The Emerging Importance of Massively Parallel Genomic Sequence Search

http://www.mpiblast.org/

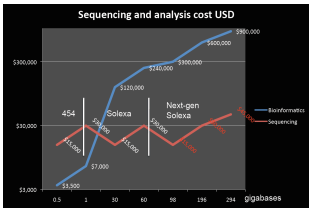
Heshan Lin and Wu-chun Feng

Department of Computer Science, Virginia Tech

http://synergy.cs.vt.edu/

Motivation

- **Sequence Search: Fundamental Tool in Computational Biology**
  - Infer functional and evolutionary relationships between genetic sequences.
  - Help identify members of gene families and discover new drugs
- **Astronomical Growth of Sequence Databases**
  - Rapidly *increasing* speed in sequencing and *decreasing* cost
  - Consequences
    - ✓ Outstripping our computational capability to analyze the data
    - ✓ Creating new opportunities in metagenomics and personalized genomics



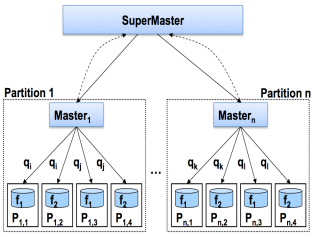
Source: F. Meyer et al.

- **Need: Petascale Supercomputing**
- **Example Application: Bioenergy and Biofuel Research**

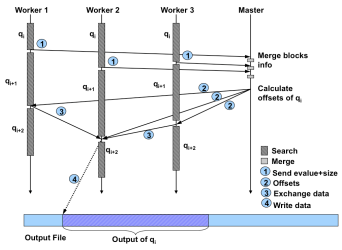
Sequence Search → Discovering Missing Genes → Converting Biomass Sugars into Energy

Approach

- **mpiBLAST: Massively Parallel Genomic Sequence Search**
- **Petascale Design**
  - A scalable architecture based on hierarchical scheduling
  - Fine-grained, low-overhead load balancing
  - Asynchronous, two-phase parallel I/O – collectively write output *without* synchronization
- **Open-Source Parallelization of NCBI BLAST**
  - Available at www.mpiblast.org



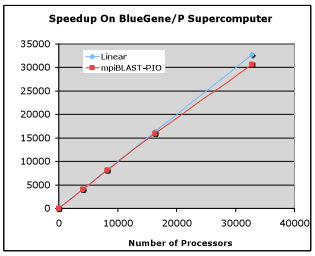
Hierarchical Scheduling



Asynchronous, Two-Phase I/O

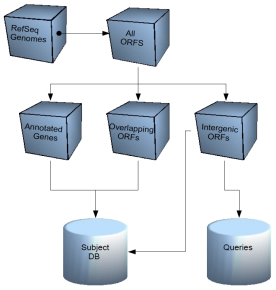
Performance Results

- **mpiBLAST Performance**
  - *Platform:* IBM Blue Gene/P
  - *Reference Database (DB):* 16-million microbial sequences
  - *Query:* 0.25 million randomly sampled sequences from DB
  - *Results:* 93% parallel efficiency across 32,768 cores

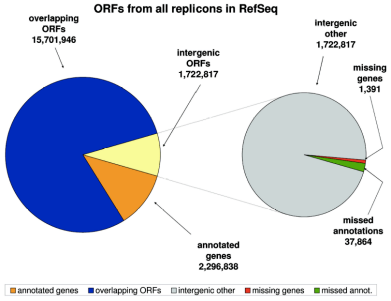


Discovering Missing Genes

- **Goal: Identify Genes Missed by Traditional Gene Detection**
- **Methodology: Compare Raw Genomes Against Each Other**
- **Computational Requirements**
  - Effectively, an all-to-all BLAST search of over 16-million sequences
- **Results**
  - Computational
    - ✓ Completed in 12 hours on 32,768 cores of Blue Gene/P
  - Biological
    - ✓ 1,391 potential missing genes
    - ✓ 37,864 potential missed annotations



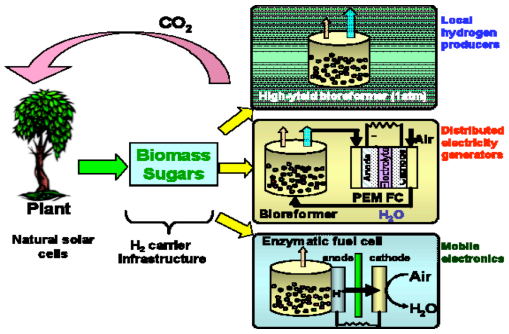
Constructing Search Data



Gene-Finding Results

Application of Discovering Missing Genes

- **Accelerated Production of Bioenergy and Biofuels**
  - Conversion of biomass sugar into hydrogen or electrical energy
  - Exhaustive data mining to identify the thermostable enzymes from the mega-genome database that will efficiently and cheaply do the biomass sugar to energy conversion, as denoted by the ⇨ below.



Collaborators

- Xiaosong Ma, NCSU and ORNL
- Nagiza Samatova, NCSU and ORNL
- Joao Setubal, Virginia Tech
- Carlos Sosa, IBM

VirginiaTech *Invent the Future*

printed by MegaPrint Inc. www.postersession.com