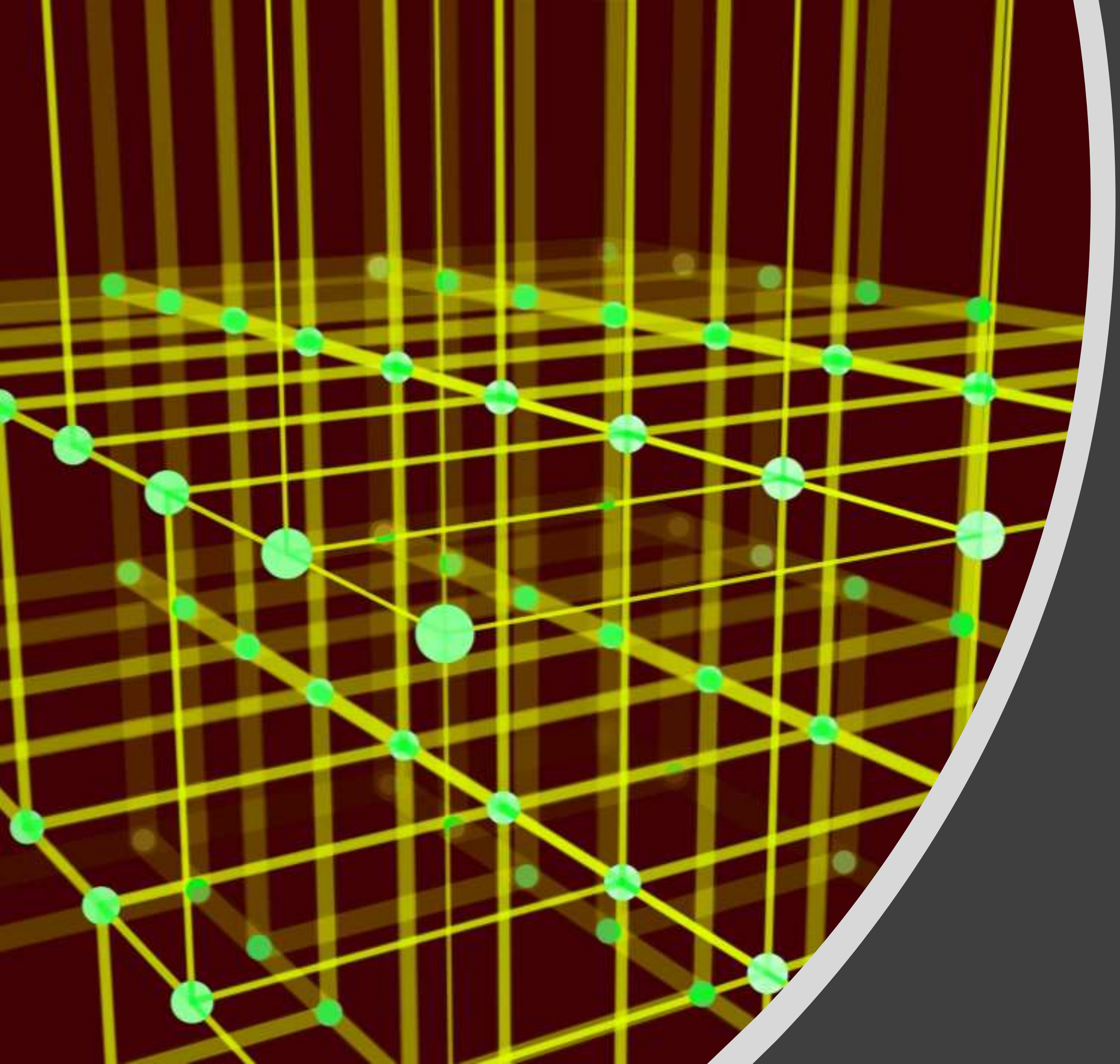


Distributed Framework for Gene Finding using Open-MPI

Shenghua Chen

University of Chicago

MPCS 56430 Introduction to Scientific Computing

Introduction

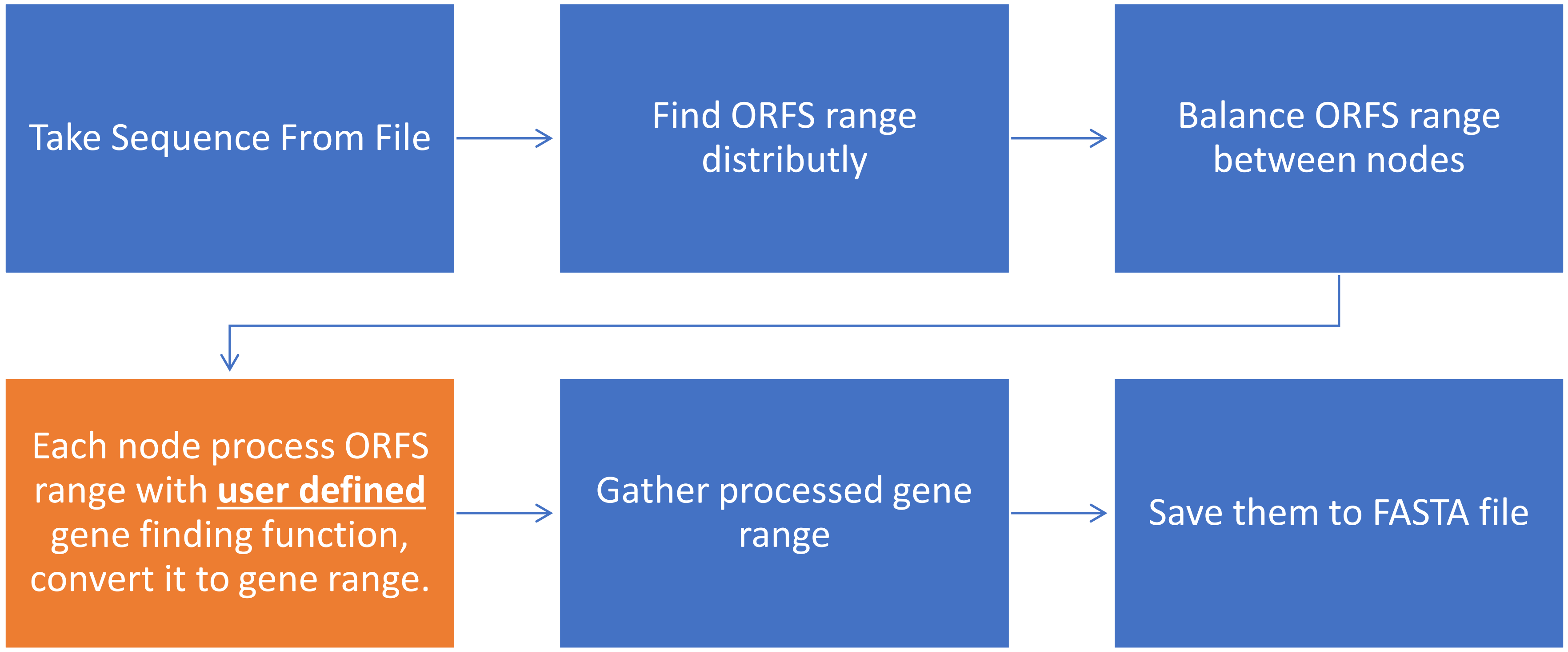
Project Objectives

- What is gene?
  - Genes are parts of DNA whose job is to make specific proteins that play a key role in the structure and function of the body.
- What is gene finding?
  - In computational biology, gene prediction or gene finding refers to the process of identifying the regions of genomic DNA that encode genes.
- We need a framework that can help us deploy gene finding job as fast as possible.

Background

- There are few gene finding software exist:
  - FINDER (https://github.com/sagnikbanerjee15/Finder)
  - GeneParser (http://stormo.wustl.edu/src/GenParser/)
  - mGene (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2775605/)
  - ……
- Those software has limitation:
  - Single algorithm
    - More Flexible
  - Single node
    - Can’t run it on cluster

My Solution



Take Sequence From File

Find ORFS range distributly

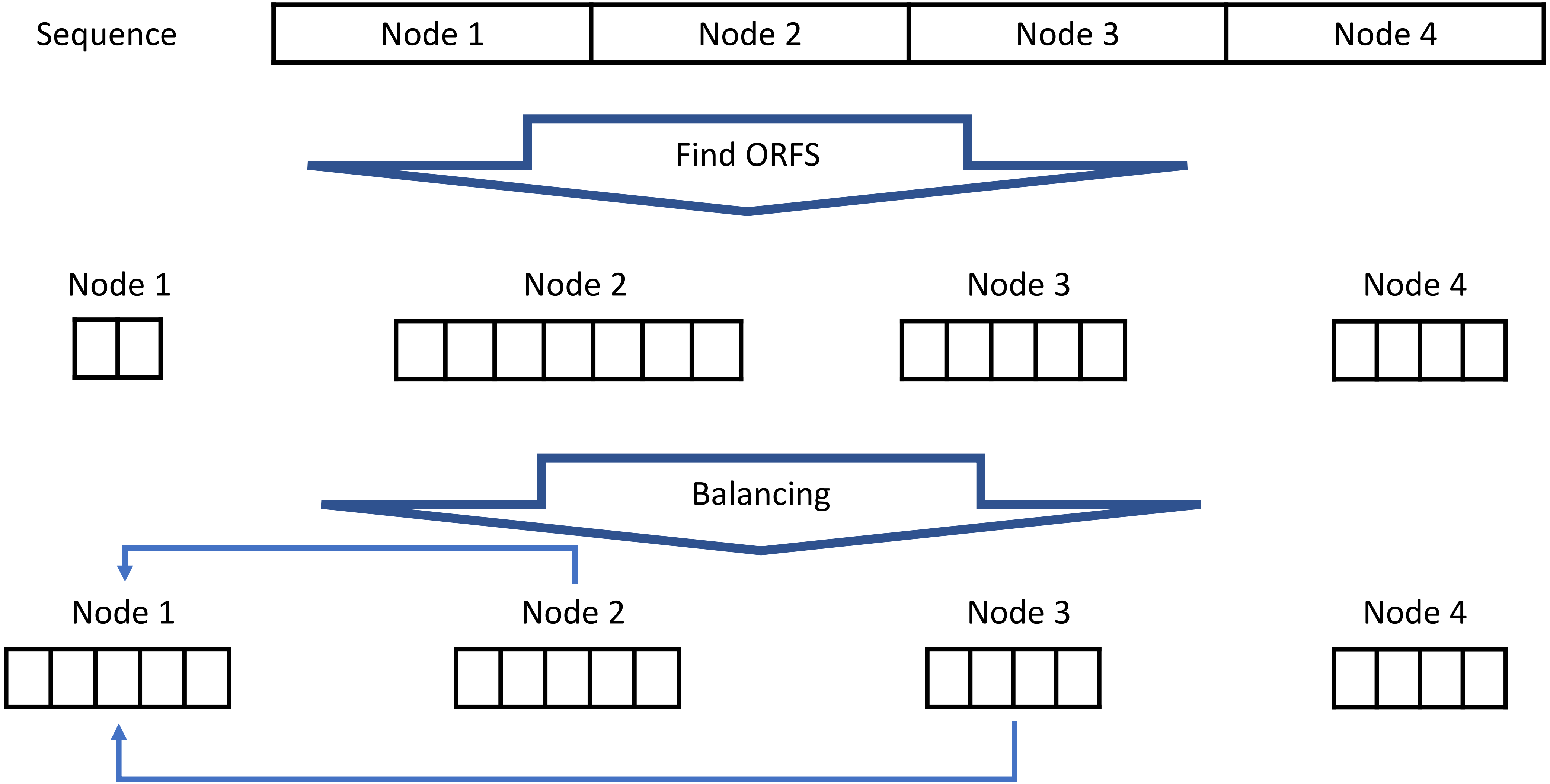
Balance ORFS range between nodes

Each node process ORFS range with **user defined** gene finding function, convert it to gene range.

Gather processed gene range

Save them to FASTA file

My Solution


Sequence
Node 1
Node 2
Node 3
Node 4
Find ORFS
Node 1
Node 2
Node 3
Node 4
Balancing
Node 1
Node 2
Node 3
Node 4

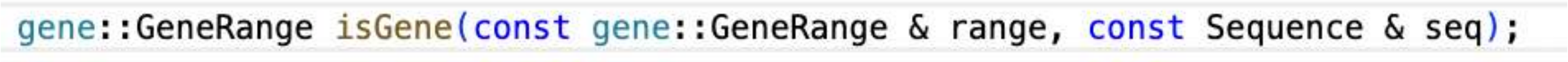
User Defined Function

Take:

- ORF
  - *GeneRange*[Start,End]
- Sequence

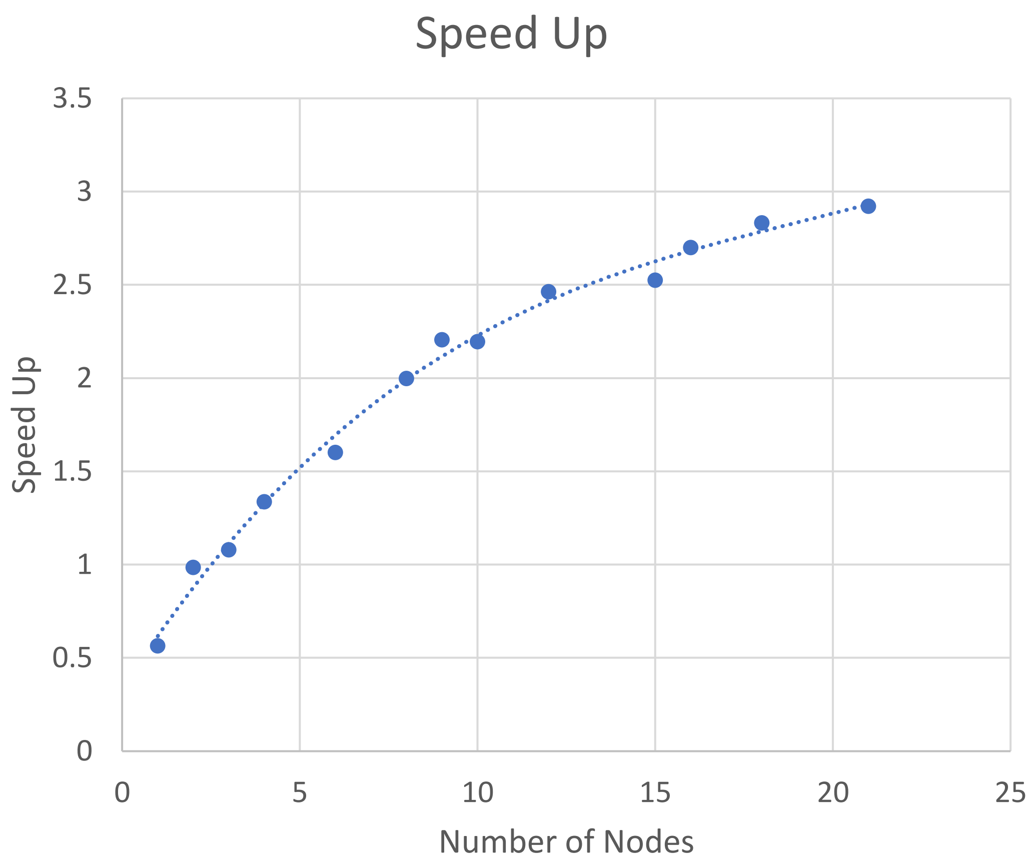
Return:

- Not Gene Signal
  - An invalid *GeneRange*
- Gene
  - *GeneRange*[AdjustedStart,AdjustedEnd]

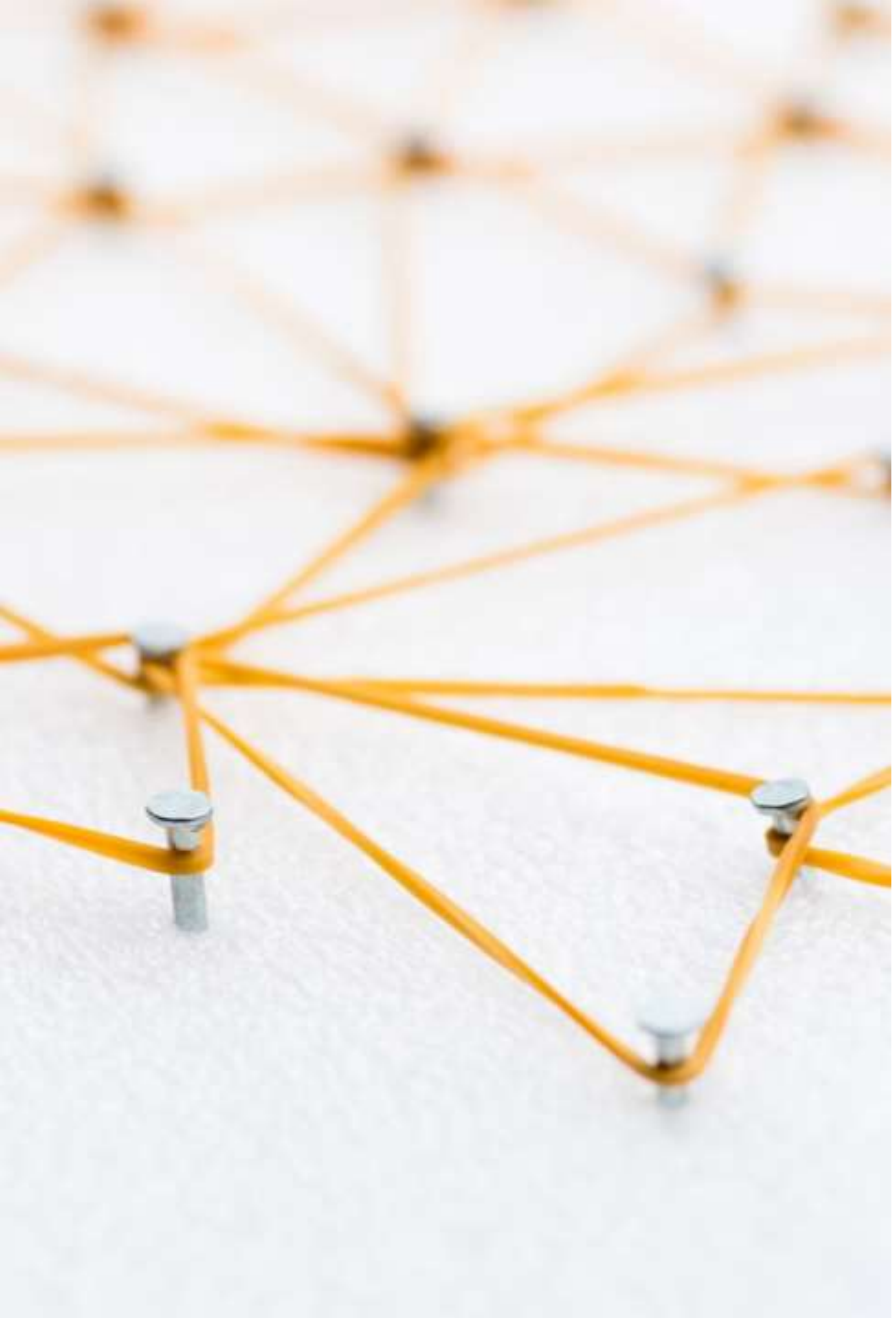


```
gene::GeneRange isGene(const gene::GeneRange & range, const Sequence & seq);
```

Summary - Performance



- $Speedup = \frac{T_{single\_node}}{T_{multi\_node}}$
- Speed up is based on complexity of user defined gene finding function.
  - This step is fully parallizble
- With highly complex gene finding function speed up factor can over 4 when there is 5 nodes.



Future Work

- Edge Computing
  - Rely on stable connection between nodes
- Heterogeneous computing
  - Balancing are based on assumption: Every nodes has similar computational power.

Thank you

- Code: https://github.com/kaseidis/MPI_Simple_Gene_Finder