NORMALIZED RELATIVE COMPRESSION IN METAGENOMES

Algorithmic Information Theory

André Oliveira, 107637
Alexandre Cotorobai, 107849
Francisco Ferreira, 124467

INTRODUCTION

GOAL

1

Analyze a metagenomic sample (meta.txt)

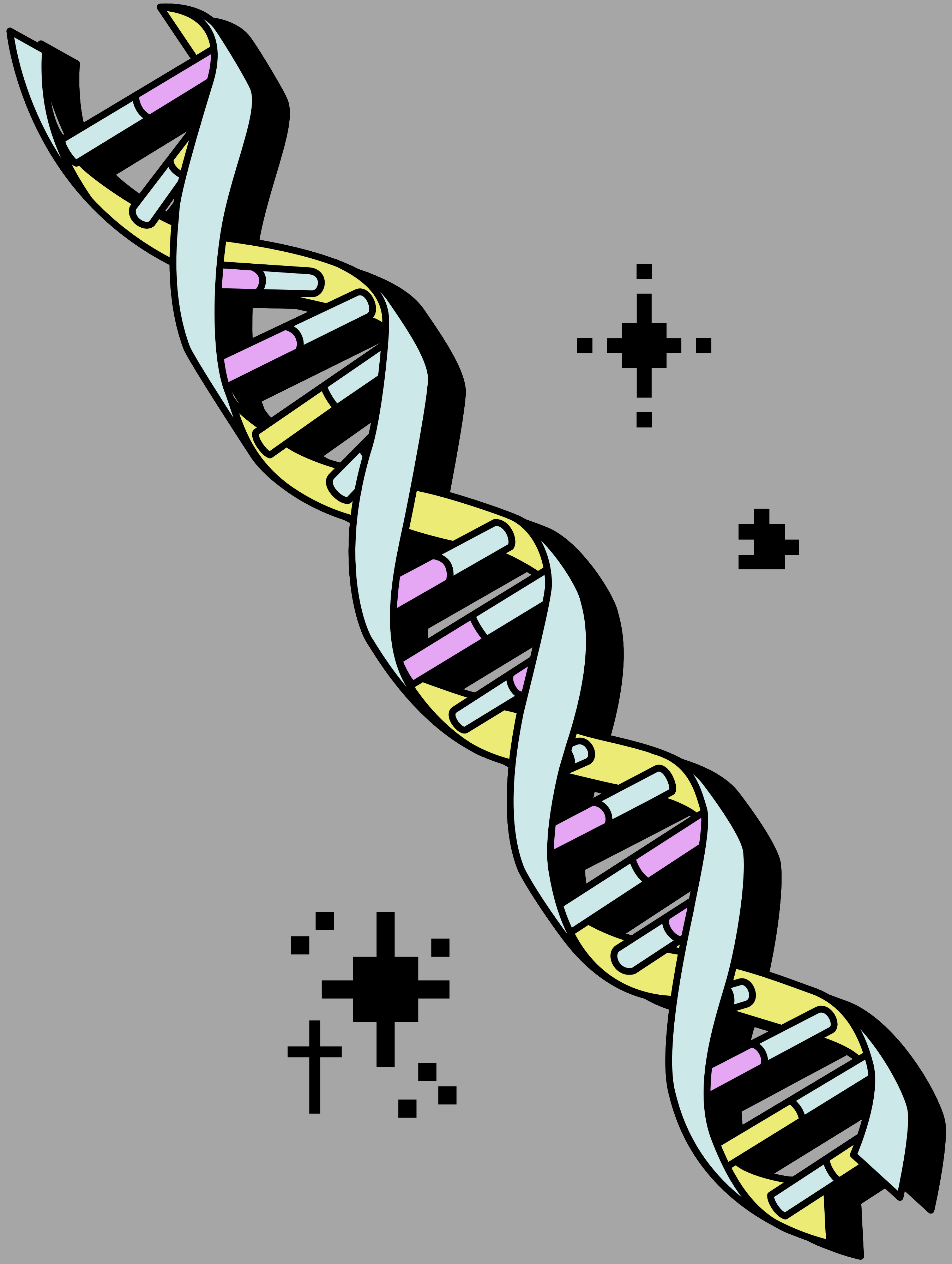
2

Identify similarities with known organisms or potential new life forms using NRC

IMPLEMENTATION: MODELS_GENERATOR.CPP

Trains a finite-context model (Markov model) with meta.txt

- Model is frozen and saved as a binary file
- Example: ./src/bin/models_generator.out -meta txt_files/meta.txt -k 11



IMPLEMENTATION: MAIN.CPP & METACLASS

- MetaClass handles the main program workflow;

- Compresses each database sequence using the trained model;

- Computes NRC values to estimate relative similarity;

- Outputs the top k most similar sequences;

- Example:

  ```
  ./src/bin/main.out -db txt_files/db.txt -m
  models/k11.bin -a 0.001 -t 20
  ```

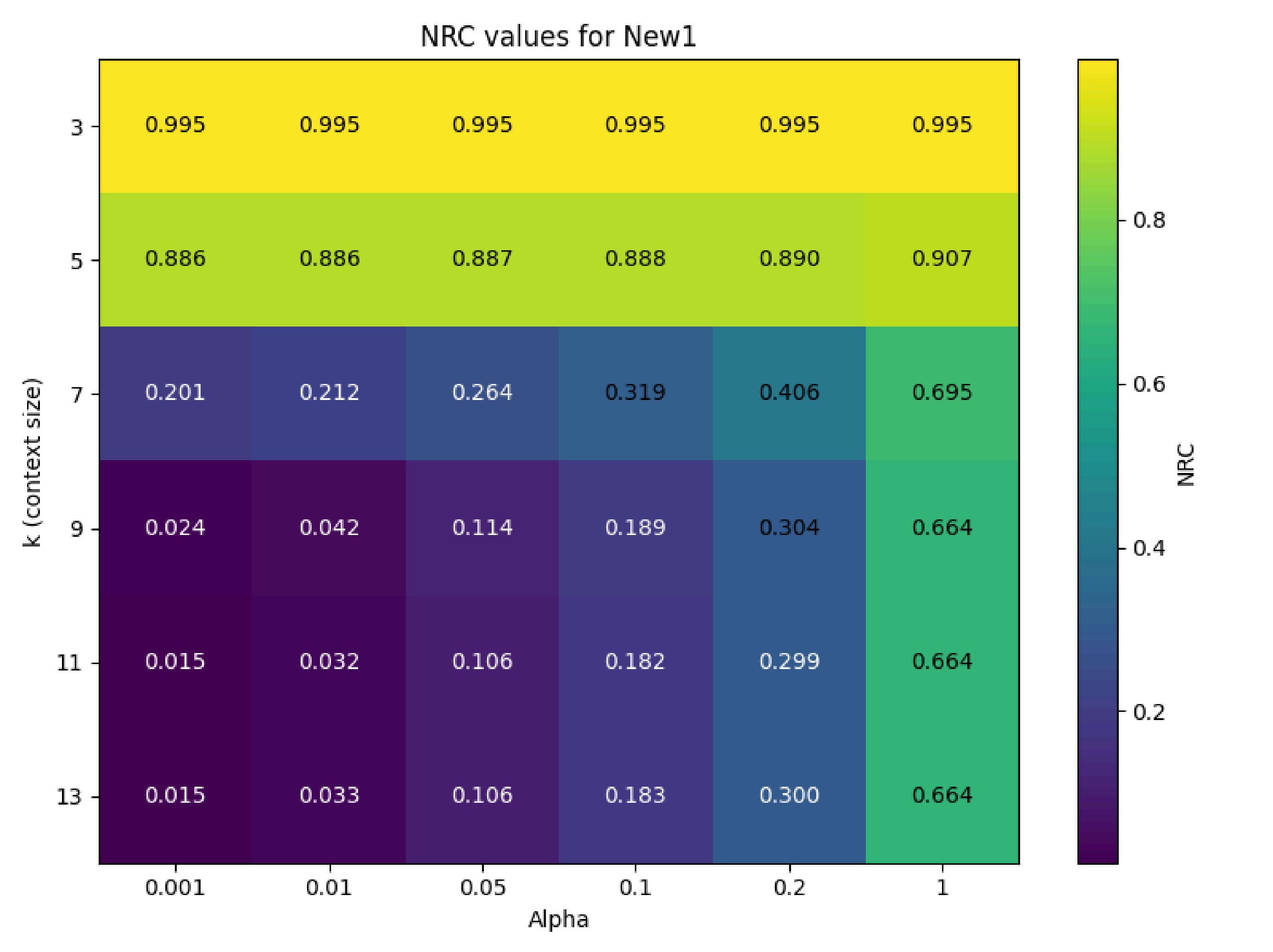
PARAMETER TUNING

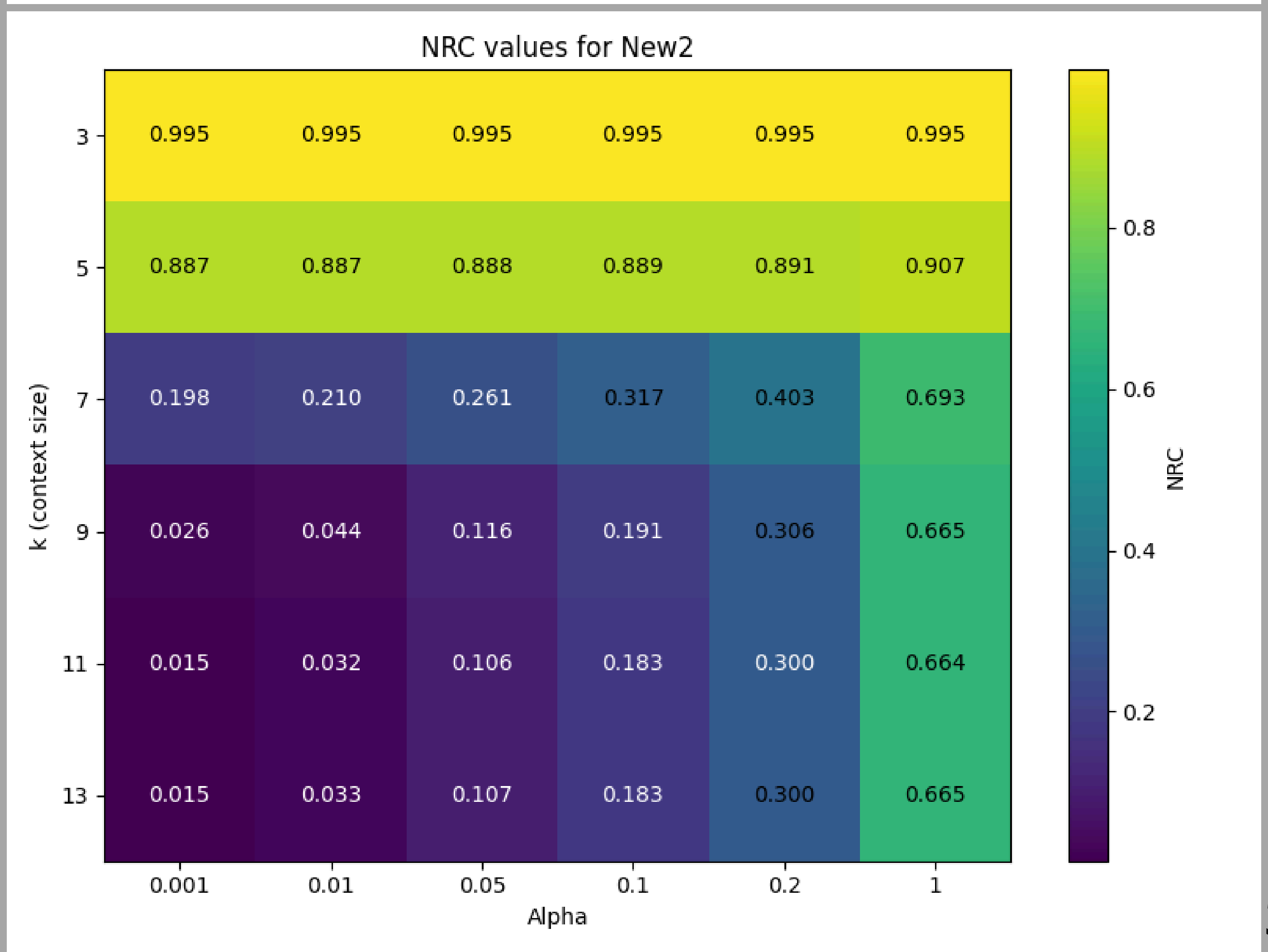
Extensive testing with **synthetic sequences** to find optimal k and alpha values

Heatmaps show optimal values:

- k = 11
- alpha = 0.001

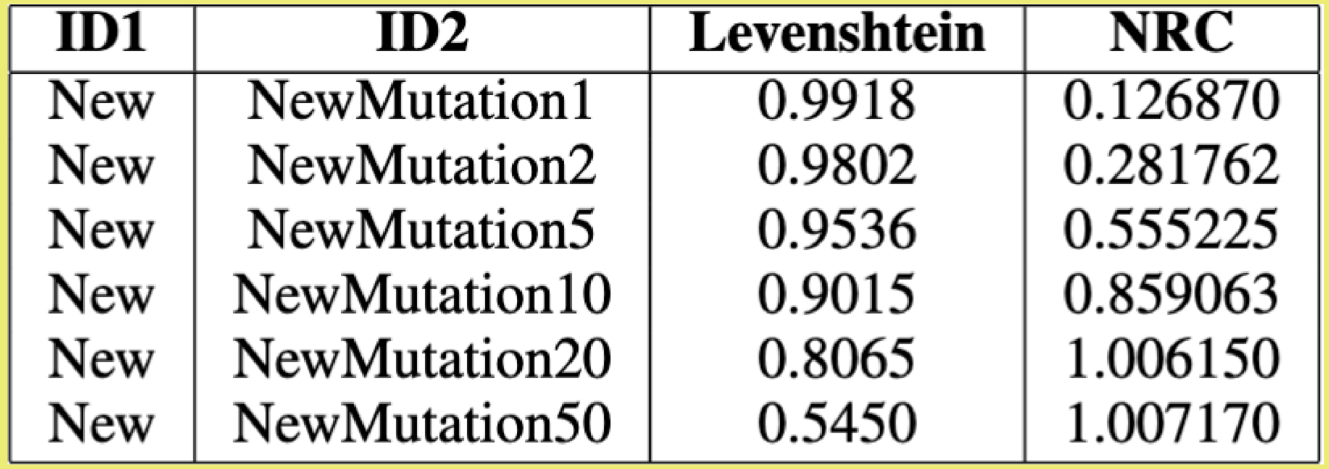
Higher k increases **space usage** without meaningful accuracy gains


NRC values for New1

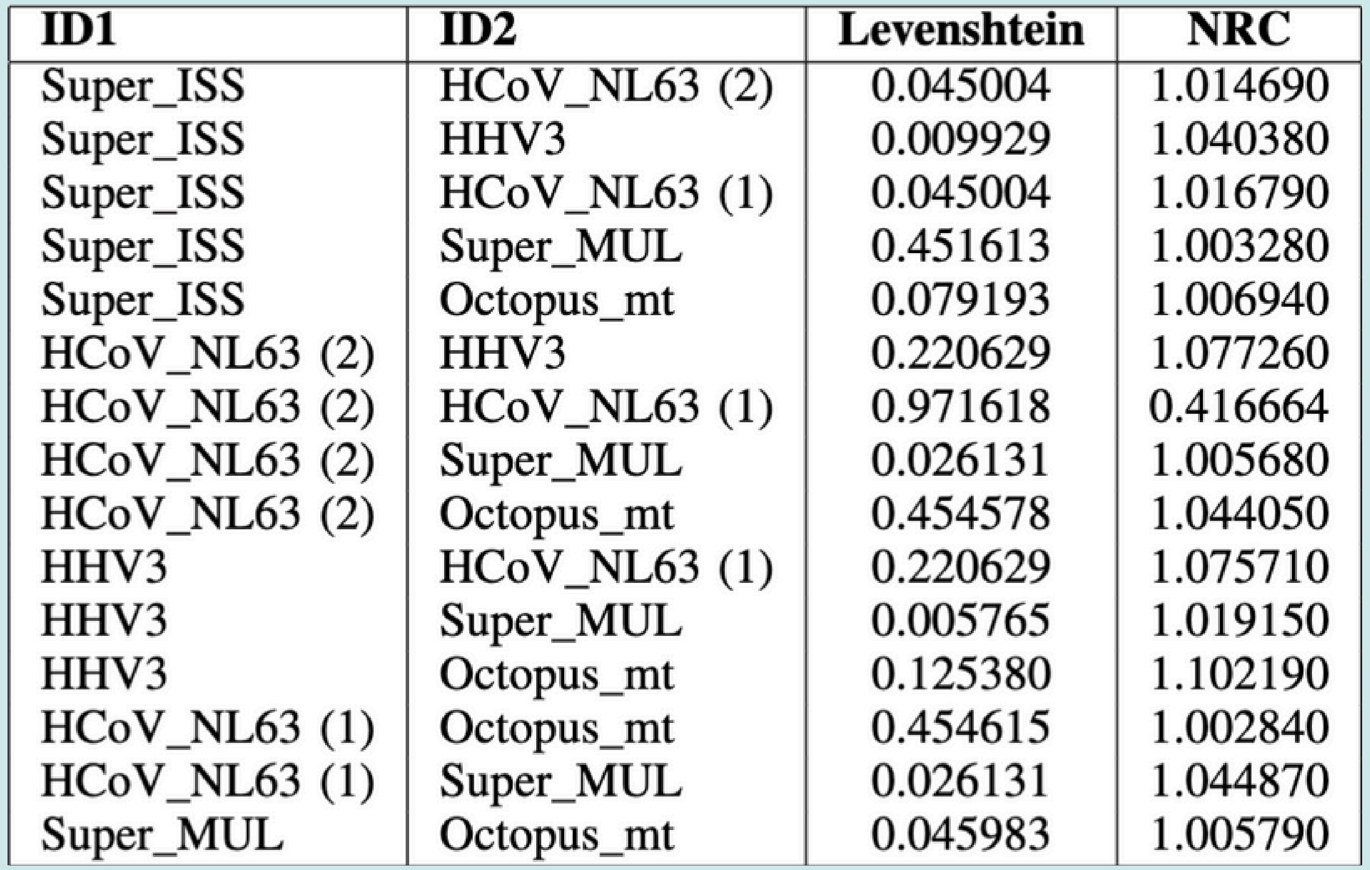

NRC values for New2

SEQUENCE SIMILARITY

Levenshtein distance accurately reflected the real relationships between sequences, while NRC was sensitive to specific patterns, sometimes overestimating due to repetitions or small changes with significant impact.



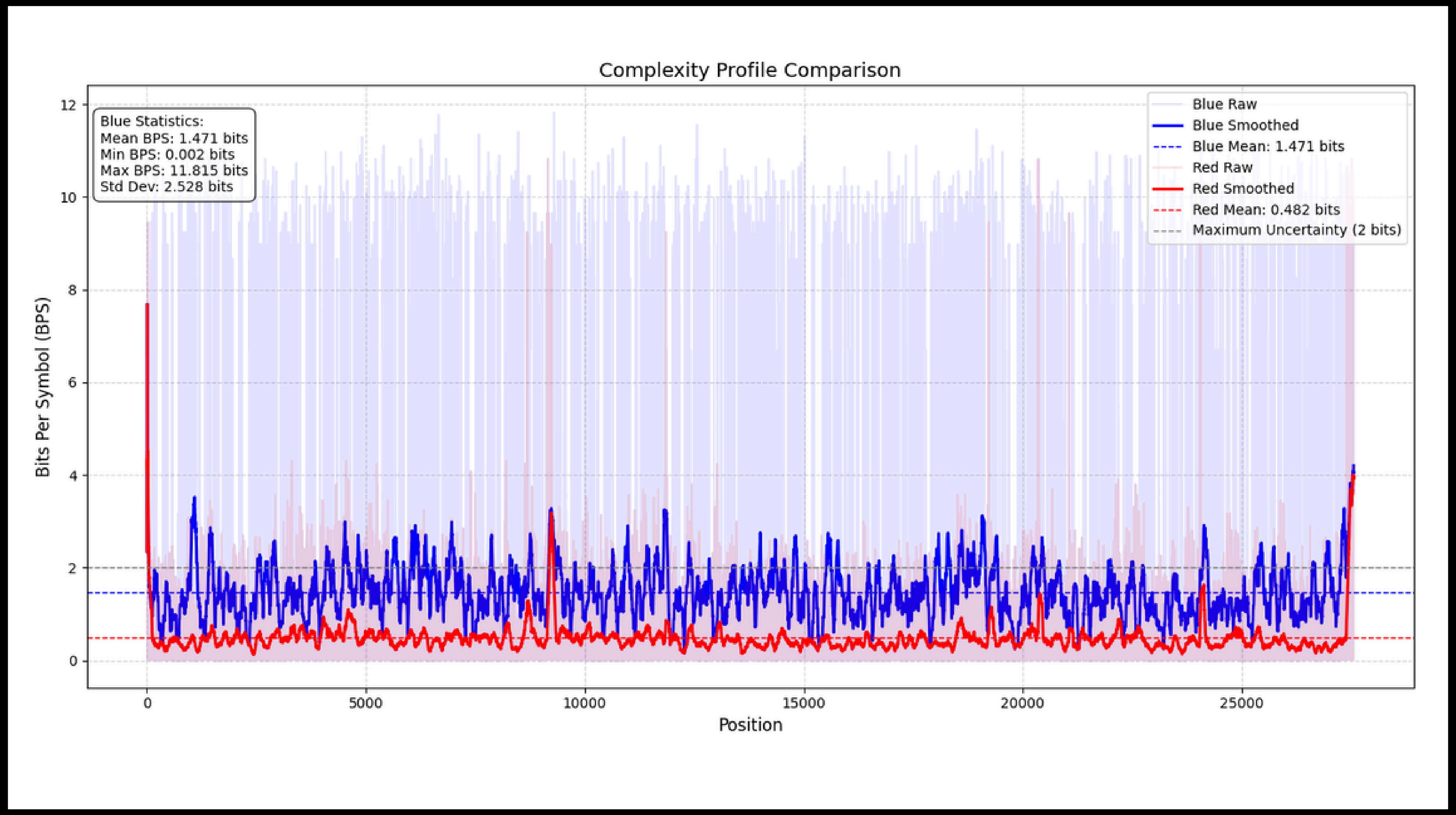
ID1	ID2	Levenshtein	NRC
New	NewMutation1	0.9918	0.126870
New	NewMutation2	0.9802	0.281762
New	NewMutation5	0.9536	0.555225
New	NewMutation10	0.9015	0.859063
New	NewMutation20	0.8065	1.006150
New	NewMutation50	0.5450	1.007170



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

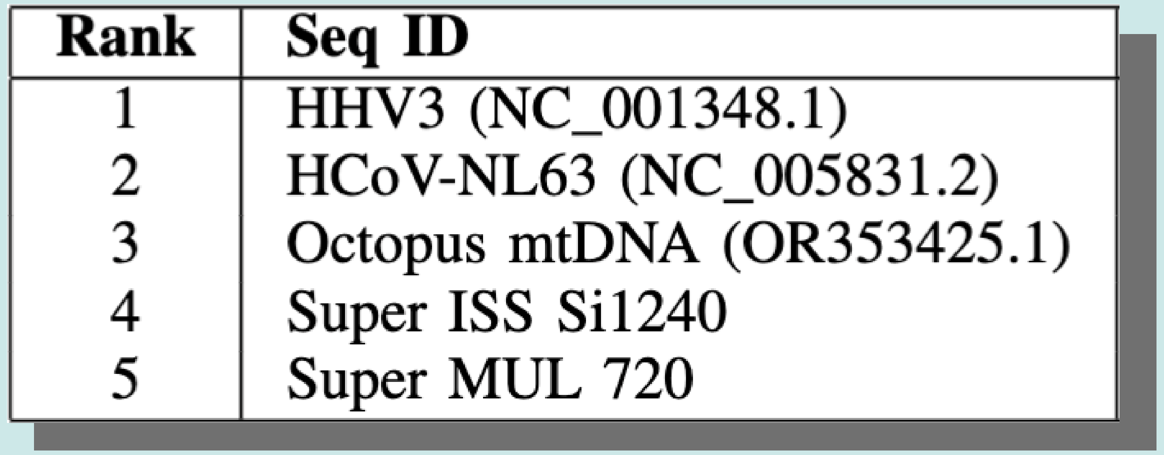
COMPLEXITY PROFILES

- Visualized complexity profiles of top sequences;
- Revealed conserved and information-dense regions;
- Top6 (blue) shown to be noisy and not truly present in the sample;
- Compared with top2 (red), which is reliably compressible



CONCLUSION

- Top 5 sequences most likely present in the sample;
- Top 6 excluded as it is a mutated variant of top 2;
- NRC combined with complementary techniques proved effective;
- Enables identification of contamination or unknown life in extreme environments.



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