STU FIIT
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

NOVEL DNA ALLIGNMENT-FREE COMPARISON
BASED ON SIGNAL PROCESSING APPROACHES

Tomáš Farkaš
supervisor: Mária Lucká

ABSTRACT

Computing similarity between two nucleotide sequences is one of fundamental problems in bioinformatics. Current methods are based on two major approaches: Full sequence alignment, which is computationally expensive, and faster, but less accurate alignment-free methods based on various statistical summaries, e.g. short word counts.
We propose three novel methods based on signal processing transforms designed for overcoming specific sequence features while requiring only modest omputational resources. Our approaches include spectral transforms computed across a smoothed sequence, sliding windows or multiple resolution windows. The experiments reveal that the novel methods are up to three times more accurate than current alignment-free methods, while they are equally computationally inexpensive.

INTRODUCTION

Hierarchical clustering
- Creating a tree diagram of species to show mutual relations
- Need to evaluate similarity of biological species - entites only by DNA sequencies belonging to them.

DNA sequence specifities
- Subsequence insertions, deletions, transpositions ...
  - Similar sequence parts (that need to be evaluated as similar) differ in spatial position and order
  - Inaccurate spatial information

CURRENT METHODS

Full sequence allignment
- Sequence allignement = way of arranging sequences by adding spaces to allign similar parts
- Number of spaces added -> metric of similarity
- Cannot process sequences with transposed parts
  - Not applicable on raw data
- Very computationally expensive

Statistic methods
- Comparing numercal characteristics computed from original string sequences
  - Word Frequencies (A)
  - Spaced Word Frequencies (B)
- Inaccurate

Early spectral transformation methods
- Relaxing strict dependency on clumsy spatial information
  - Comparing raw signal spectra gathrered from a sequence
  - Comparing numerical characteristics computed from signal spectrum (C)
- Need of good numeric representation of the input string sequencies

NEW METHODS

SLIDING WINDOW METHOD (E,F)

Windowed processing
- No problem with transposed sequence parts

1. Split the sequence into windows
2. Compute spectral transform of each window (FFT, ...)
3. Sum all spectral vectors into one resulting vector
4. Normalize by the number of windows procecessed.

5. Compare resulting vectors by generic distance fuctions (e.g. Euclidean) to evaluate metric dissimilarity.

The method can be further expanded by using multiple resolution windows to compute higher spectal coefficients with better precision (F)

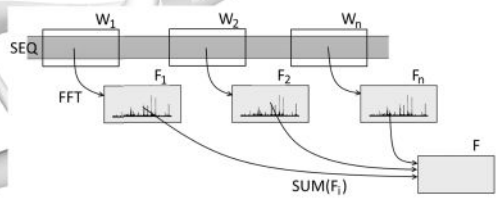


Fig. 4: Visualization of the Sliding window method. Sequence is denoted as SEQ, windows W1, ... , Wn. Spectral vectors F1, ... , Fn (with equal length) are summed into resulting vector F.

DOMINANT COEFFICIENT METHOD (D)

Processing whole sequences
- Expecting better accuracy on data with high level of mutual dissimilarity

1. Smooth the representative sequence by a smoothing window (e.g. Blackmann-Nutall) to benefit from side effects (loss of resolution, scalloping)
2. Pad the sequence with zeros to the highest common sequence length
3. Compute spectral transform (FFT, WHT, ...)
4. Select a predefined number of dominant = highest spectral coefficients in absolute value (e.g. top 10%)

5. Evaluate number of position matches between each pair of sequences. The result becomes the metric of mutual sequence similarity

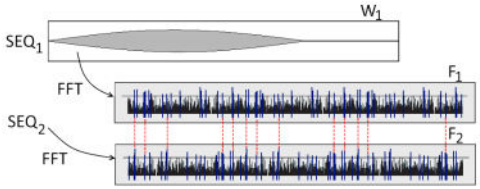
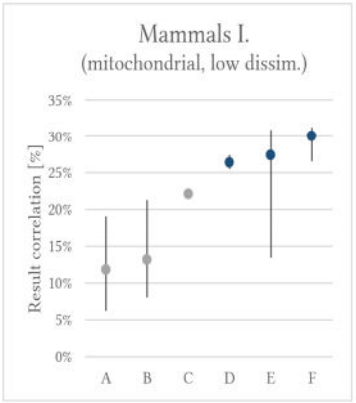
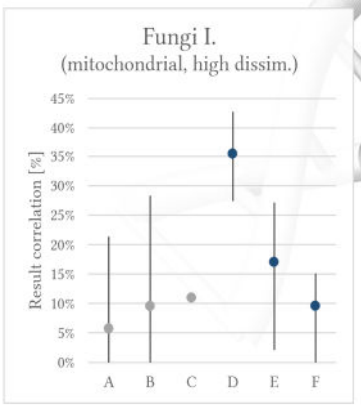


Fig. 5: Visualization of the Dominant spectral coeffiecient method. Padded and smoothed sequence is denoted as SEQ1, its spectrum as F1.Dominant coefficients are marked blue, position matches dashed red.





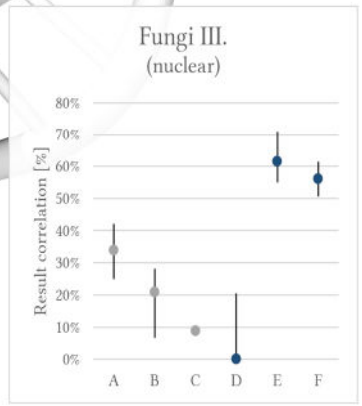


Fig. 1 (right) : Method accuracy on Mammals I. data. Data contain no transpositions implying so good accuracy of methods C and D that would otherwise fail. Our methods are marked D, E and F.

Fig. 2 (middle): Method accuracy on Fungi I. data. Data contain no or few transpositions and sequences have higher amount of dissimilarity. This theoretically favours method D, which shows clear superiority, however E still performs better than current established methods.

Fig. 3 (left): Method accuracy on Fungi III. data. Raw nuclear sequences contain frequent rearrangements leading to weak accuracy of C and D. However our windowed methods E and F show clear superiority against other windowed methods (A, B).

REFERENCES

- Yin, C., Chen, Y., and Yau, S. S.-T. S. (2014). A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering. Journal of theoretical biology, 359, 18–28
- Song, K., Ren, J., Reinert, G., Deng, M., Waterman, M. S., and Sun, F. (2013). New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. Briefings in Bioinformatics, 15(3), 343.

ACKNOWLEDGEMENT
This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG 1/0752/14.