Algorithms in Computational Biology and Sequence Analysis

Module 1: Introduction

Demo - Rank Data Structure

Dr. Chirag Jain
Department of Computational and Data Sciences
Indian Institute of Science
Web: https://at-cg.github.io/

Teaching Assistant:

Daanish Mahajan
Interdisciplinary Mathematical Sciences
Indian Institute of Science

Bitvector rank operation

B

0
1
0
1
1
0
1
1
0

- **Input**: A bitvector is an array $B[1\ldots n]$ of boolean values, i.e., $B[i] \in \{0,1\}$

- **Objective**: Design a space-efficient data structure to support queries of the following type

  - $rank_1(B)[i] = \left| \{i' \,|\, 1 \leq i' \leq i, B[i'] = 1\} \right|$

  - $rank_0(B)[i] = \left| \{i' \,|\, 1 \leq i' \leq i, B[i'] = 0\} \right|$

  - NOTE: $rank_0(B)$ is directly known from $rank_1(B)$

- An important building block of space-efficient data structures for general data as well as genomic data indexing

Naive Algorithm

B

0	1	1	0	1	0	1	1	0	0	0	1	1	1	0	1	0	1	0	1	0	1	1	0

$rank_1(B)[i]$

0	1	2	2	3	3	4	5	5	5	5	6	7	8	8	9	9	10	10	11	11	12	13	13

Succinct Algorithm (4-Russian technique)

B

0	1	1	0	1	0	1	1	0	0	0	1	1	1	0	1	0	1	0	1	0	1	1	0

$l = ((\log n)/2)^2$

first | 0 | | 9 |

$k = (\log n)/2$

second | 0 | | 2 | | 5 | | 6 | | 0 | | 2 | | 4 |

third $[1\dots2^{k-1}][1\dots k-1]$

i

$$rank_1(B)[i] = first[i/l] + second[i/k] + third[c_i][(i \bmod k)]$$

"/" indicates integer division