Alignment-free Recombination Detection Using Genomic Database Distributions of Exact Protein Matches

Program name: **Redcarpet (Recombination Detection using Comparative Analysis of Regional Patterns of Exact Match Targets)**

Ahmed M Moustafa[1,2], Arnav Lal[3], Andries Feder[1], Apurva Narechania[4], Paul J. Planet[1,4,5]

[1]Division of Infectious Diseases; [2]Division of Gastroenterology, Hepatology & Nutrition, Children's Hospital of Philadelphia, Philadelphia, PA; [3]School of Arts and Sciences, University of Pennsylvania , Philadelphia, PA; [4]American Museum of Natural History, New York, NY; [5]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

Email: moustafaam@chop.edu

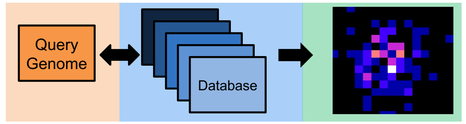
BACKGROUND

Genomic recombination is a generator of biological diversity and plays an important role in microbial adaptation to new environments, hosts, and niches. Recombination detection previously relied on genomic sequence alignment and phylogenetic or comparative techniques that are computationally expensive, especially with vast increases in available whole genome sequences.

Here we present a database-driven technique that is alignment-free, does not rely on phylogeny or sequence similarity, and can be used rapidly on single genomes on a standard desktop. The tool, Redcarpet, combines the analytic output from our recently developed WhatsGNU algorithm[1] with a MinHash technique[2]. This operation is predicated on the idea that identical genes are more likely to appear in the same set of genomes if they share a common evolutionary history

METHODOLOGY

Redcarpet takes in a **single query genome**, and for each encoded protein, determines **the set of genomes in a database** that contain an exact protein sequence match. It then **computes the Jaccard similarity coefficient between genome sets** for all pairwise protein comparisons in the genome.



CALCULATION THEORY

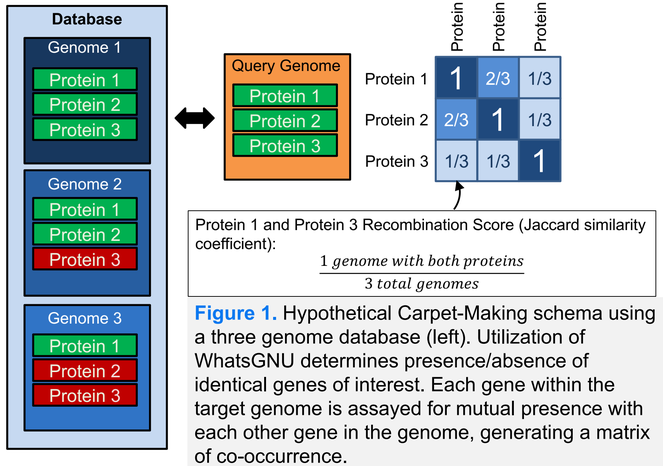


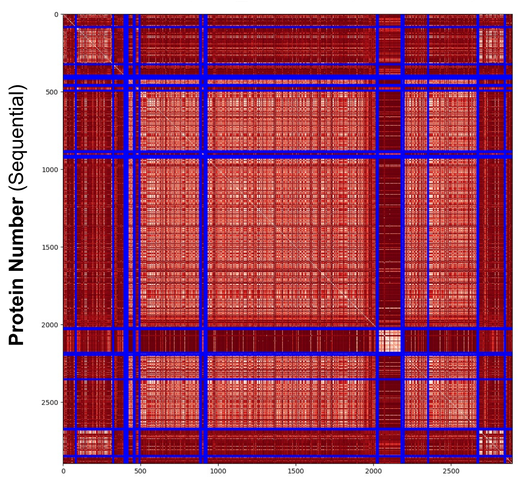
Figure 1. Hypothetical Carpet-Making schema using a three genome database (left). Utilization of WhatsGNU determines presence/absence of identical genes of interest. Each gene within the target genome is assayed for mutual presence with each other gene in the genome, generating a matrix of co-occurrence.

RESULTS

Recombination

Recombination Heatmap

S. aureus TW20 (GCA_00027045.1):Sequence Type 239



The output of Redcarpet is a pairwise, genome-set similarity matrix that can be visualized as a 2-D heatmap ordered by the gene location on the chromosome. The heatmap provides a visual tool for identifying recombination tracts.

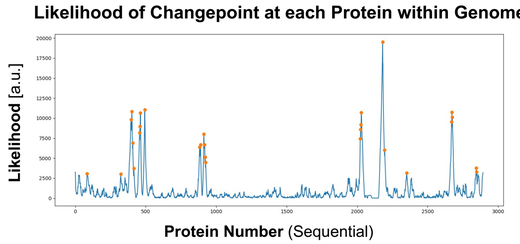
Beyond recombination detection, this tool can be further expanded to explore both (1) the phylogenetic history of intra-genomic recombinant regions and (2) the likely genomic regions of recombination.

Figure 2. Similarity Matrix (Redcarpet) for *Staphylococcus aureus* genome (GCA_000027045), within ST239 that is known for large-scale evolutionary recombination. Use of 10,350 publicly-accessible NCBI *S. aureus* genomes were utilized as the comparative database in the generation of this heatmap.

This analysis reveals the potential locations of recombination ("squares" within the heatmap), with blue lines added to depict the edge of the squares (i.e., the changepoints), as calculated within Figure 3.

Changepoint Analysis

Likelihood of Changepoint at each Protein within Genome



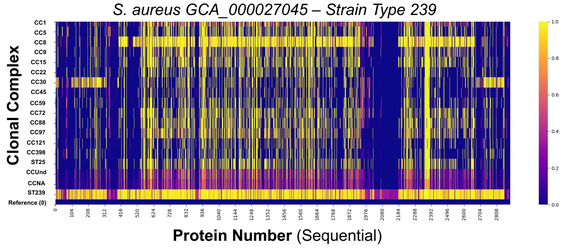
The recombinant changepoints, defined as likely location of recombination were calculated by comparing the protein values within a set number (~30) of proteins away from each protein in the genome. The local maxima of the generated curve (Fig. 3), approximated the likely region of recombination.

Figure 3. Regional Logarithmic-scaled parameter of changepoint likelihood for the same genome as studied within Fig. 2 surveying +/- 30 proteins determines likelihood of recombination: this can be utilized to analytically predict changepoints (orange dots) within the similarity matrix.

Evolution and Recombinant Ancestry

Historical Recombination Tracer

S. aureus GCA_000027045 – Strain Type 239



Within genomic datasets with known phylogenetic metadata, recombination analysis can be extended into hypothesizing the likely evolutionary history of recombination. In Fig. 4, utilization of known phylogenetic classifications within the *S. aureus* database, genomic recombination can be traced back to genome subgroupings (eg., clonal complexes) of origin.

Figure 4. Evolutionarily analysis recapitulated previous studies findings that recombinant regions within this ST239 genome likely originate from Clonal Complex 8 (CC8; a genomic subgroup) and CC30. Further, the genomic region within protein number 2100-2300 likely corresponds to genomic information originating from outside of the canonical *S. aureus* gene set.

RESULTS (Contd.)

Simulated Genome Analysis

We studied the applicability of Redcarpet with simulated genomes and databases. Redcarpet was able to detect an artificially-introduced recombinant region within 10^5 base-pair genomes.

Redcarpet analysis for two simulated genomes

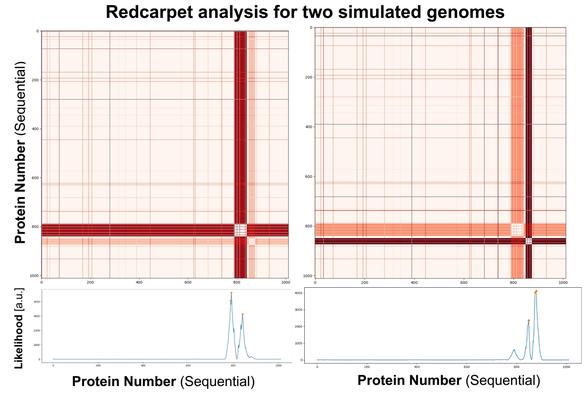


Figure 5. Similarity Matrix (top) and likely changepoints (bottom) for two simulate genomes, each with a unique major recombinant event and an additional less significant one. Changepoint analysis was able to recognize the primary region of recombination. Note the smaller peak in (D) which corresponds with a less significantt yet possible recombinant region. Changepoint specification may allow the user to select the approximate region within which they wish to find the most likely single recombination, at which point either recombinant event may be detectable.

CONCLUSIONS

The comparative analysis of identical protein matches within a large database of genomes offers an opportunity to rapidly assess the comparative recombinatory history of any given genome.

FUTURE DIRECTIONS

It is our goal to develop this program into a freely accessible publicly-available tool. This will require streamlining and automating some of the steps (most notably changepoint analysis) within the procedure.

REFERENCES

1- Moustafa AM & Planet PJ. WhatsGNU: a tool for identifying proteomic novelty. Genom Biol 21, 58 (2020).

2- Ondov et al. Mash: fast genome and metagenome distance estimation using MinHash. Genom Biol 17, 132 (2016).