Finding all Significant Closed Connected Subgraphs at Scale

HECTOR ROUX DE BÉZIEUX[1,2], FANNY PERRAUDEAU[2], ARNAUD MARY[3], SANDRINE DUDOIT[1], LAURENT JACOB[3]
[1]UC Berkeley, [2]Pendulum Therapeutics, Inc., [3]Université de Lyon

BIOLOGICAL MOTIVATION

n samples with binary phenotypes 0/1, (resistance to antibiotic), and a set of genomics sequences for each sample. Testing for association between all k-mers and the phenotype is redundant and hard to interpret.

Figure 1: Example of setting: one strain is sensitive to antibiotic, one is not. Genetic sequences are sequenced for each strain and lead to the table of presence-absence for each k-mer for the two strains, with $k = 4$.

Compacted De Bruijn Graphs (DBGs) allow for non-redundant compressed format without loss of information.

Figure 2: Compaction of the k-mers into a De Bruijn Graph.

However, testing only individual nodes makes results hard to interpret since genetic features such as genes can be represented by several nodes.

	TTCG TCGTA	TCGCTCG	TCGATCG
	1	1	0
	1	0	1

Figure 3: Table of presence-absence for each unitig for the two strains. No information is lost compared to the k-mer table but this is described with fewer sequences.

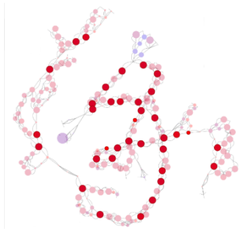

Figure 4: Each dark red node is significantly associated with the phenotype.

A typical bacterial genome graph contains millions of nodes, the subgraph above is just a small part representing a gene. That sequence is not linear because of small mutations along that gene between samples.

METHODS

- **Consider all closed connected subgraphs (CCS) of the DBG.**
- **Tarone's idea of testability [1]**: for a discrete distribution, the smallest possible p-value $p^\star$ can be strictly bigger than 0. If it is higher than the rejection threshold, the hypothesis is not-testable and can be discarded, decreasing the number of hypotheses being tested. It is therefore possible to control the Family-Wise Error Rate at the same nominal level while strictly increasing the power.
- **Enumerate all CCS by building an appropriate tree structure rooted on $\emptyset$.** We define a tree structure from a `Children` function. Instead of enumerating all connected subgraphs and discarding the non-closed ones, we directly enumerate the CCS using a double alphabetical order on samples and nodes. This leads to a faster enumeration.
- **Building a tree structure that can be pruned using testability** The `Children` needs to verify the following property: for all CCS $\mathcal{S}, \mathcal{S}', \mathcal{S}' \in \texttt{Children}(\mathcal{S}) \implies p^\star(\mathcal{S}) \leq p^\star(\mathcal{S}')$.

ALGORITHM

Algorithm 1 `CALDERA`: List all significant closed connected subgraphs [2]

▷ Find all testable CCS
procedure ENUM($\mathcal{S}$, Testables, k_0)
  for $\mathcal{S}' \in \texttt{Children}(\mathcal{S})$ **do**
    if $p^\star(\mathcal{S}') \leq \alpha/k_0$ **then**
      Add $\mathcal{S}'$ to Testables
      $k_0 \leftarrow k_0 + 1$
      Update Testables given new α/k_0
      Enum($\mathcal{S}'$, Testables, k_0)
    end if
  end for
  return Testables
end procedure
Testables $\leftarrow$ Enum($\emptyset, \emptyset, 1$)
▷ Actually test them
Sols $\leftarrow \emptyset$
for $\mathcal{S} \in$ Testables **do**
  If $p(\mathcal{S}) < \alpha/k_0$, add $\mathcal{S}$ to Sols
end for

TOY EXAMPLE

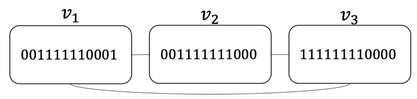

Figure 5: Toy example with 3 nodes and 12 samples. Each node has a vector of presence-absence of samples.

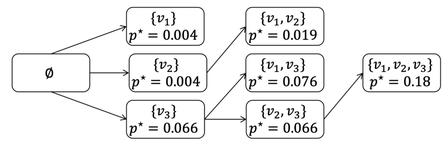

Figure 6: `CALDERA` defines a reduction on the CCS that can be inverted to explore all CCS starting from $\emptyset$.

Breadth-First Search. Since $\{v_3\}$ is not testable once we finish exploring the first stage (step 3), we can prune the branch: we do not explore its children $\{v_1, v_3\}$, $\{v_2, v_3\}$ and $\{v_1, v_2, v_3\}$.

Subgraph	Testables	k_0	α/k_0
$\{v_3\}$	$\{\{v_3\}\}$	1	.15
$\{v_2\}$	$\{\{v_3\}, \{v_2\}\}$	2	.075
$\{v_1\}$	$\{\{v_2\}, \{v_1\}\}$	3	.05
$\{v_1, v_2\}$	$\{\{v_2\}, \{v_1\}, \{v_1, v_2\}\}$	3	.05

DISCUSSION

The method scales to bacterial genome samples (2 million nodes DBG) but not to metagenome samples (100 million nodes DBG). Pre-processing of the data, including filtering of low-frequencies k-mers might help.

RESULTS ON SIMULATIONS

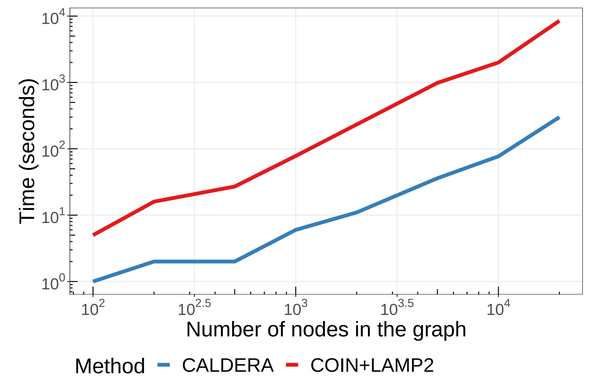

Figure 7: Runtimes for increasing graph sizes against state-of-the-art:`COIN` [3] and `LAMP2` [4]

RESULTS ON REAL DATA

$n = 280$ *Pseudomonas Aeruginosa genomes* from Jaillard et al. [5], along with their amikacin resistance phenotype. `CALDERA` runs in ~ 5 hours while `COIN+LAMP2` is only at 10% of the exploration after 9 days.
The top two hits match the only genes genes linked to resistance phenotype for those strains.

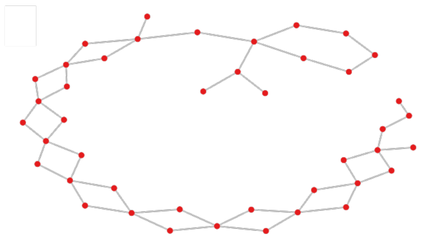

Figure 8: Subgraph with lowest p-value, matches to the AAC(6′) gene

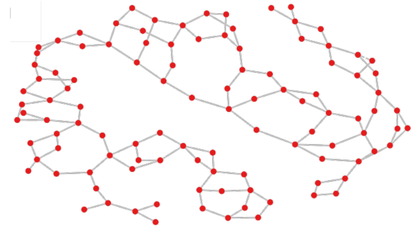

Figure 9: Subgraph with second lowest p-value, matches to the pHS87b plasmid

MAIN REFERENCES

[1] R. E. Tarone. A Modified Bonferroni Method for Discrete Data. *Biometrics*, 1990. doi: 10.2307/2531456.
[2] Felipe Llinares-López, Dominik G. Grimm, Dean A. Bodenham, Udo Gieraths, Mahito Sugiyama, Beth Rowan, and Karsten Borgwardt. Genome-wide detection of intervals of genetic heterogeneity associated with complex traits. *Bioinformatics*, 2015. doi: 10.1093/bioinformatics/btv263.
[3] Jun Sese, Aika Terada, Yuki Saito, and Koji Tsuda. Statistically significant subgraphs for genome-wide association study. *SDM*, 47:1–7, 2014.
[4] Shin Ichi Minato, Takeaki Uno, Koji Tsuda, Aika Terada, and Jun Sese. A fast method of statistical assessment for combinatorial hypotheses based on frequent itemset enumeration. In *Lecture Notes in Computer Science*, 2014. doi: 10.1007/978-3-662-44851-9_27.
[5] Magali Jaillard, Leandro Lima, Maud Tournoud, Pierre Mahé, Alex van Belkum, Vincent Lacroix, and Laurent Jacob. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS genetics*, 2018. doi: 10.1371/journal.pgen.1007758.