PanPA: Construction and Alignments of Panproteome Graphs

Fawaz Dabbaghie[1,2], Sanjay K. Srikakulam[2,3,4], Olga V. Kalinina[2,5,6], Tobias Marschall[1]

[1]Heinrich Heine University Düsseldorf, Medical Faculty, Institute for Medical Biometry and Bioinformatics, 40225 Düsseldorf, Germany, [2]Helmholtz Institute for Pharmaceutical Research Saarland (HIPS) / Helmholtz Centre for Infection Research, 66123 Saarbrücken, Germany, [3]Graduate School of Computer Science, Saarland University, Campus E3.1, 66123 Saarbrücken, Germany, [4]Interdisciplinary Graduate School of Natural Product Research, Saarland University, 66123 Saarbrücken, Germany, [6]Drug Bioinformatics, Medical Faculty, Saarland University, 66421 Homburg, Germany, and [6]Center for Bioinformatics, Saarland University, 66123 Saarbrücken, Germany

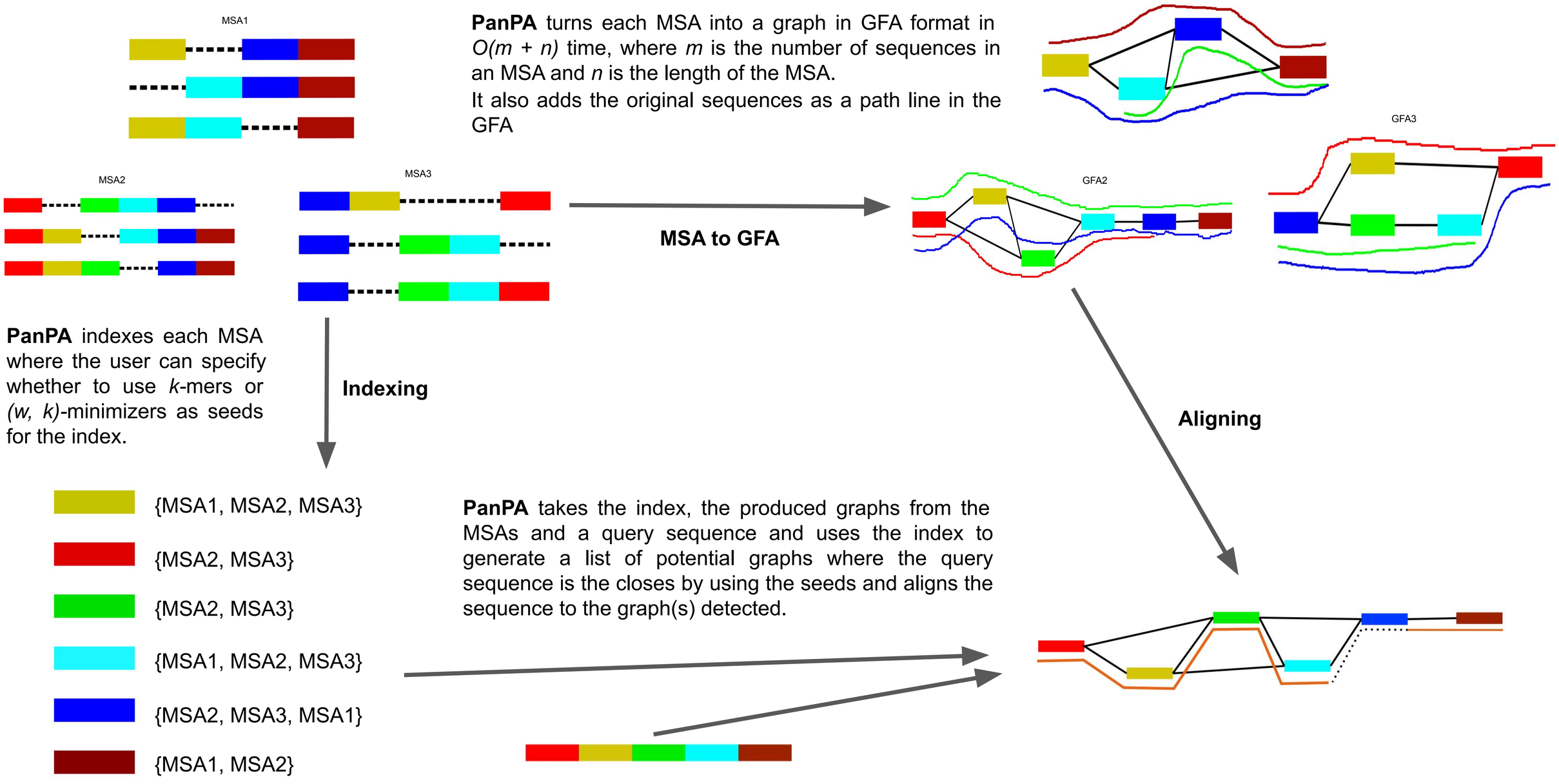
Compared to *eukaryotes,* ***prokaryote*** genomes are much more diverase through different mechanisms, including a higher mutation rate and horizontal gene transfer. Therefore, using a linear DNA reference cannot capture the full diversity spectrum within and between prkaryotic species and workflows relying on a linear reference will exhibit strong biases.

Graph-based methods have been developing rapidly to combat the linear reference bias issue. However, working with DNA sequences is still challenging in interspecies and within clades comparison.

In contrast, amino acid sequences have higher similarity as they are more susceptible to selection and several DNA codons result in the same amino acid

As coding regions cover the majority of prokaryotic genomes, we can build panproteome graphs from individual genes and gene clusters instead of the complete genome, which leverages the fact that amino acid are more conserved to make comparisons over longer phylogenetic distances.

PanPA is designed to tackles this problem in three main steps.

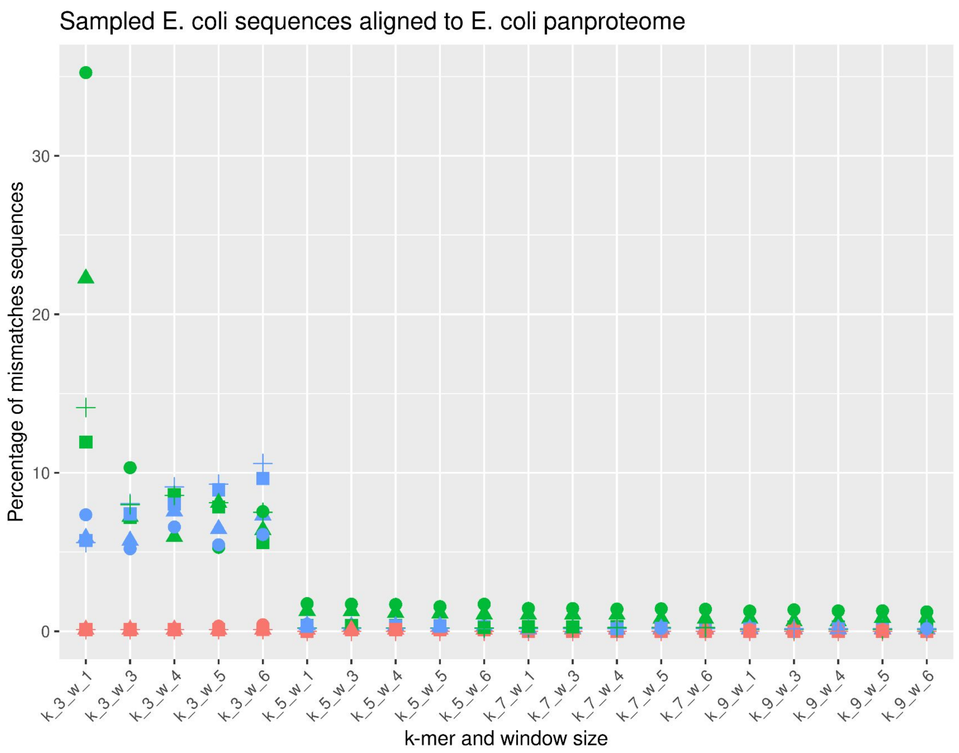


PanPA turns each MSA into a graph in GFA format in $O(m + n)$ time, where m is the number of sequences in an MSA and n is the length of the MSA.

It also adds the original sequences as a path line in the GFA

PanPA indexes each MSA where the user can specify whether to use *k*-mers or *(w, k)*-minimizers as seeds for the index.

PanPA takes the index, the produced graphs from the MSAs and a query sequence and uses the index to generate a list of potential graphs where the query sequence is the closes by using the seeds and aligns the sequence to the graph(s) detected.



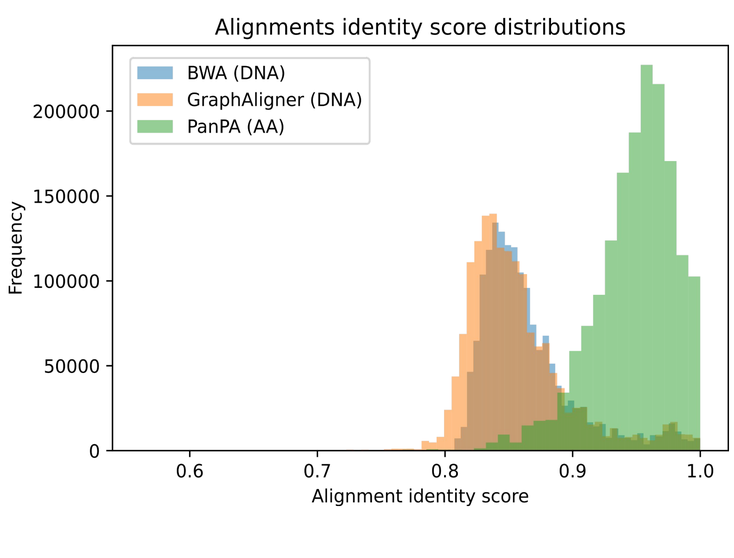
We used 1,350 *E. coli* assemblies from RefSeq to test **PanPA**. We extracted all proteins from each assembly and clustered them into individual cluster, each cluster turned into MSA, then turned into a graph and indexed.

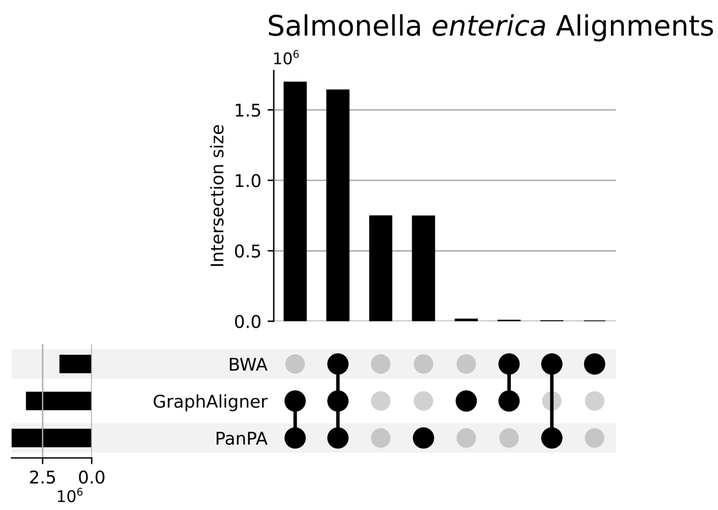
To test if PanPA is doing correct alignment, we extracted ~30,000 sequences randomly from the clusters and aligned them back to the graph, we see in the figure above that for most parameter combinations, we have 0 or almost 0 percentage of mismatches, i.e. most of the sequences matched back to the graphs where they came from

To demonstrate that working in the amino acid space allows us to make comparison with more distant organisms. We extracted proteins as DNA and amino acid sequences from 1,073 *Salmonella enterica* assemblies from RefSeq and aligned these sequences back in three different ways using three difference methods::

1. *E. coli* reference genome using **BWA**
2. *E. coli* pangenome from the 1,350 assemblies using **GraphAligner**
3. *E. coli* panproteome generated by **PanPA**

The results show that our alignments had higher identity scores, and PanPA was able to capture about 22% more alignments that were not detected by the other methods





Website

Twitter