Supplementary Material
Sparse Binary Relation Representations for Genome Graph Annotation

Mikhail Karasikov[1,2,3], Harun Mustafa[1,2,3], Amir Joudaki[1,2], Sara Javadzadeh No[1], Gunnar Rätsch[1,2,3], and André Kahles[1,2,3]

[1] Department of Computer Science, ETH Zurich, Zurich, Switzerland
[2] University Hospital Zurich, Biomedical Informatics Research, Zurich, Switzerland
[3] SIB Swiss Institute of Bioinformatics, Zurich, Switzerland {raetsch,andre.kahles}@inf.ethz.ch

1 Binary relation matrix simulation

To benchmark our compression techniques systematically, we generated three series of random binary matrices satisfying different properties. Given fixed matrix dimensions $n \times m$, an expected column density d, and a uniqueness factor u, we define our generation schemes as follows:

1. **Random:** generate m random columns of length n with expected density d
2. **Uniform rows:** generate m random columns of length $\frac{n}{u}$, then duplicate each row u times
3. **Uniform columns:** generate $\frac{m}{u}$ columns of length n, then duplicate each column u times

For each generated column, its indices are iterated through linearly and the values of the indices are set by drawing observations from a random variable $X \sim \text{Bernoulli}(d)$. For all experiments, values of $n = 1,000,000$, $u = 5$, and $d = 0.01$ were used. The values $m \in \{500, 1000, 3000\}$ were used.

1.1 Sizes of compressed representations

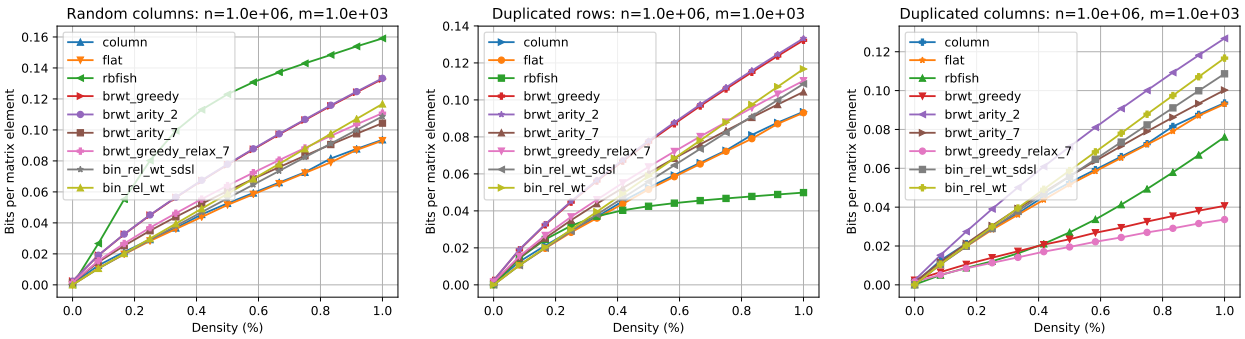

Fig. 1. Size of the representation of $A \in \{0,1\}^{10^6 \times 1000}$ with densities $d < 0.01$ using different approaches: a) Random columns; b) Duplicated rows; c) Duplicated columns

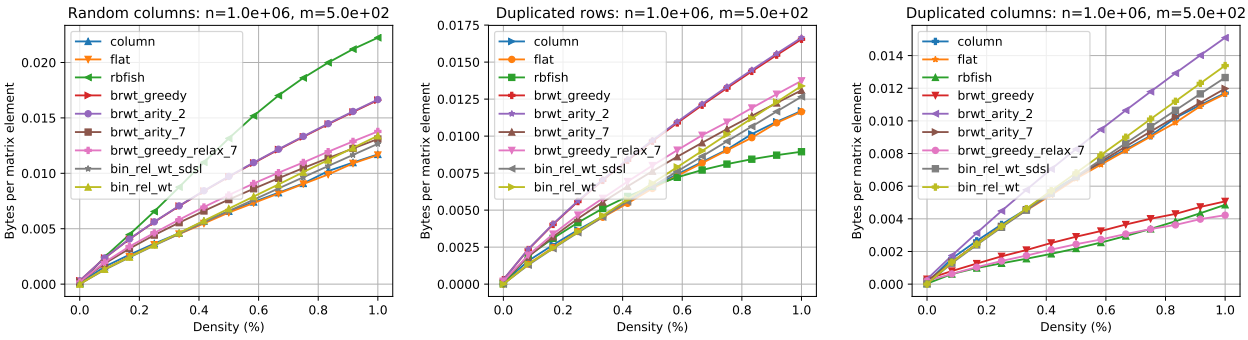

Fig. 2. Size of the representation of $A \in \{0,1\}^{10^6 \times 500}$ with densities $d < 0.01$ using different approaches: a) Random columns; b) Duplicated rows; c) Duplicated columns

2 Subsampling lemma

Proof. Suppose for $i \in \{1, ..., n\}$ we introduce random variable X_i as follows:

$$X_i = \begin{cases} 1 & \text{if } i \in \mathcal{S}, \\ 0 & \text{otherwise.} \end{cases}$$

In other words X_i is 1 if i is subsampled and 0 otherwise. Besides, for $V \subseteq \{1, ..., n\}$ we define $X_V = \sum_{i \in V} X_i$, which basically counts how many elements of V are subsampled. Based on these definitions can derive

$$X_{o_i \cup o_j} = |(o_i \cup o_j) \cap [n]_p| = |\tilde{o}_i \cup \tilde{o}_j| = p\,\tilde{u}_{i,j},$$

$$\mathbb{E}[X_{o_i \cup o_j}] = p \cdot |o_i \cup o_j| = p\,u_{i,j}.$$

If we denote the mean by $\mu_{i,j} = \mathbb{E}[X_{o_i \cup o_j}]$ we have

$$\begin{aligned} \Pr(|\tilde{u}_{i,j} - u_{i,j}| \geq \epsilon u_{i,j}) &= \Pr\left(|\frac{1}{p} X_{o_i \cup o_j} - \frac{1}{p}\mu_{i,j}| \geq \frac{1}{p}\epsilon\mu_{i,j}\right) \\ &= \Pr\left(|X_{o_i \cup o_j} - \mu_{i,j}| \geq \epsilon\mu_{i,j}\right). \end{aligned}$$

Using Chernoff bound we have

$$\begin{aligned} \Pr\left(|X_{o_i \cup o_j} - \mu_{i,j}| \geq \epsilon\mu_{i,j}\right) &\leq 2e^{-\epsilon^2 \mu_{i,j}/3} && \triangleright \text{ Chernoff} \\ &= 2e^{-\epsilon^2 |o_i \cup o_j| \frac{12 \log m}{d\epsilon^2} \frac{1}{3}} && \triangleright \text{ plugging } p \\ &\leq 2m^{-12/3} && \triangleright\ |o_i \cup o_j| \geq |o_i| \geq d \end{aligned}$$

and, therefore,

$$\Pr(|\tilde{u}_{i,j} - u_{i,j}| \geq \epsilon u_{i,j}) \leq 2m^{-4}. \tag{1}$$

Now back to the joint probability we can write

$$\begin{aligned} \Pr\Big(\bigvee_{i,j \in [m]} |\tilde{u}_{i,j} - u_{i,j}| \geq \epsilon u_{i,j}\Big) &\leq \sum_{i<j} \Pr(|\tilde{u}_{i,j} - u_{i,j}| \geq \epsilon u_{i,j}) && \triangleright \text{ union bound} \\ &< \frac{m^2}{2} \cdot 2m^{-4} && \triangleright \text{ by inequality 1} \\ &= m^{-2}. \end{aligned}$$

Finally, taking the compliment of the probability gives us the claimed bound:

$$\Pr\Big(\bigwedge_{i,j \in [m]} |\tilde{u}_{i,j} - u_{i,j}| \leq \epsilon u_{i,j}\Big) > 1 - m^{-2}.$$