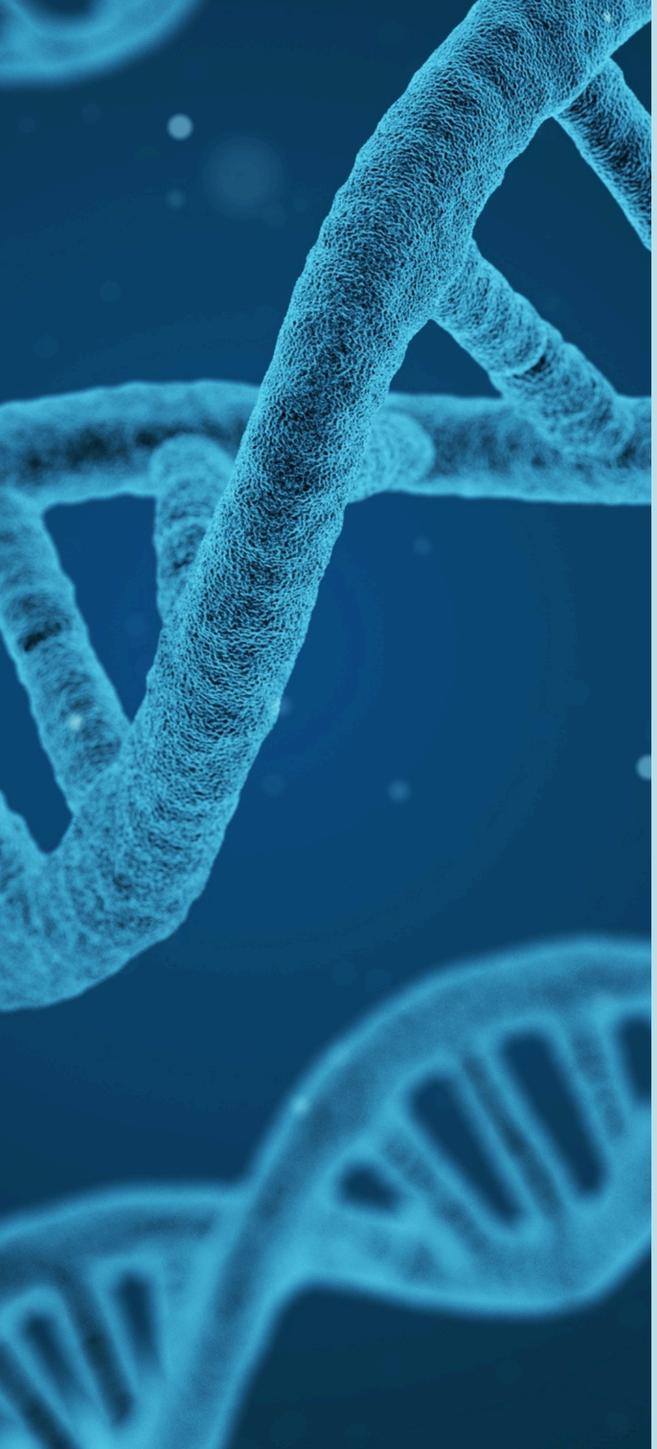
Information Models for Prediction

Algorithmic Information Theory (2024/2025)
University of Aveiro

Alexandre Ribeiro (108122)
Maria Sardinha (108756)
Miguel Pinto (107449)

Index:



INTRODUCTION

- Project extends our previous work on Finite-Context Models (FCMs).
- Transitions from natural language to genetic sequences.
- **Goal:** Develop a metagenome classifier using Normalized Relative Compression.
- Estimate which known organisms (from a given reference database) share the highest similarity with the genomes found in the sample.
- **Evaluate:**
  - Classifier’s performance.
  - Impact of model parameters.
  - Optimal configurations.

THEORETICAL FOUNDATION

- Based on information theory (Shannon) and algorithm complexity (Kolmogorov).
- Normalized Information Distance (NID) provides a theoretical measure of similarity between any 2 objects.
- DNA sequences particularly suitable for compression-based analysis:
  - Limited alphabet (A, C, G, T).
  - Contains repetitive elements at different scales.
  - Evolutionary conservation creates shared patterns.

NORMALIZED RELATIVE COMPRESSION

- Measures similarity between sequences, based on compression efficiency.
- Formula: $$\mathrm{NRC}(x \parallel y) = \frac{C(x \parallel y)}{|x| \log_2(A)}$$
  - Where:
    - *C(x || y)* is the number of bits needed to losslessly compress *x* given exclusively a model trained with *y*.
    - *|x|* is the size of the sequence *x*.
    - *log2(A)* is the log of the alphabet size of sequence *x*.
- Lower NRC values indicate greater similarity

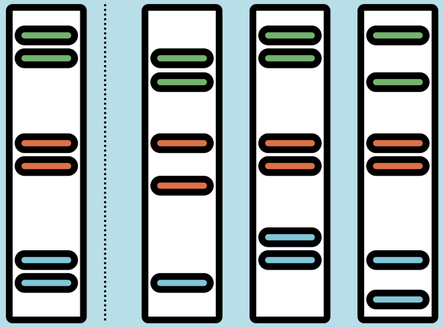
ADVANTAGES OF COMPRESSION-BASED CLASSIFICATION

- **Alignment-Independence:** Detects similarity even with rearrangements.
- **Computational Efficiency:** Scales well for large datasets.
- **Sensitivity to Higher-Order Patterns:** Detects subtle shared patterns.
- **Natural Confidence Measure:** Ranking provides classification confidence.
- **Handles Mixed Samples:** Identifies dominant organisms without separation.

SYSTEM ARCHITECTURE (1)

- **FCMModel:** Core implementation of the Finite-Context Model:
  - Maintains frequency tables for context-symbol occurrences.
  - Implements Laplace smoothing with parameter α.
  - Supports model serialization/ /deserialization.
  - Methods for entropy calculation and sequence prediction.

- **MetaClass:** Command-line application:
  - Processes parameters.
  - Manages training workflows and calculating metrics.
  - Ranks and reports results.

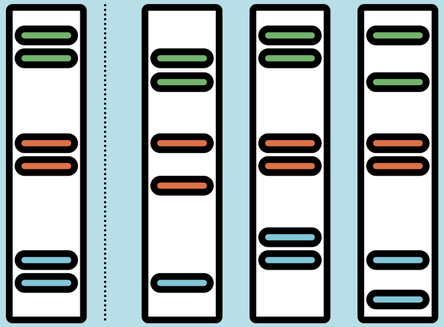


SYSTEM ARCHITECTURE (2)

- **Testing Framework:**
  - Parameter testing (*k* and α).
  - Generates performance reports.
  - Model performance metrics across parameter combinations.

- **Utility Files:**
  - DNA sequence preprocessing.
  - Reference database loading.
  - Assist other modules (provides reusable code).

- **Visualization Tools:**
  - Based on obtained results.
  - Better perception of tests.



METRICS COLLECTED

NRC (Normalized Relative Compression)**:** Primary similarity metric.

KLD (Kullback-Leibler Divergence)**:** Measures distribution differences.

Compression Bits: Theoretical minimum bits needed for encoding.

Time: Processing Time for each parameter combination.

Ranking: Orders references by similarity to the query sequence.

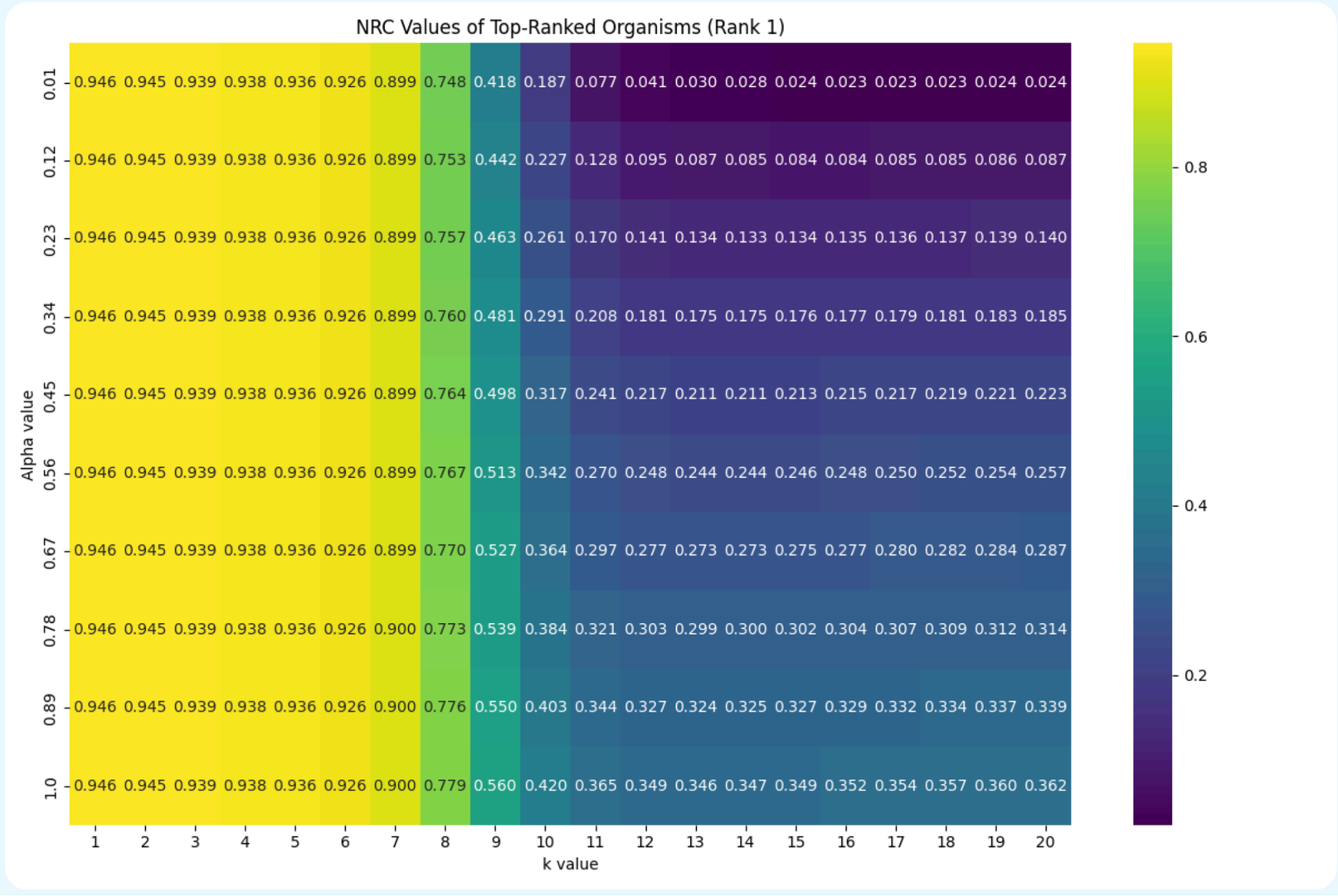
Classification Metrics: Accuracy, Precision, Recall, F1-Score, and ROC AUC (for synthetic testing).

METHODOLOGY

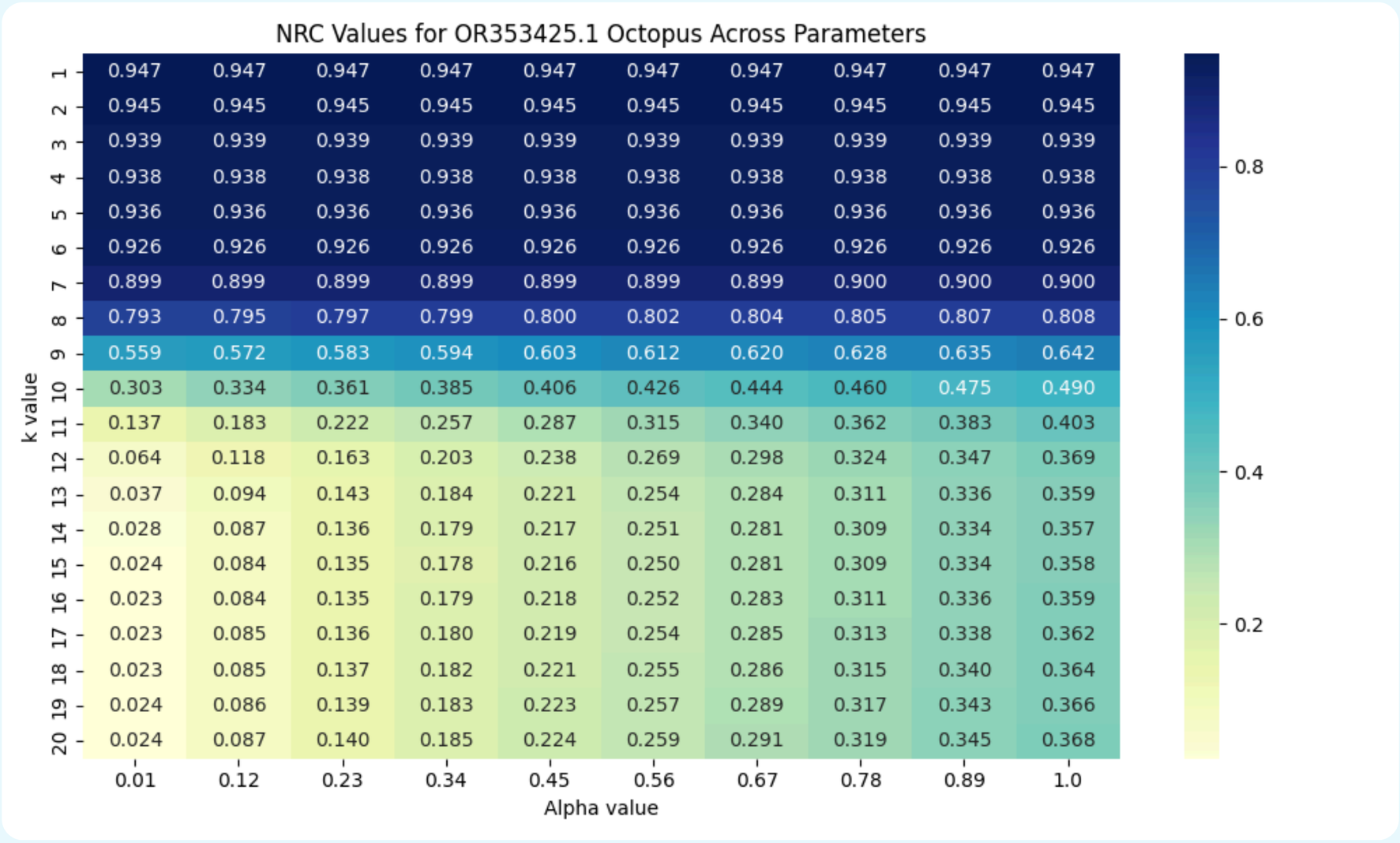
1. Train the Finite-Model on metagenomic sample (*y*).
2. Freeze the model counts (no further updates).
3. For each reference sequence (*xi*):
    - Estimate the bits required to compress *xi* using the model trained on *y*.
    - Calculate the NRC to measure relative compression.
4. Sort the reference sequences by their NRC value.
5. Output the top 20 most similar sequences.

RESULTS VISUALIZATION & ANALYSIS

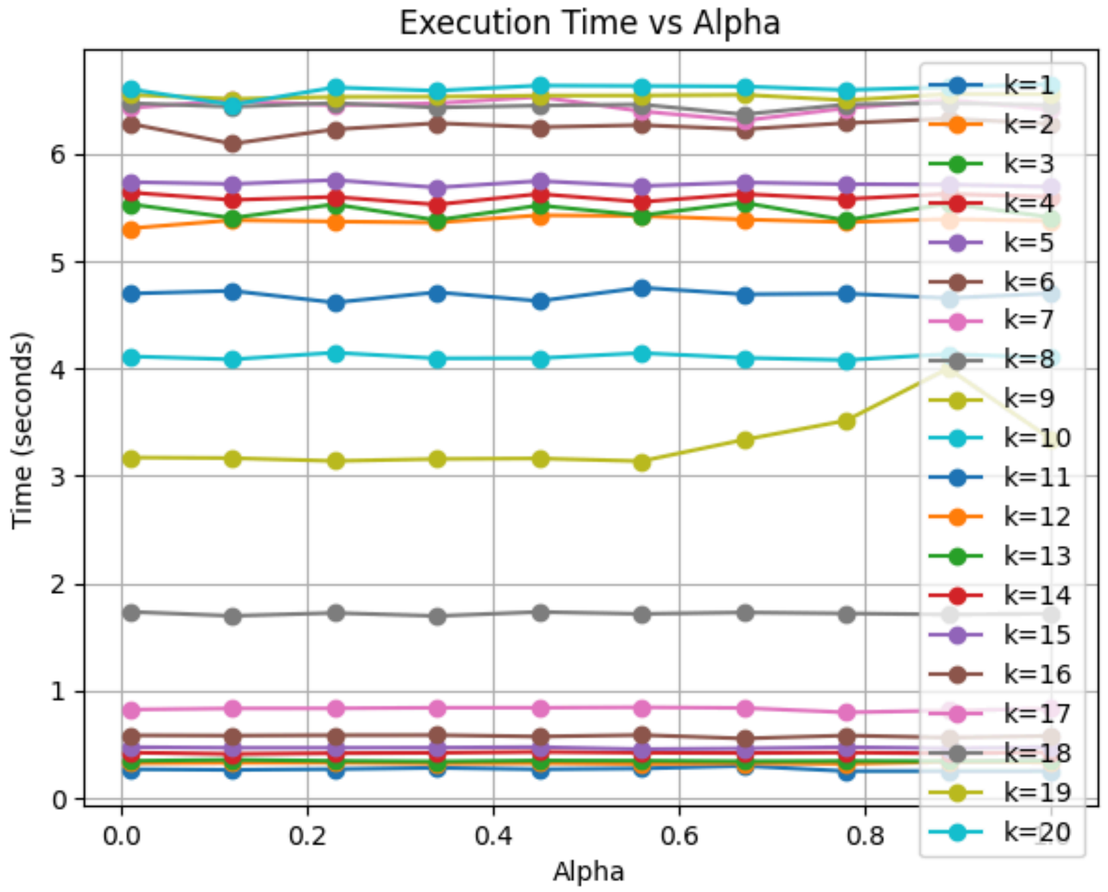
PARAMETER IMPACT ANALYSIS - HEATMAP (NRC)

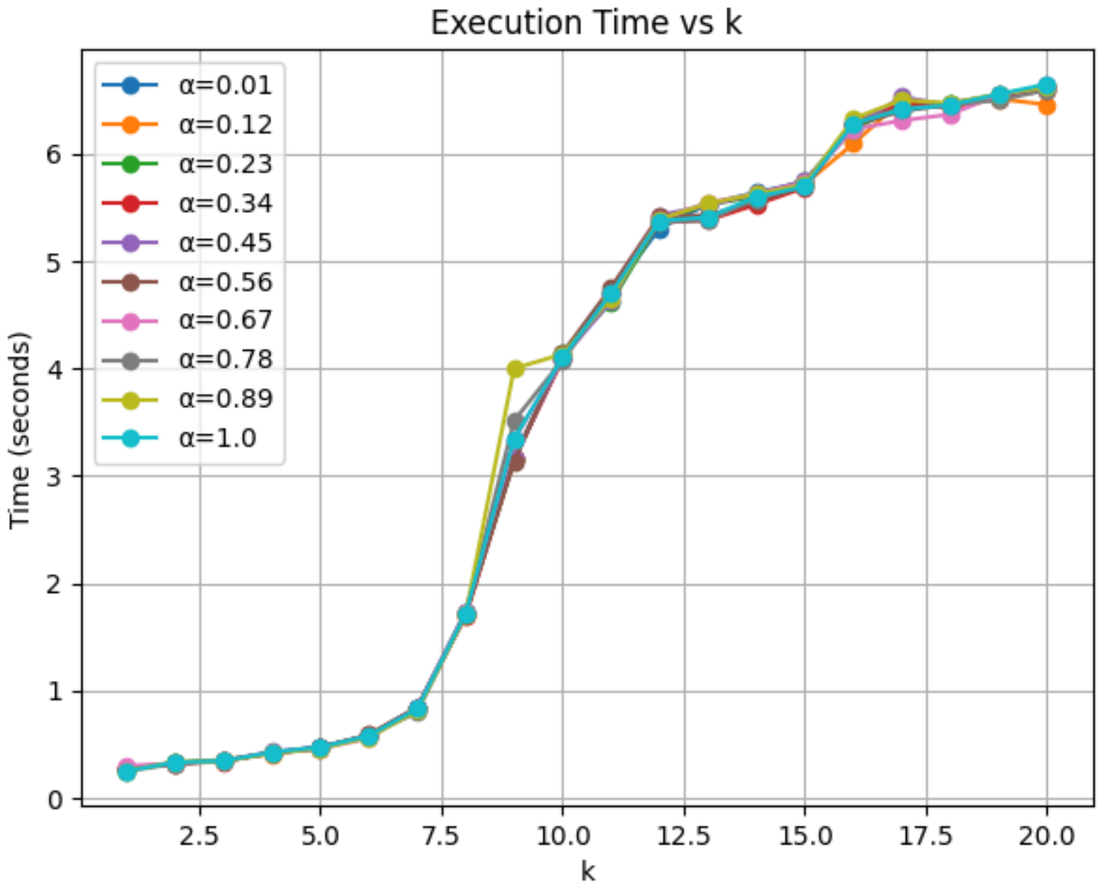


PARAMETER IMPACT ANALYSIS - OCTOPUS EXAMPLE

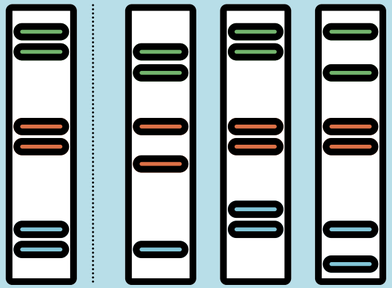


EXECUTION TIME





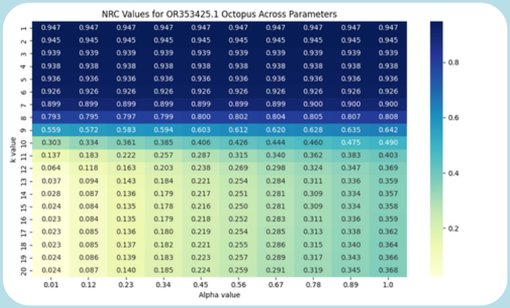
PARAMETER IMPACT ANALYSIS



- **Context Size (*k*):**
  - NRC values are more stable and lower for *k* around 17.

- **Smoothing Parameter (*α*):**
  - NRC values are more stable and lower for $\alpha \approx 0.01$.
  - Higher values of *α* tend to over-smooth

- **Optimal Values Obtained:**
  - *k* = 17.
  - α = 0.01

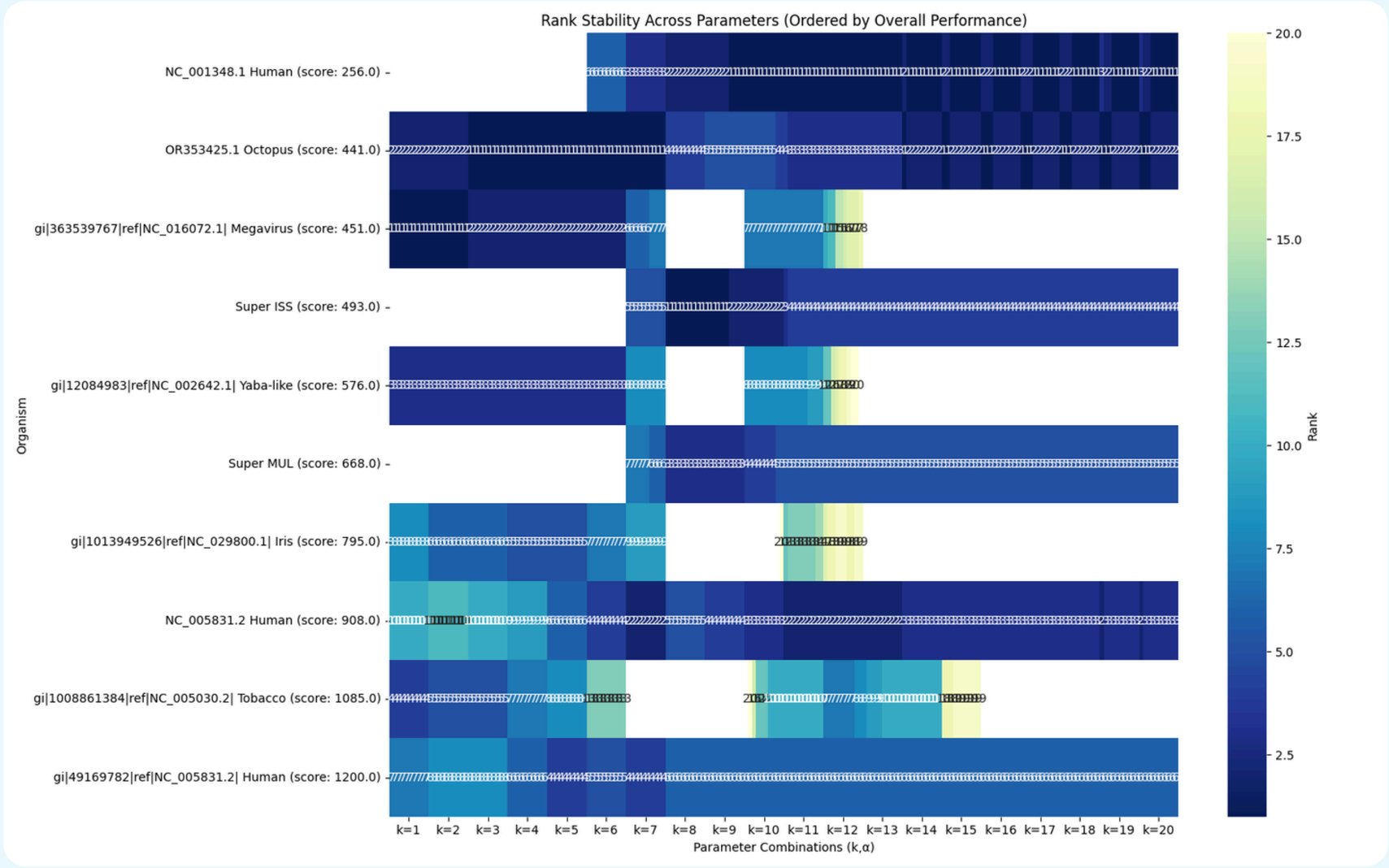


RESULTS - TOP 10 NRC-RANKED ORGANISMS

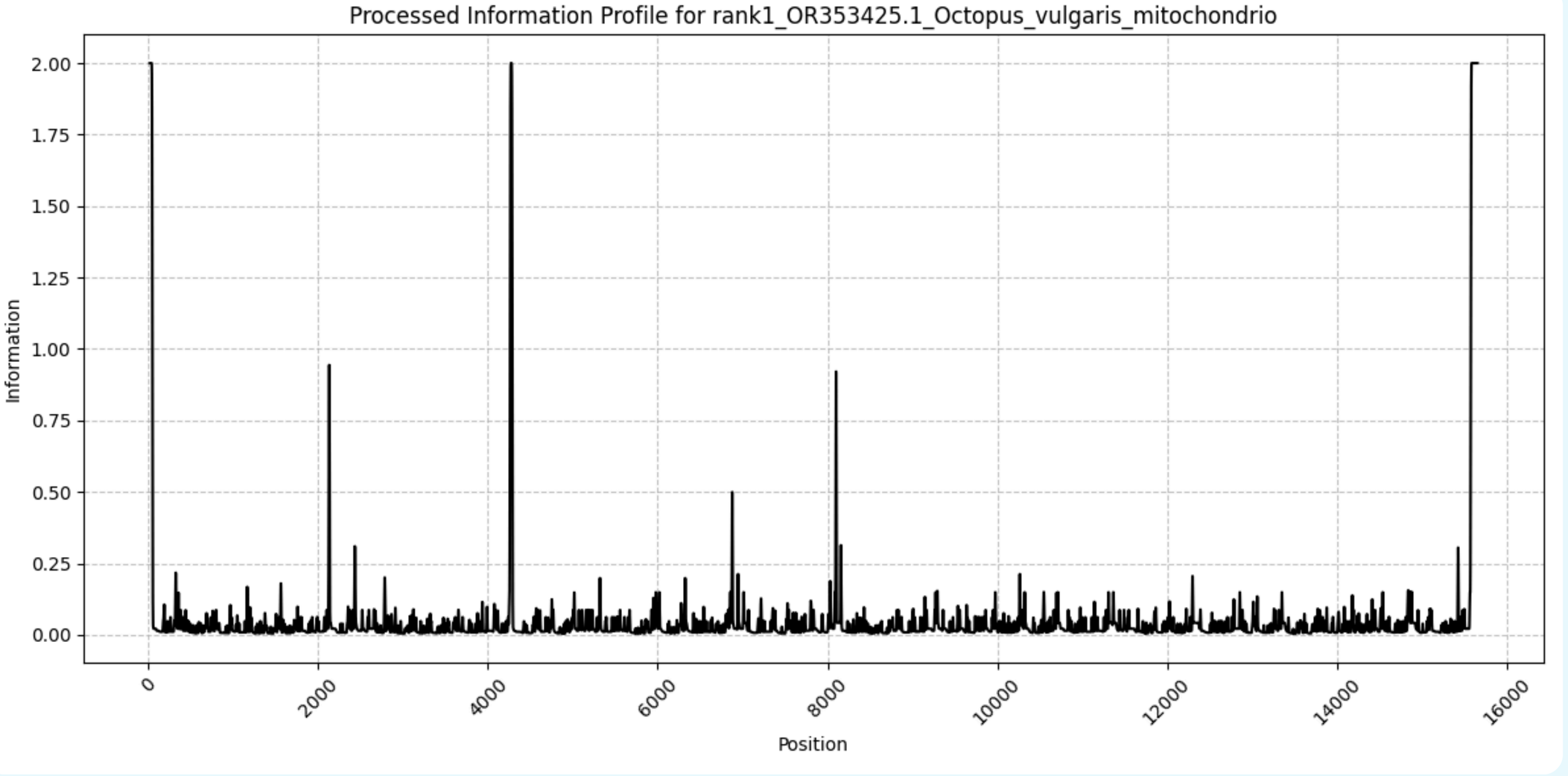
K = 17, ALPHA = 0.01

Rank	Organism Name	NRC	KLD
1	OR353425.1 Octopus vulgaris mitochondrion, complete genome...	0.023185	719.722054
2	NC 001348.1 Human herpesvirus 3, complete genome...	0.028285	6976.290343
3	NC 005831.2 Human Coronavirus NL63, complete genome...	0.028549	1570.720935
4	Super ISS Si1240...	0.106828	264.934522
5	Super MUL 720...	0.197829	284.873124
6	NC 005831.2— Human Coronavirus NL63, complete genome...	0.426500	23500.206077
7	NC 025220.1— Sweet potato leaf curl virus associated satelli...	0.976808	1432.000000
8	NC 003847.1— Panicum mosaic satellite virus, complete genome...	0.979419	1618.000000
9	NC 024075.1— Cat Que Virus strain VN04-2108 nucle-oprotein ge...	0.982706	1932.000000
10	NC 029618.1 Lake Sarah-associated circular molecule 7 ...	0.984489	2158.000000

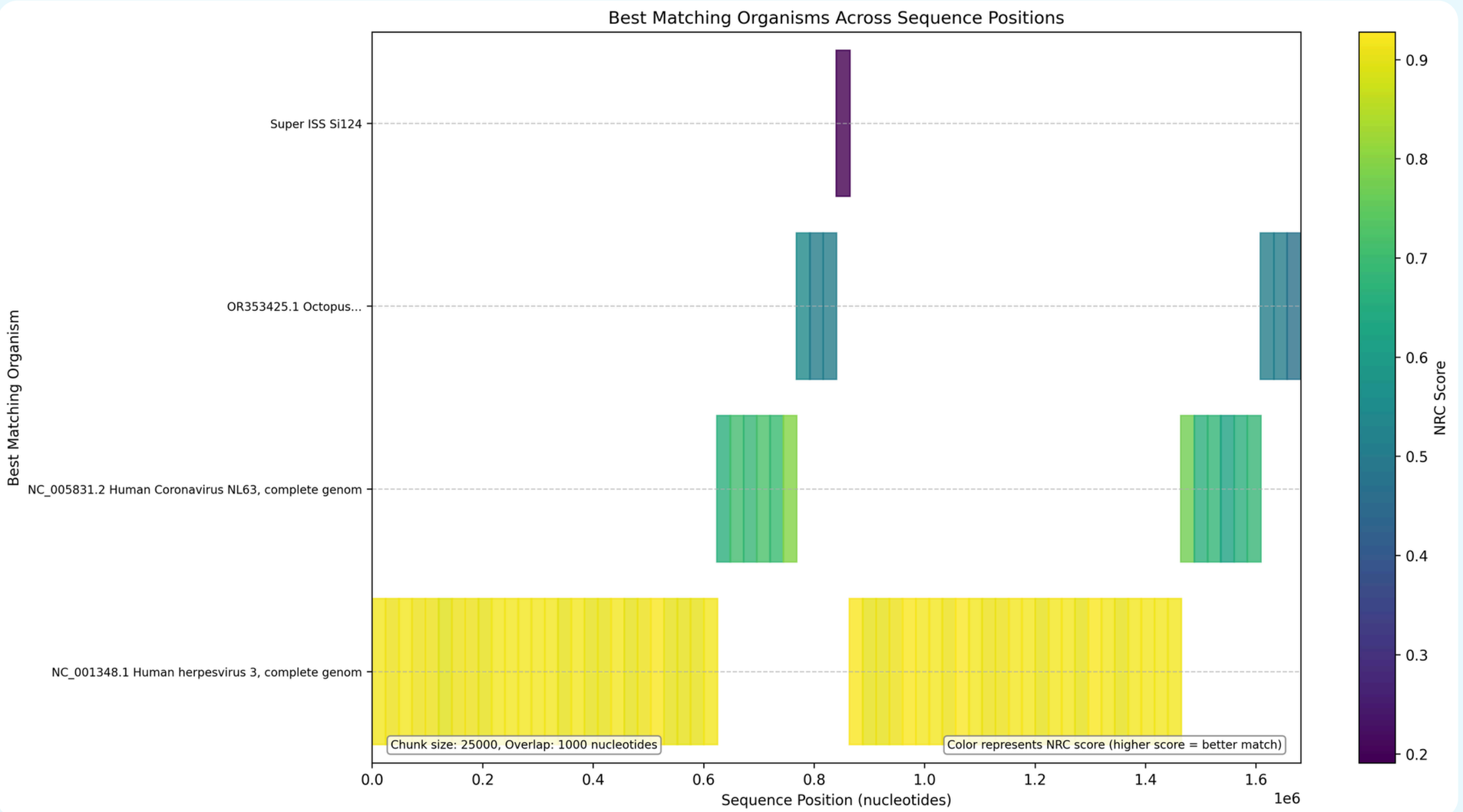
RANK STABILITY ACROSS PARAMETERS



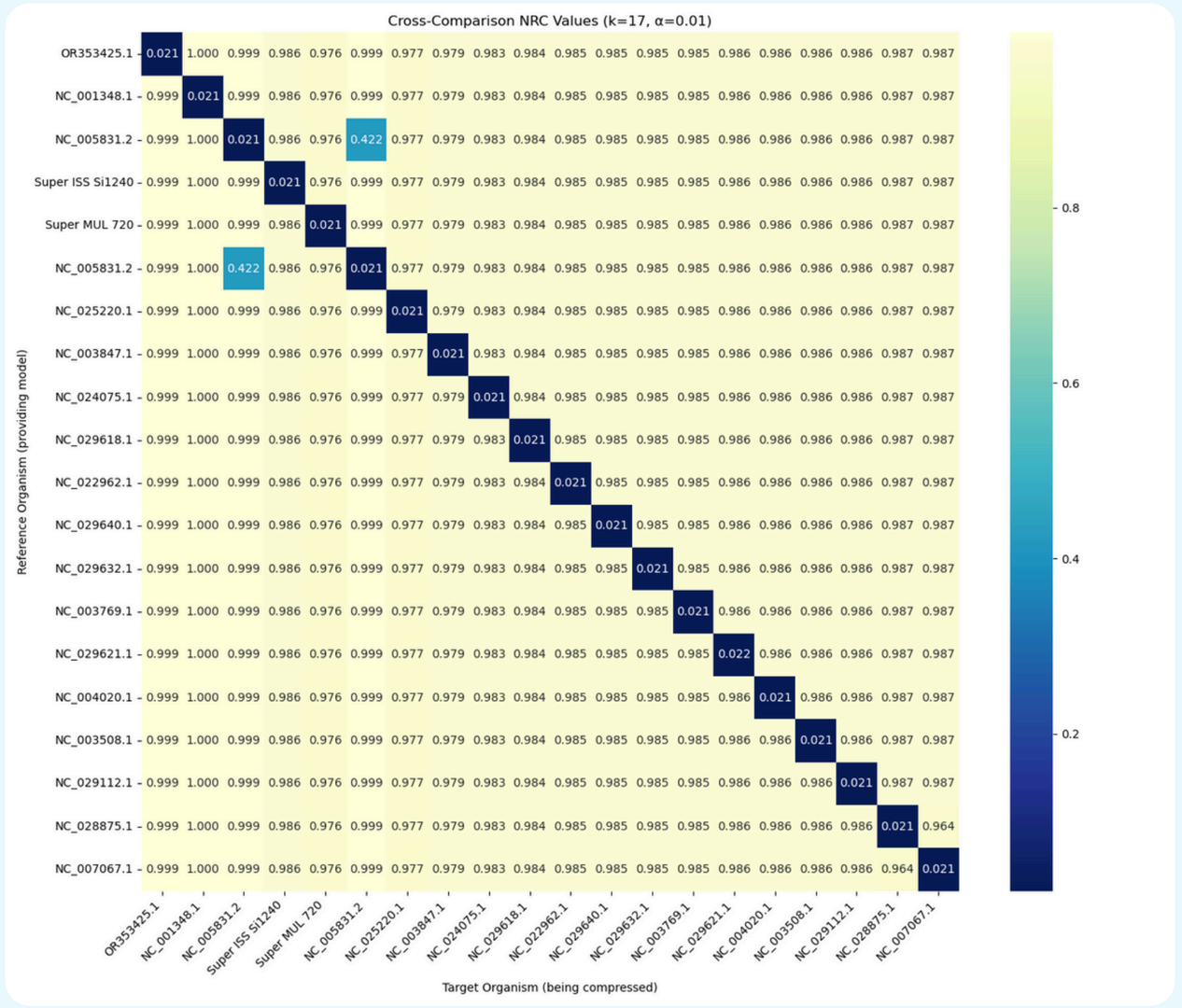
INFORMATION PROFILE - OCTOPUS EXAMPLE



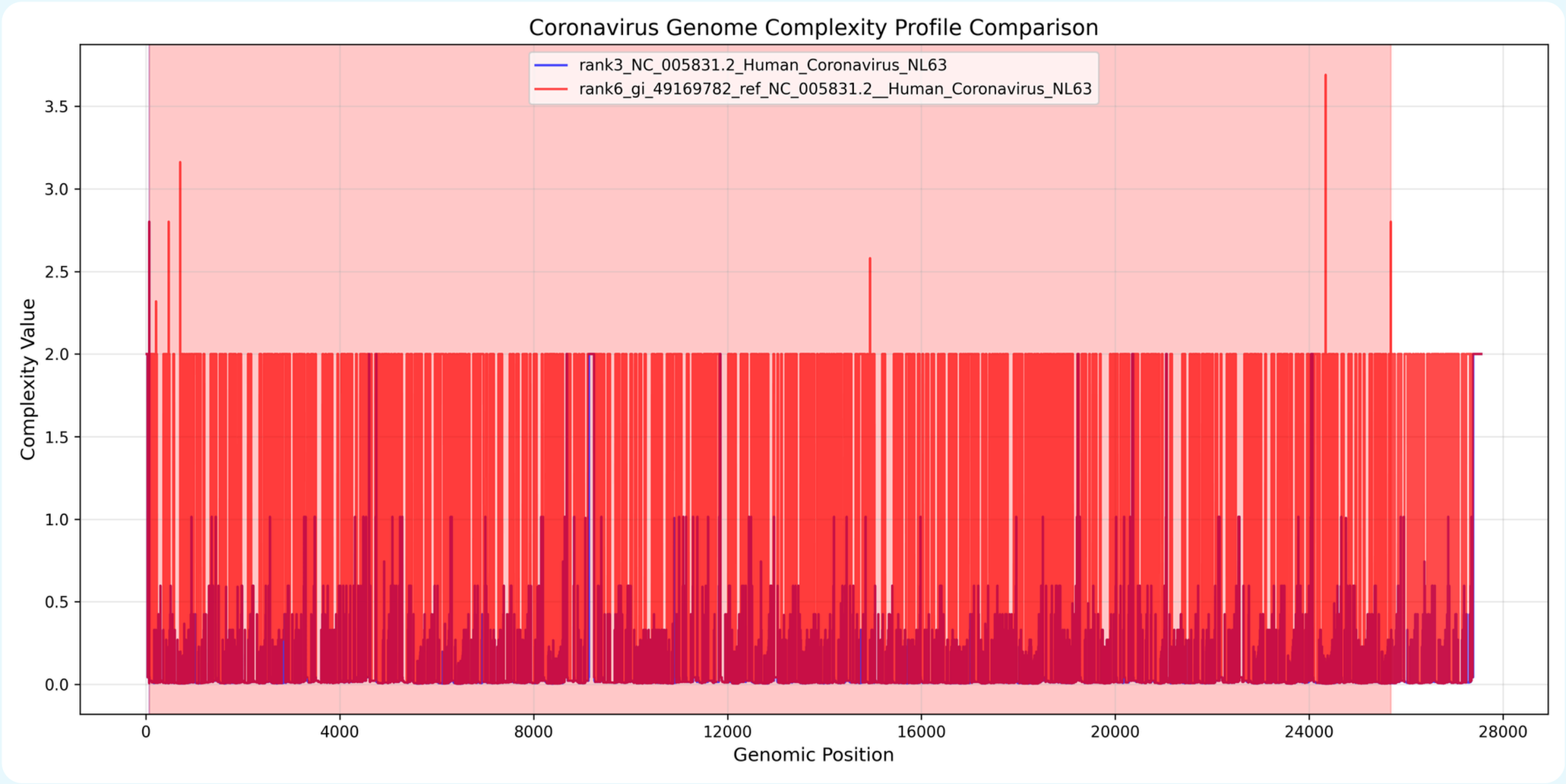
SEQUENCE-LEVEL CHUNK ANALYSIS



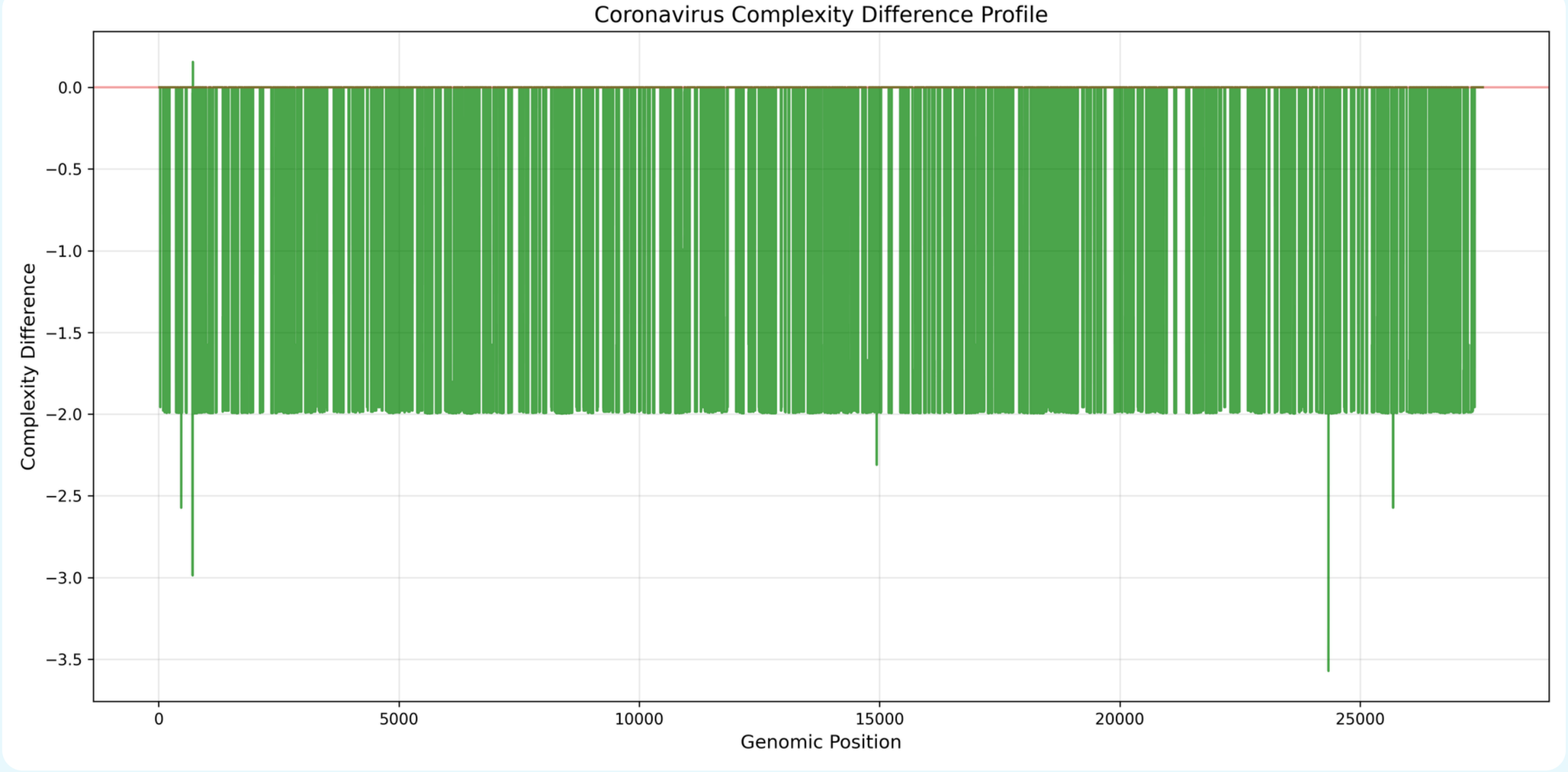
NRC HEATMAP - CROSS COMPARISON



HUMAN CORONAVIRUS GENOMES - COMPARISON



HUMAN CORONAVIRUS GENOMES - COMPARISON



SYNTHETIC DATA - ANALYSIS

K = 8, ALPHA = 0.01, THRESHOLD = 0.5

```
==================================================
                EVALUATION RESULTS
==================================================
NRC Threshold: 0.500000

Confusion Matrix:
--------------------------------------------------
                |     Actual     |     Actual     |
                |    Positive    |    Negative    |
--------------------------------------------------
 Predicted      |         17     |          0     |
 Positive       |                |                |
--------------------------------------------------
 Predicted      |          3     |         80     |
 Negative       |                |                |
--------------------------------------------------

Metrics:
--------------------------------------------------
Accuracy:  97.0000%
Precision: 100.0000%
Recall:    85.0000%
F1 Score:  0.9189
ROC AUC:   1.0000
--------------------------------------------------
```

```
Top 20 matches by NRC:
--------------------------------------------------
Rank |        NRC |     Status | Reference
--------------------------------------------------
   1 |   0.151432 |   TRUE POS | sequence_83_meta
   2 |   0.151577 |   TRUE POS | sequence_86_meta
   3 |   0.152661 |   TRUE POS | sequence_84_meta
   4 |   0.153892 |   TRUE POS | sequence_87_meta
   5 |   0.154849 |   TRUE POS | sequence_81_meta
   6 |   0.160097 |   TRUE POS | sequence_85_meta
   7 |   0.160648 |   TRUE POS | sequence_88_meta
   8 |   0.161236 |   TRUE POS | sequence_82_meta
   9 |   0.162995 |   TRUE POS | sequence_90_meta
  10 |   0.166849 |   TRUE POS | sequence_89_meta
  11 |   0.231097 |   TRUE POS | sequence_98_mut_88
  12 |   0.282013 |   TRUE POS | sequence_91_mut_81
  13 |   0.412662 |   TRUE POS | sequence_97_mut_87
  14 |   0.470191 |   TRUE POS | sequence_94_mut_84
  15 |   0.484995 |   TRUE POS | sequence_96_mut_86
  16 |   0.487677 |   TRUE POS | sequence_99_mut_89
  17 |   0.489892 |   TRUE POS | sequence_93_mut_83
  18 |   0.543829 |  FALSE NEG | sequence_95_mut_85
  19 |   0.548875 |  FALSE NEG | sequence_100_mut_90
  20 |   0.631181 |  FALSE NEG | sequence_92_mut_82
--------------------------------------------------

False negatives (missed sequences that should be detected): 3
```

KEY FINDINGS

- Compression-based metrics effectively capture biological relationships without alignment.
- Parameter optimization is crucial for meaningful results.
- Combinations of NRC and KLD provide complementary information.
- Chunk analysis potentially indicates horizontal gene transfer or mixed-origin content.

FUTURE WORK

- **Adaptive parameter selection:** Automatically select optimal parameters.
- **Comprehensive Benchmarking:** Compare with establish methods (BLAST, Kraken, etc...).
- **Hierarchical Classification:** Hierarchy-aware classification to improve accuracy by considering evolutionary relationships.
- **Pipeline Integration:** Integrating with existing bio-informatics workflows for broader adoption.

CONCLUSIONS

- Successfully adapted FCMs for DNA sequence analysis.
- Implemented NRC as an effective distance measure.
- Developed a comprehensive testing framework.
- Validated our model with synthetic data.
- Provided accessible command-line interface.

METACLASS DEMO

```
→ TAI_Project2 git:(dev) ✗ mkdir -p build
→ TAI_Project2 git:(dev) cd build
→ build git:(dev) cmake ..
-- The CXX compiler identification is GNU 14.2.0
-- Detecting CXX compiler ABI info
-- Detecting CXX compiler ABI info - done
-- Check for working CXX compiler: /usr/bin/c++ - skipped
-- Detecting CXX compile features
-- Detecting CXX compile features - done
-- Configuring done (0.1s)
-- Generating done (0.0s)
-- Build files have been written to: /home/maria/Desktop/TAI_Project2/build
→ build git:(dev) cmake --build .
[ 10%] Building CXX object src/CMakeFiles/tai_src.dir/core/FCMModel.cpp.o
[ 20%] Building CXX object src/CMakeFiles/tai_src.dir/utils/dna_compressor.cpp.o
[ 30%] Building CXX object src/CMakeFiles/tai_src.dir/utils/io_utils.cpp.o
[ 40%] Building CXX object src/CMakeFiles/tai_src.dir/utils/interface_utils.cpp.o
[ 50%] Building CXX object src/CMakeFiles/tai_src.dir/utils/test_utils.cpp.o
[ 60%] Linking CXX static library libtai_src.a
[ 60%] Built target tai_src
[ 70%] Building CXX object apps/CMakeFiles/MetaClass.dir/MetaClass.cpp.o
[ 80%] Linking CXX executable MetaClass
[ 80%] Built target MetaClass
[ 90%] Building CXX object apps/CMakeFiles/tests.dir/tests.cpp.o
[100%] Linking CXX executable tests
[100%] Built target tests
→ build git:(dev) ./apps/MetaClass -d ../data/samples/db.txt -s ../data/samples/meta.txt -k 17 -a 0.01 -t 20
Reading metagenomic sample from: ../data/samples/meta.txt
Metagenomic sample length: 1696500 nucleotides
Training FCM model with k=17, alpha=0.01
Reading reference database from: ../data/samples/db.txt
Found 239 reference sequences
Calculating metrics using 16 threads...
```

THANK YOU!

GitHub Repository: https://github.com/mariiajoao/TAI_Project2