Approximating Haploid and Diploid Inferred Paths in Genome Graphs

William T. Doan, Daniel J. Gibney

THE UNIVERSITY OF TEXAS AT DALLAS
Undergraduate Research Apprenticeship Program

Introduction

Genome graphs encode collections of reference DNA sequences in graph form. This compact representation reduces space and unifies computation across reference collections. A central problem is: given reads from a newly sequenced DNA strand, how can we use the genome graph to reconstruct the sampled sequence?

This is formalized as the "path inference problem": given a genome graph with known reference paths and a set of reads from a novel sequence, identify the 'best' path through the graph.

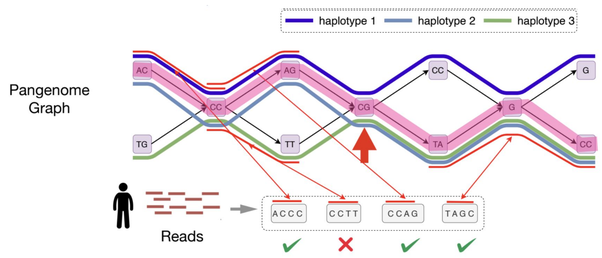

Fig. 1. The Problem Formulation. Credit: *RECOMB 2025*.

The notion of 'best' incorporates prior knowledge of reference haplotypes and the likelihood of recombinations between them. This problem was addressed in earlier work (RECOMB 2025), which provided a mathematical formulation and an integer linear programming framework.

The diploid variant of the path inference problem extends this by asking for a pair of paths that best explain a set of reads sampled from a diploid individual. Some reads are shared between the two sequences (homozygous), while others are unique to one (heterozygous). This reflects the diploid nature of human genomes.

Our approach combines well-studied sketching algorithms and dynamic programming in a novel way. We develop the first known polynomial-time approximation algorithms for both path inference and diploid path inference. Our results include a randomized polynomial-time approximation scheme and a deterministic algorithm with approximation ratio governed by a parameter that remains small for low-coverage read sets.

Problem Formulation

Given a set of reads $R \subseteq \Sigma^+$, where Σ is an alphabet and each read $r \in R$ is a string over Σ, we define the set of unique kmers S as follows: for a fixed integer $k \geq 1$, a kmer is a substring of length k, and the set of kmers extracted from a read $r \in R$ is $\text{Kmers}(r) \leftarrow \{ r[i..i+k-1] \mid 1 \leq i \leq |r| - k + 1 \}$. The global set of kmers is then $S \leftarrow \bigcup_{r \in R} \text{Kmers}(r)$, retaining only unique kmers. We partition S into two disjoint subsets: $S \leftarrow S_{homo} \cup S_{hetero}$ with $S_{homo} \cap S_{hetero} = \varnothing$.

Problem 1 (Diploid Inference Paths Problem with Fixed Recombinations)**.**

Input: A labeled haplotype-aware directed acyclic graph (DAG) $G = (V, E, \sigma, H)$ is given, along with a set of strings S, a non negative integer c called the recombination penalty, and $\mathcal{R}$, which represents the maximum allowed recombinations.

$$Cost(\mathcal{P}_1, \mathcal{P}_2) = \max_{(r(\mathcal{P}_1)+r(\mathcal{P}_2)) \leq \mathcal{R}} \sum_{x \in \mathcal{S}_{hom}} (\chi(x, \sigma(\mathcal{P}_1)) \wedge \chi(x, \sigma(\mathcal{P}_2))) + \sum_{x \in \mathcal{S}_{het}} (\chi(x, \sigma(\mathcal{P}_1)) \oplus \chi(x, \sigma(\mathcal{P}_2))).$$

where $r(\mathcal{P}_i)$ is the number of recombinations of $\mathcal{P}_i$, $\chi(x, \sigma(\mathcal{P}_i)) = 1$ if string x occurs as a substring of $\sigma(\mathcal{P}_i)$, and 0 otherwise. Here, $\sigma(\mathcal{P}_i)$ for $i \in \{1, 2\}$ denotes the string spelled by path $\mathcal{P}_i$.

We approach this NP-Hard problem using approximation algorithms. Our goal will be to output a $(1 - \varepsilon)$ approximation in polynomial time with a probability of at least $(1 - \delta)$. Our algorithm uses HyperLogLog sketches and dynamic programming.

To do that, let SKETCHES(v) denotes a set of sketches and corresponding minimum path weights associated with vertex $v \in V$. We assume we can efficiently check if a given sketch is in SKETCHES(v), as well as add a new sketch/weight pair into SKETCHES(v). We initialize SKETCHES(v) for every source vertex v. Using a hash table, wherein each key is a sketch, we are able to identify that path $\mathcal{P}$ of total weight of at most $\mathcal{R}$ which contains approximately the most unique kmers.

Analysis and Conclusion

Our Polynomial-time Randomized Approximation Scheme (PRAS) is described as:—

— Input: an instance I, an error tolerance $\varepsilon > 0$, and a confidence interval of $\delta > 0$

— Output: a solution S such that

$$\Pr[\text{Score}(S) \geq (1 - \varepsilon) \cdot \text{OPT}] \geq 1 - \delta$$

Suppose:

$|S|$: number of strings,

$|E|$: number of edges in the graph,

$\mathcal{R}$: recombination budget.

We run a dynamic programming-style algorithm over all paths of cost $\leq \mathcal{R}$, storing sketches at each step. In the diploid case, the algorithm considers pairs of such paths, which leads to a per-trial time complexity of

$$|S|^{3 + O(1/\varepsilon)} \cdot |E|$$

per trial for $\log(1/\delta)$ trials.

Theorem 1.
There exists a randomized algorithm that returns a pair of paths $(\mathcal{P}_1, \mathcal{P}_2)$ in the expanded graph G_E such that the number of strings in S that are hit by either $\mathcal{P}_1$ or $\mathcal{P}_2$ is at least $(1 - \varepsilon) \cdot \text{OPT}$ with probability at least $1 - \delta$, where OPT is the maximum number of strings hit by any pair of paths under the same recombination constraint and the algorithm runs in time

$$|S|^{3 + O(1/\varepsilon)} \cdot |E| \cdot \log(1/\delta).$$

Administrivia

William T. Doan
doan@utdallas.edu
https://wtd-1.github.io/web/

Acknowledgments
I would like to thank Dr. Daniel Gibney for his patience and guidance during this project. This work was brought about in part by the generous support of the Undergraduate Research Apprenticeship Program at the University of Texas at Dallas.