Finding all Significant Closed Connected Subgraphs at Scale
HECTOR ROUX DE BÉZIEUX1,2, FANNY PERRAudeau2, ARNAUD MARY3, SANDRINE DUDOIT1, LAURENT JACOB3
1UC Berkeley, 2Pendulum Therapeutics, Inc., 3Université de Lyon

BIOLICAL 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.

Consider all closed connected subgraphs (CCS) of the DBG.

Taron’s idea of testability [1]: for a discrete distribution, the smallest possible p-value \( p^* \) 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 1-frontal 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 follow the above procedure: for all CCS \( S, S', S'' \in \text{Children}(S) \Rightarrow p^*(S) \leq p^*(S') \).

METHODS

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.

RESULTS ON REAL DATA

RESULTS ON SIMULATIONS

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.

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.

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 \( k = 4 \).

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.

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.

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

MINI REFERENCES


TOY EXAMPLE

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

1: if Find all testable CCS
2: procedure Enum(S, Testables, k0)
3: end for
4: if \( p^*(S') \leq \alpha/k_0 \) then
5: Add \( S' \) to Testables
6: \( k_0 \leftarrow k_0 + 1 \)
7: Update Testables given new \( \alpha/k_0 \)
8: Enum(S, Testables, k0)
9: end if
10: end for
11: return Testables
12: end procedure
13: Testables ← Enum(\( \emptyset \), \( \emptyset \), 1)
14: if Actually test them
15: Sols ← \( \emptyset \)
16: for \( S \in \text{Testables} \) do
17: if \( p(S) < \alpha/k_0 \), add \( S \) to Sols
18: end for

SUBGRAPH WITH SECOND LOWEST P-VALUE, MATCHES TO THE AAC(6') GENE

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.

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

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

Figure 9: Subgraph with second lowest p-value, matches to the pHS67 plasmid