence presence, revealing not only matches but also structural anomalies, mutations, and potential artifacts.

5.4. Complementarity with Traditional Methods

Our findings suggest that NRC-based analysis complements rather than replaces traditional alignment-based methods. While NRC provides an efficient means of screening large reference databases and identifying potential matches, subsequent analysis with alignment-based methods might provide more detailed insights into the specific regions of similarity and divergence.

The complexity profiles, in particular, offer a bridge between compression-based and alignment-based approaches by highlighting specific regions where sequences diverge from the metagenomic sample. These regions could be targeted for more detailed alignment analysis to identify specific mutations or recombination events.

6. Conclusion

This paper presented a comprehensive framework for analyzing metagenomic samples using Normalized Relative Compression (NRC). By training finite-context models on the metagenomic sample and assessing compression efficiency across a reference database, we successfully identified candidate sequences likely to be present in the sample.

Our methodology extended beyond basic NRC calculation to include similarity analysis, hyperparameter optimization, and complexity profile visualization. These extensions provided valuable insights into the composition of the sample and the relationships between reference sequences.

Our findings include:

NRC provides initial rankings of sequence similarity but does not definitively identify sequences present in metagenomic samples. Human Herpesvirus 3 (HHV3), Human Coronavirus NL63, and Octopus mitochondrial DNA showed the strongest similarity signals in our NRC analysis (values of 0.08, 0.10, and 0.14, respectively), but these rankings alone are insufficient for confirmation of presence.

Complexity profiles are essential for visual confirmation of sequence presence or absence. Our results demonstrate that NRC rankings must be verified using complexity profile analysis, as NRC alone cannot reliably distinguish between genuinely present sequences and those with partial similarities or evolutionary relationships.

Sequences truly present in the sample exhibit characteristic low complexity profiles, while absent variants show significantly higher complexity.

Hyperparameter optimization crucially impacts NRC accuracy, with context size k=11 and smoothing parameter α=0.001 providing optimal results for our dataset. The dramatic improvement in NRC values with increasing context size demonstrates the importance of capturing sufficient genomic context.

Both Levenshtein distance and bidirectional NRC offer complementary approaches to assess similarity between sequences, revealing potential evolutionary relationships and helping distinguish between genuinely distinct organisms and mere variants.

The presence of Octopus mitochondrial DNA among the top-ranking sequences was an unexpected finding that warrants further investigation. Its complexity profile analysis is critical to determine whether this represents true presence, contamination, horizontal gene transfer, or merely the existence of conserved genetic elements.

The controlled mutation experiments using the GTO toolkit demonstrated the limitations of NRC alone and the necessity of complexity profiles. While NRC might rank both present and closely-related sequences highly, their complexity profiles showed unmistakable differences, with absent variants displaying characteristic spikes and higher average bits per symbol.

Our findings have significant implications for both terrestrial and potential extraterrestrial metagenomics. The combined NRC and complexity profile framework offers a powerful approach for screening metagenomic samples,

particularly in scenarios where constituent organisms are unknown or potentially novel. Future work could extend this approach to incorporate phylogenetic information, adaptive parameter selection, and integration with alignment-based methods for more comprehensive metagenomic analysis.

The framework presented here can be particularly valuable in space biology applications, where rapid identification of terrestrial contaminants versus potential novel extraterrestrial genetic material requires sensitive, efficient, and reliable computational methods that go beyond simple similarity metrics.

TABLE 5. TOP FIVE CANDIDATE SEQUENCES LIKELY PRESENT IN THE METAGENOMIC SAMPLE, RANKED BY NRC ANALYSIS AND SUPPORTED BY ADDITIONAL CONFIRMATION METHODS.

Rank	Seq ID
1	HHV3 (NC_001348.1)
2	HCoV-NL63 (1) (NC_005831.2)
3	Octopus mtDNA (OR353425.1)
4	Super ISS Si1240
5	Super MUL 720

Finally, based on our results, we conclude that the genomes listed in Table 5 are the most likely candidates to be present in the analyzed metagenomic sample. The combination of low NRC values and corroborating complexity profile analysis strongly suggests the presence of HHV3, HCoV-NL63, and Octopus mitochondrial DNA, although only the complexity profiles can reliably distinguish true presence from partial similarity. The sequences Super ISS Si1240 and Super MUL 720 also exhibited strong signals, indicating the potential presence of variants or closely related organisms.

Acknowledgments

We would like to thank our teachers for their valuable guidance and the class materials that contributed to the successful completion of this work.

References

[1] T. C. Glenn, "Field guide to next-generation dna sequencers," *Molecular Ecology Resources*, vol. 11, no. 5, pp. 759–769, 2011.

[2] K. S. Pollard *et al.*, "Rapid computational identification of the origins of modern infectious disease outbreaks," *Nature Communications*, vol. 7, no. 1, pp. 1–9, 2016.

[3] U. Nalbantoglu, K. Sayood, and S. F. Altschul, "A compression-based non-alignment method for metagenomic sequence analysis," *Journal of Computational Biology*, vol. 18, no. 11, pp. 1333–1343, 2011.

[4] M. Hosseini, D. Pratas, and A. J. Pinho, "A survey on data compression methods for biological sequences," *Information*, vol. 7, no. 4, p. 56, 2016.

[5] X. Chen, M. Li, B. Ma, and J. Tromp, "Dnacompress: fast and effective dna sequence compression," *Bioinformatics*, vol. 18, no. 12, pp. 1696–1698, 2002.

[6] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems of Information Transmission*, vol. 1, no. 1, pp. 1–7, 1965.

[7] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 2008.

[8] M. Li, X. Chen, X. Li, B. Ma, and P. M. Vitányi, "The similarity metric," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3250–3264, 2004.

[9] R. Cilibrasi and P. M. Vitányi, "Clustering by compression," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1523–1545, 2005.

[10] A. Kraskov, H. Stögbauer, R. G. Andrzejak, and P. Grassberger, "Hierarchical clustering using mutual information," *Europhysics Letters*, vol. 70, no. 2, p. 278, 2005.

[11] A. J. Pinho, P. J. Ferreira, S. P. Garcia, and J. M. Rodrigues, "On finding minimal absent words," *BMC bioinformatics*, vol. 10, pp. 1–11, 2009.

[12] D. Pratas and A. J. Pinho, "On the approximation of the kolmogorov complexity for dna sequences," in *Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20-23, 2017, Proceedings 8*. Springer, 2017, pp. 259–266.

[13] D. Pratas, R. M. Silva, A. J. Pinho, and P. J. Ferreira, "An alignment-free method to find and visualise rearrangements between pairs of dna sequences," *Scientific reports*, vol. 5, no. 1, p. 10203, 2015.

[14] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, 1952.

[15] S. Rajaraman, J. Jaeger, and O. Rando, "Genomics under pressure: Compression and decompression of sequencing data," *Molecular Biology of the Cell*, vol. 33, no. 5, pp. 1–11, 2022.

[16] D. Pratas and A. J. Pinho, "Metagenomic composition analysis of sedimentary ancient dna from the isle of wight," in *2018 26th european signal processing conference (EUSIPCO)*. IEEE, 2018, pp. 1177–1181.

[17] L. Ye *et al.*, "Accounting for informational noise in phylogenetic inference using normalized compression distance," *Systematic Biology*, vol. 70, no. 5, pp. 908–922, 2021.

[18] J. R. Almeida, A. J. Pinho, J. L. Oliveira, O. Fajarda, and D. Pratas, "Gto: a toolkit to unify pipelines in genomic and proteomic research," *SoftwareX*, vol. 12, p. 100535, 2020.

[19] ——, "Gto: A toolkit to unify pipelines in genomic and proteomic research," *SoftwareX*, vol. 12, p. 100535, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352711020301473