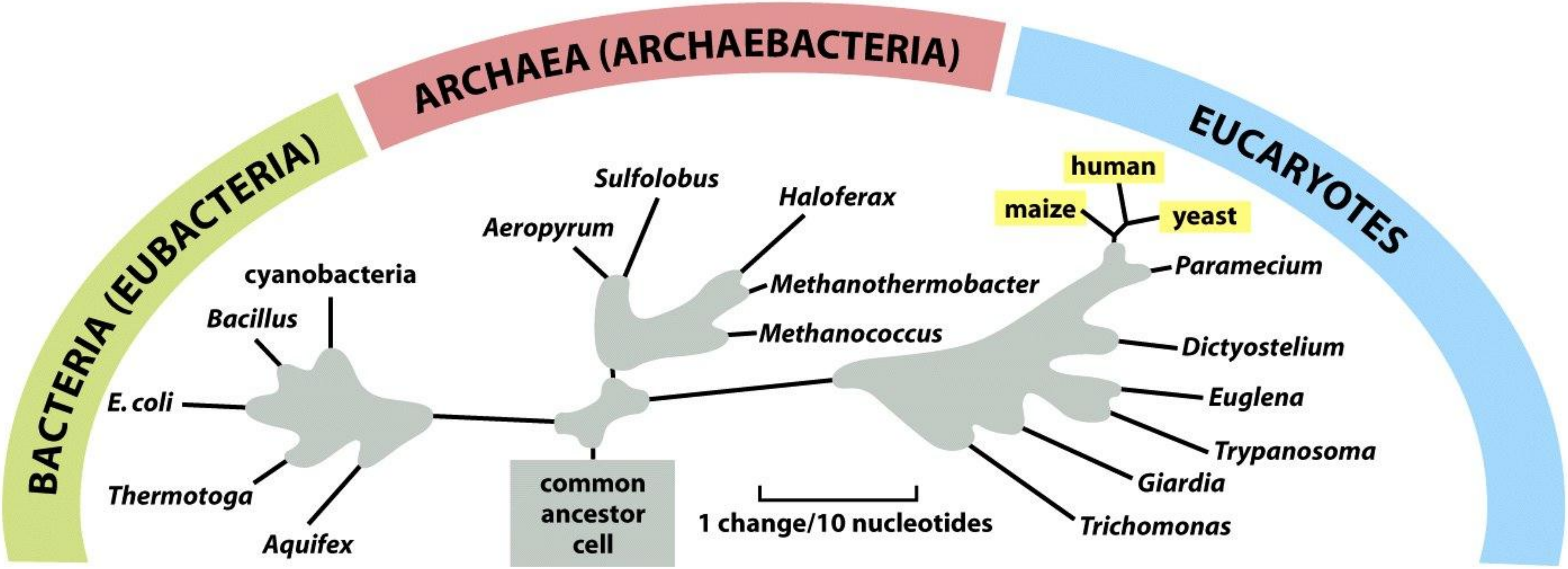
VERY LARGE SCALE OPERON PREDICTIONS VIA COMPARATIVE GENOMICS

USING MICROSOFT AZURE

Ehsan Tabari
University of NC Charlotte

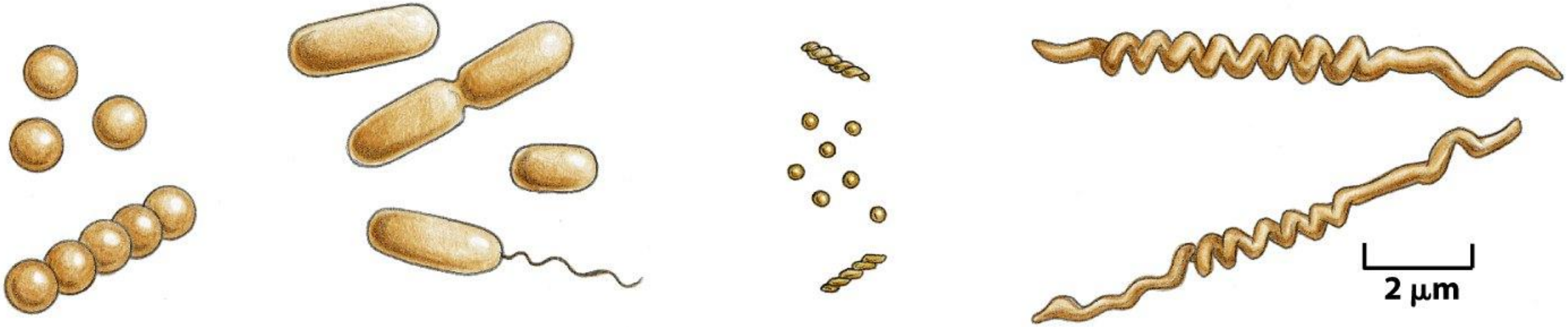
Bioinformatics

Life domains



Courtesy of Molecular Biology of the Cell 5/e (© Garland Science 2008)

Prokaryotes

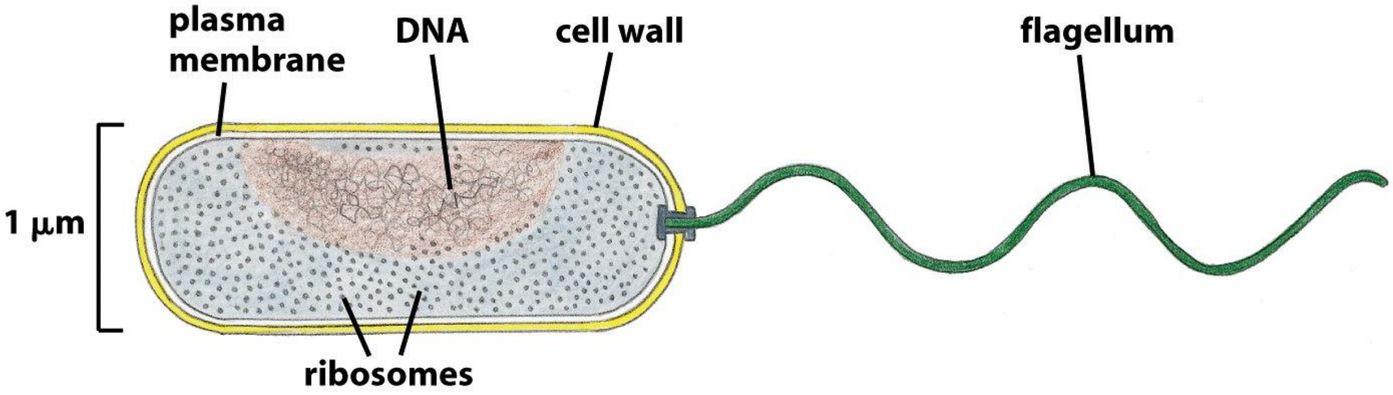


spherical cells
e.g., ***Streptococcus***

rod-shaped cells
e.g., ***Escherichia coli, Vibrio cholerae***

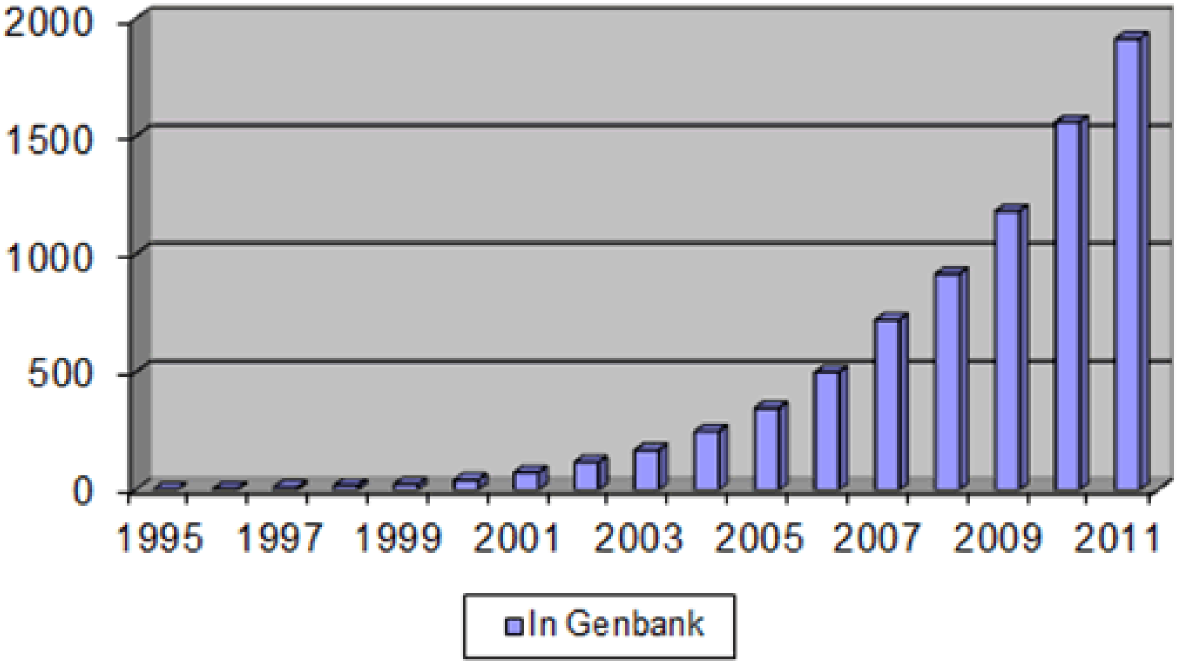
the smallest cells
e.g., ***Mycoplasma, Spiroplasma***

spiral cells
e.g., ***Treponema pallidum***



Courtesy of Molecular Biology of the Cell 5/e (© Garland Science 2008)

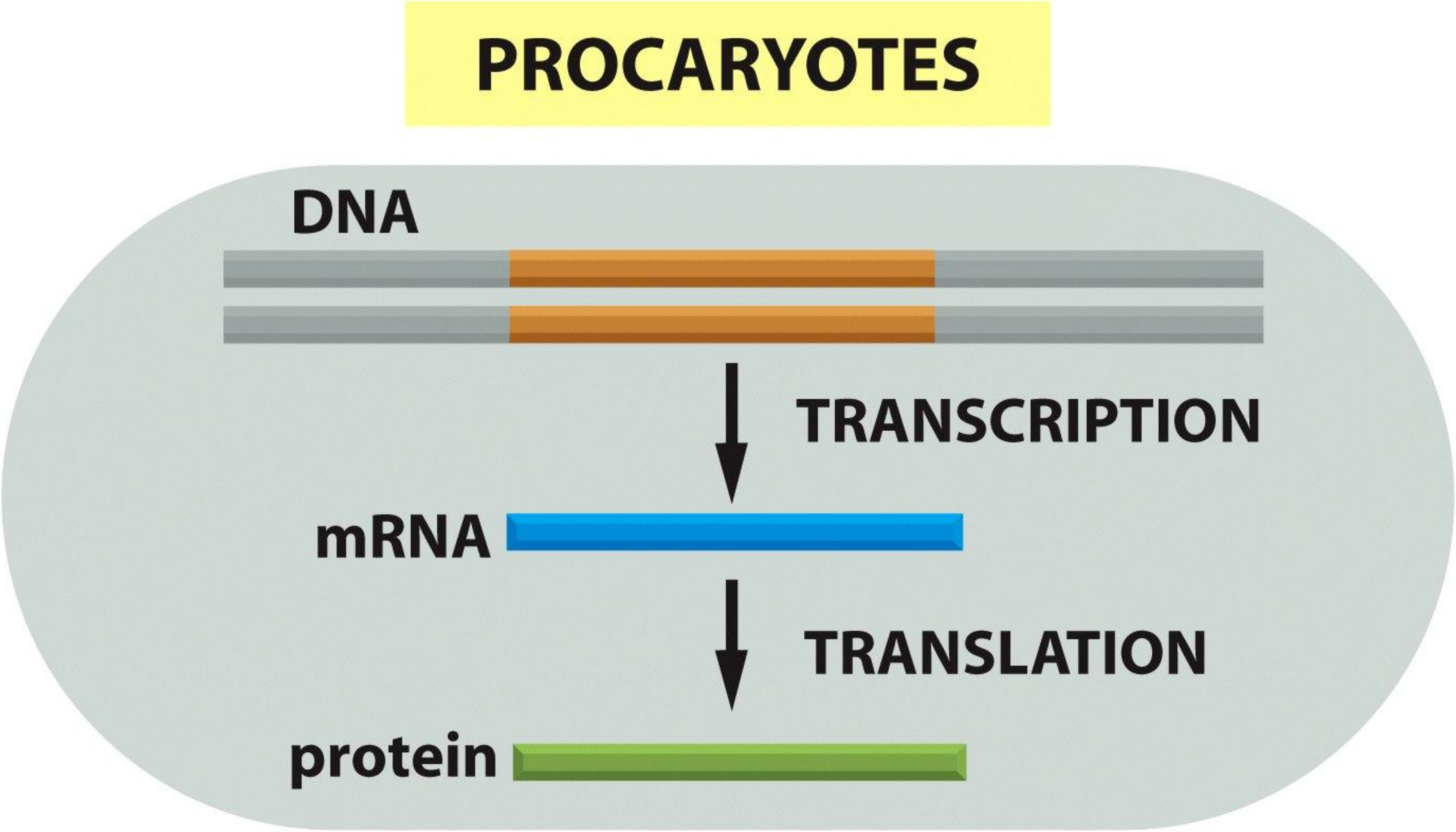
Research in Life



Domain	Currently Sequenced	In progress
BACTERIA	2,847	7,908
ARCHAEA	153	213
EUKARYOTES	173	2,385

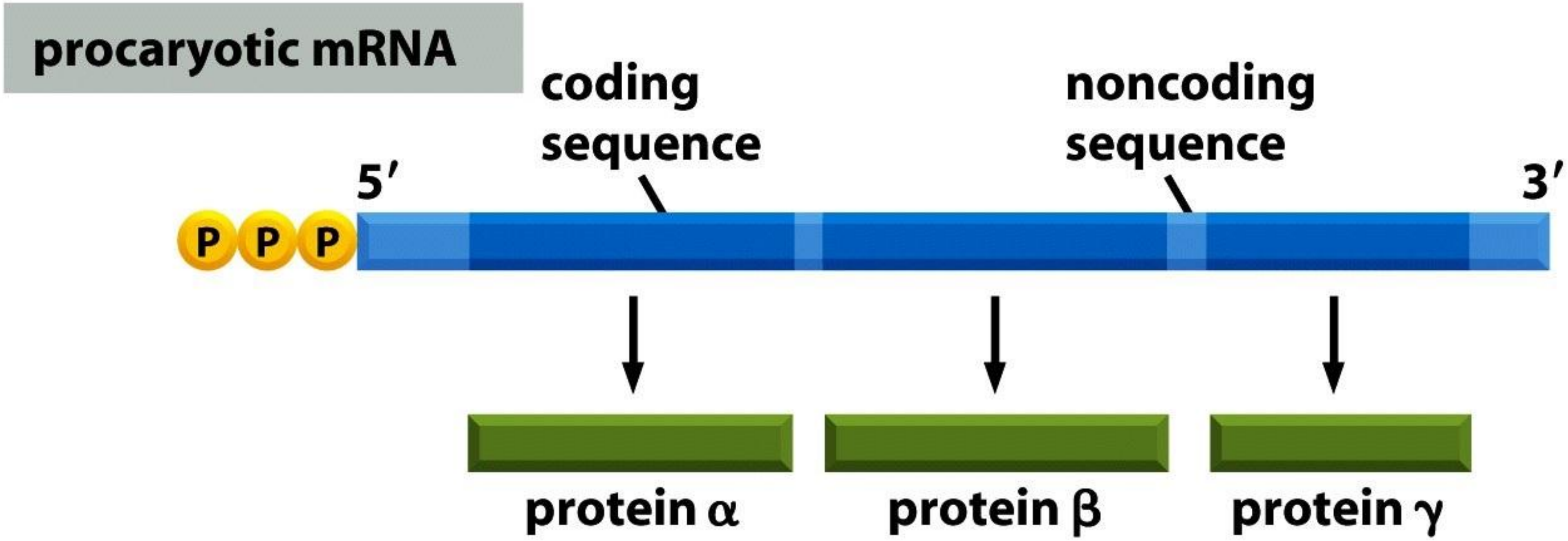
Courtesy of Genome Online Database, last update: March 12, 2012

Central Dogma of Biology



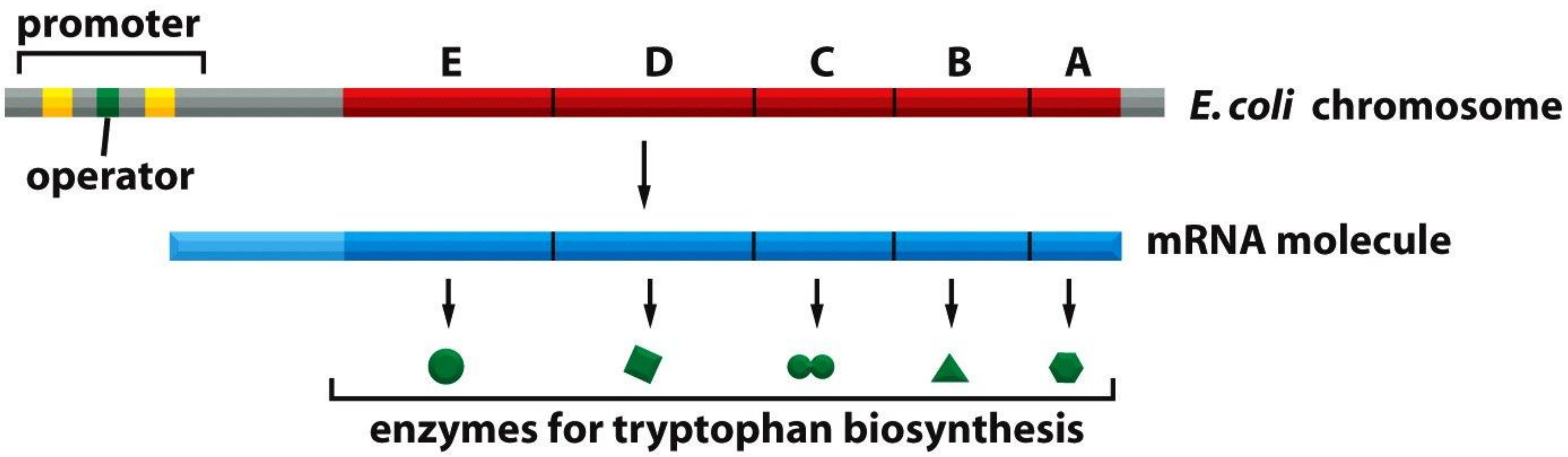
Courtesy of Molecular Biology of the Cell 5/e (© Garland Science 2008)

Operons



Courtesy of Molecular Biology of the Cell 5/e (© Garland Science 2008)

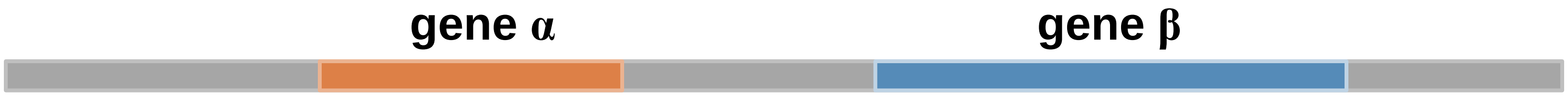
Operons



- □ Functional prediction
- □ Gene transcription
- □ Gene regulation

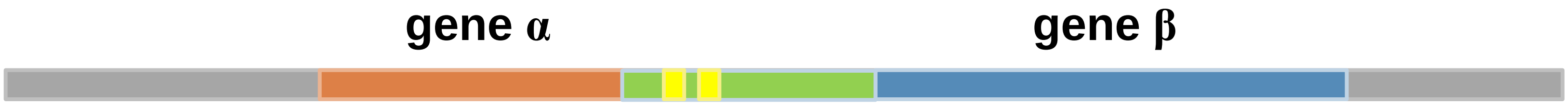
Courtesy of Molecular Biology of the Cell 5/e (© Garland Science 2008)

The Question



- Gene pair in an operon?
- Prediction vs. Verification
- Features

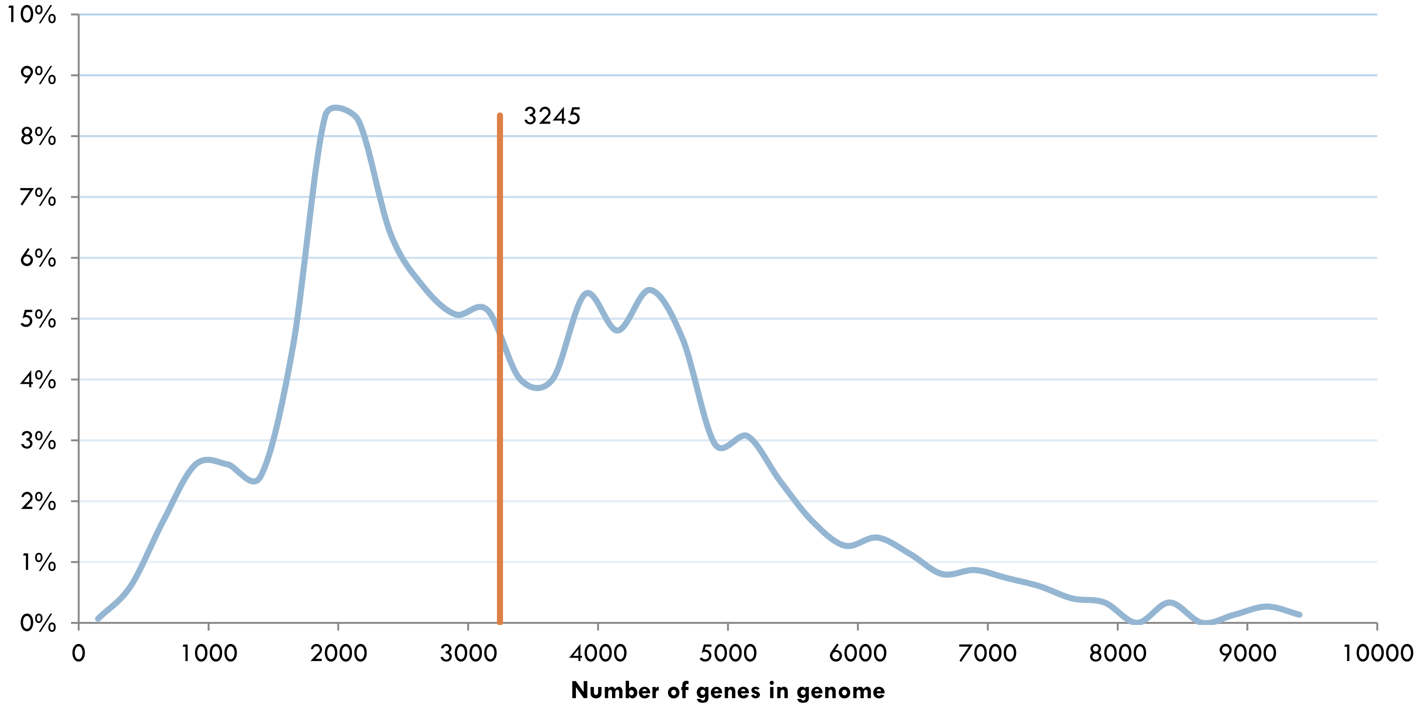
Features



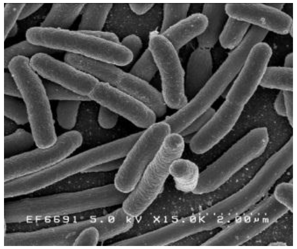
- Intergenic Distance
- Transcription Factors
- Gene Pair Neighborhood Conservation
- Mutual Expression level
- Functional Similarities

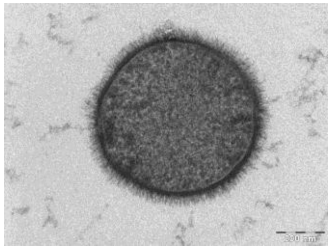
Prokaryote s

- 1499 published genomes
  - Genome
  - Predicted genes



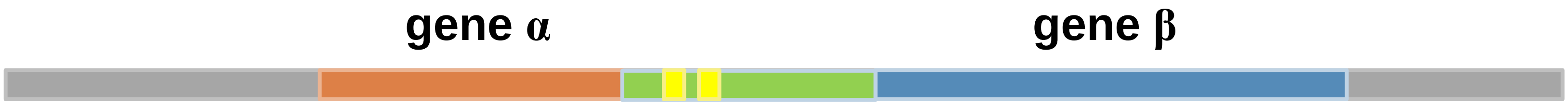
Model Organisms





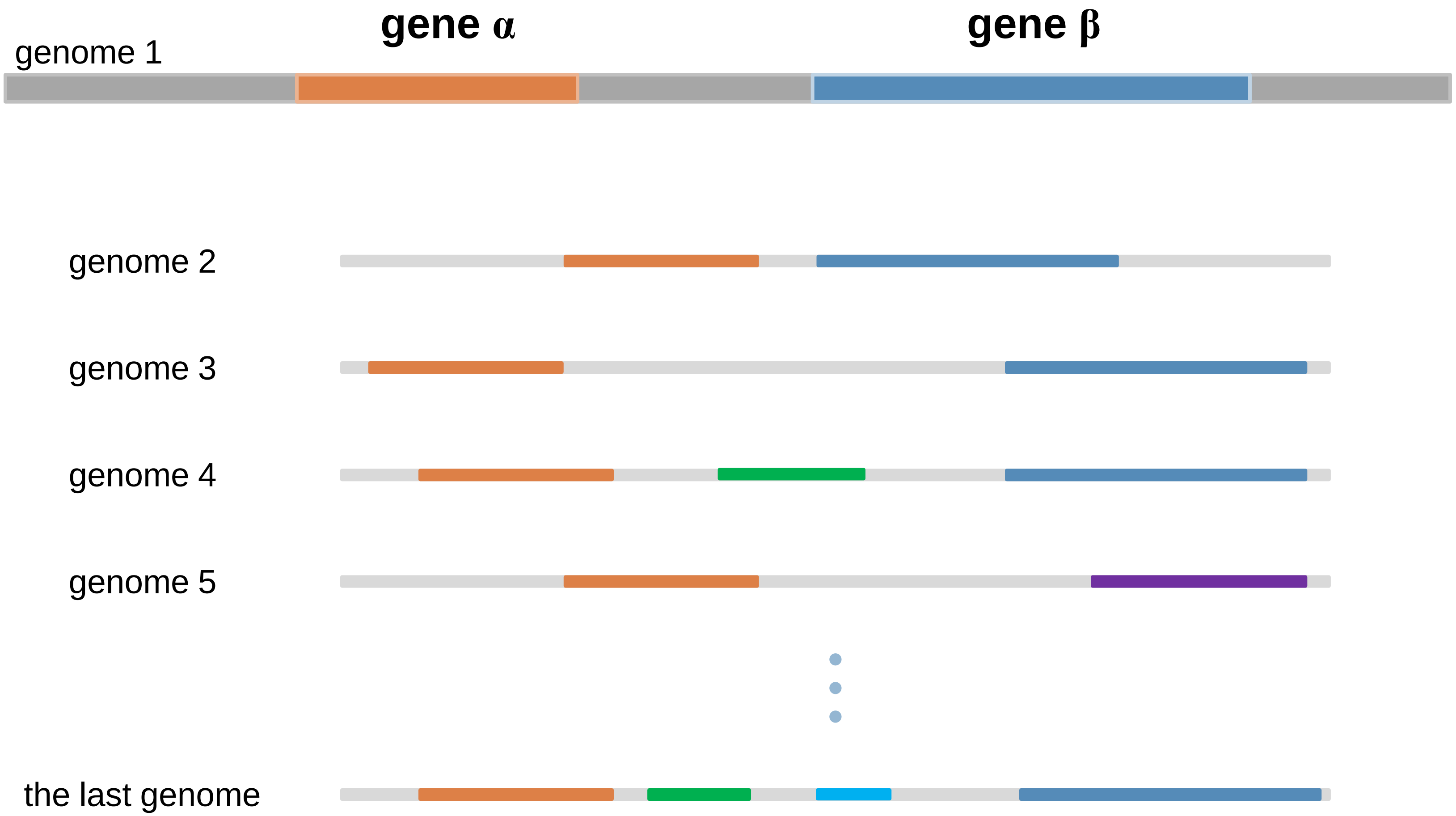
	Escherichia coli K12 MG1655	*Bacillus subtilis* W168
Predicted Genes	4144	4176
Known Operons	843	166
Gene pairs in Operon	1588	387
	38.3%	9.3%

Features

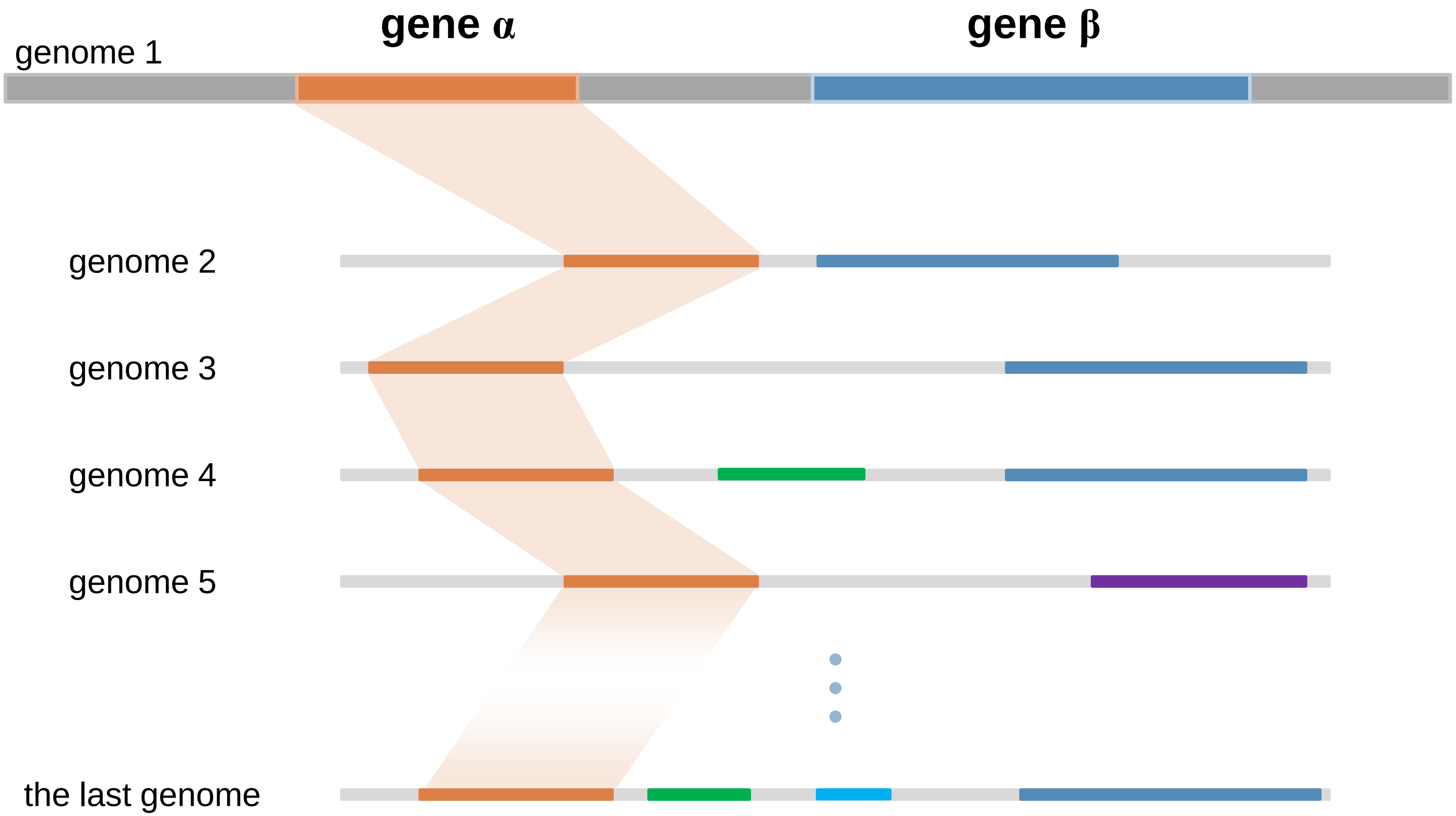


- □ Intergenic Distance
- □ Transcription Factors
- □ Gene Pair Neighborhood Conservation
- □ Mutual Expression level
- □ Functional Similarities

Gene neighborhood Conservation



Orthologous Genes



Gene Neighborhood Conservation

- Find Orthologous genes
  - pairwise BLASTs
  - Processing results
- Calculate conservation values

  for each gene pair of every genome

BLAST

- String search tool
- Inexact gapped matches
- Search a query against a database
  - Search a gene in a genome
- Bidirectional BLAST hits to deduce orthology

On Cloud

- How many?
  - 1499 x 1498 = 2,245,502
- Runtime of each BLAST?
  - 2 minutes
- How long will it take on a single computer?
  - 2.2M x 2min = 3,118 days!

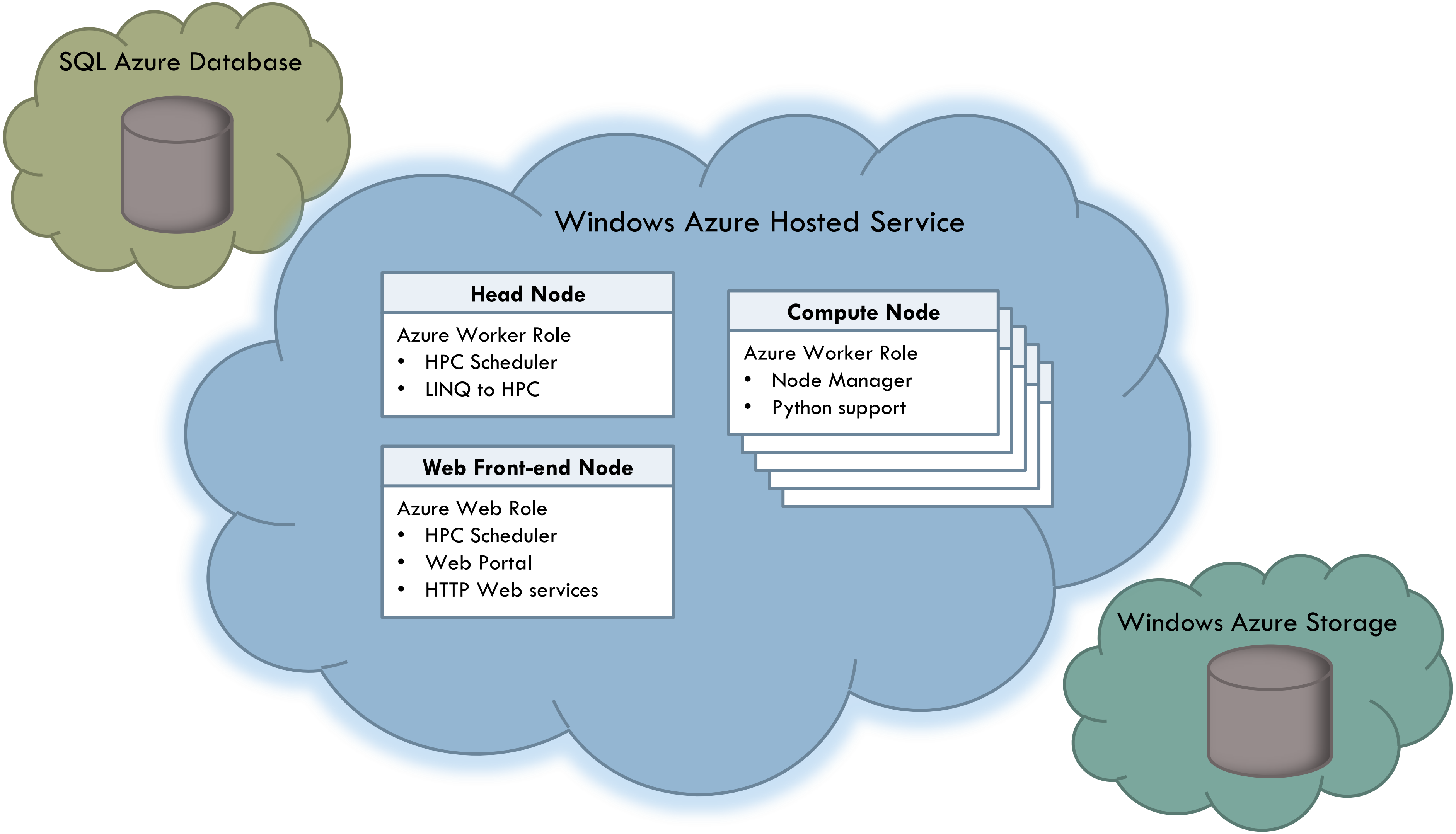
Azure Cloud

- Windows Azure HPC Scheduler
  - Parametric Sweep
- 300 small size nodes
  - 1 CPU core
  - 2GB Memory
- 20TB of Azure Storage

Windows Azure HPC Scheduler

- Parametric Sweep
  - Embarrassingly parallel problems
- SOA Services
  - Distributed applications
- MPI Applications
  - Platform-independent standard

HPC Architecture



SQL Azure Database

Windows Azure Hosted Service

Head Node

Azure Worker Role
- HPC Scheduler
- LINQ to HPC

Web Front-end Node

Azure Web Role
- HPC Scheduler
- Web Portal
- HTTP Web services

Compute Node

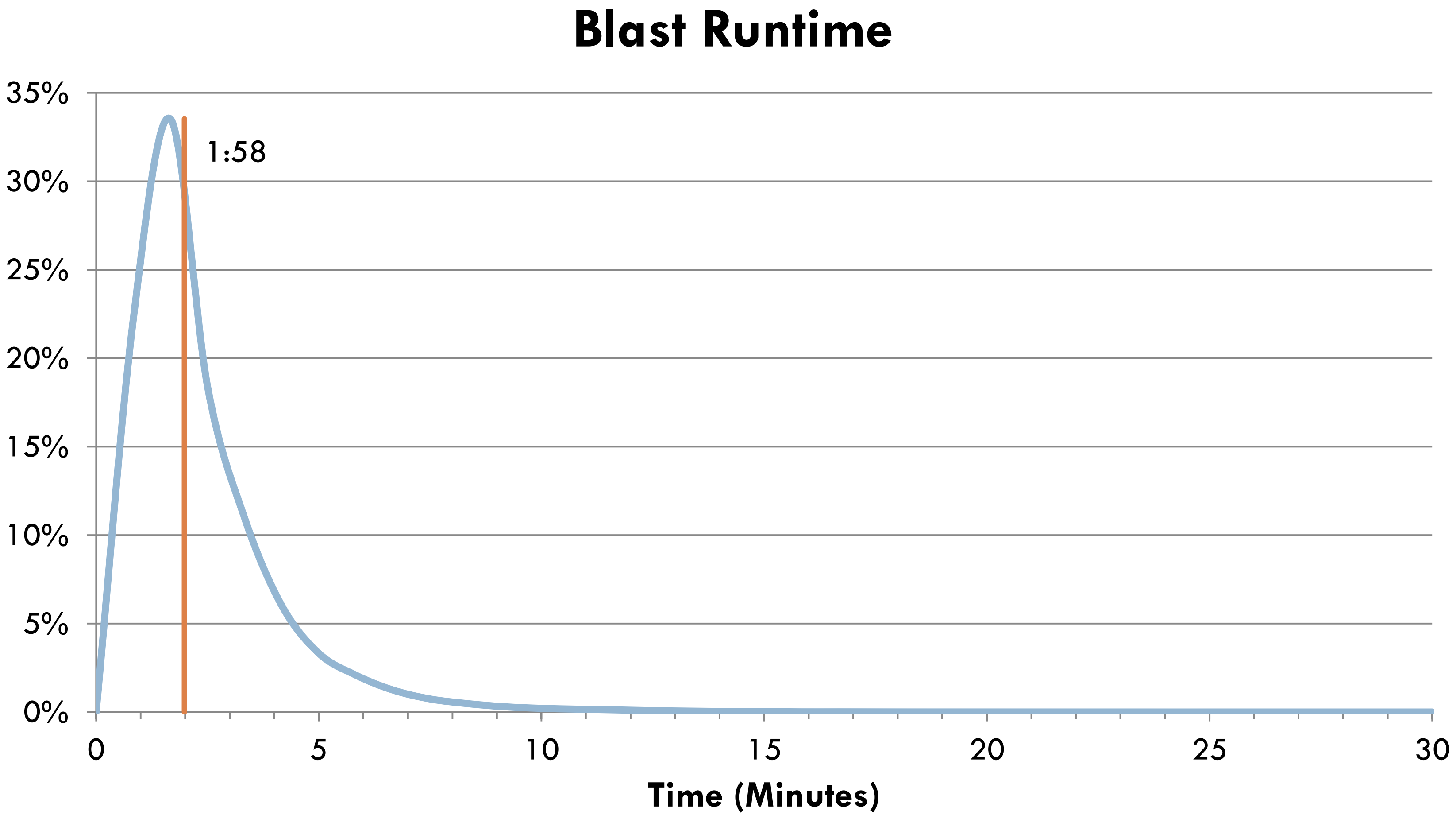
Azure Worker Role
- Node Manager
- Python support

Windows Azure Storage

Parametric Sweep BLAST

- Deploy a hosted service
- Run
  - Downloads required data from storage account to the work node
  - Run BLAST
  - Uploads the results back to storage account
- Check Fails, log files, …

Blast run time



Storage

- Genome
  - Including blast databases: 11.1 GB
- BLAST
  - 7.5 TB
- The rest
  - Calculated results: 25 GB

Total cost

- Computation
  - About $9,000 in total
- Storage
  - About $900 / month

Importance

- Computationally intensive research
  - Cost
  - Time
- Prediction Power
  - Incorporate as much data
- Future research
  - Availability of data

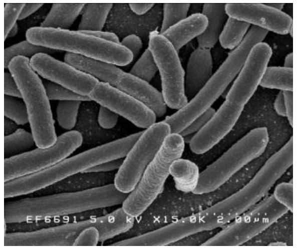
Thank You

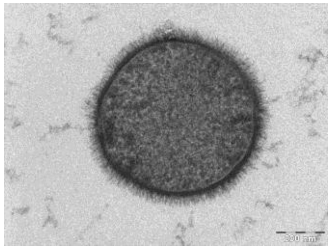
UNC CHARLOTTE

- Acknowledgments
  - Dr. ZhengChang Su
  - Bioinformatics @ UNC Charlotte
  - Microsoft support team

EXTRA SLIDES

Model Organisms





	Escherichia coli K12 MG1655	***Bacillus subtilis*** W168
Predicted Genes	4144	4176
Known Operons	843	166
Gene pairs in Operon	1588	387
	38.3%	9.3%

Conservation

Genome	2	3	4	5	6	…	1499
gene α	√	√	√	√	×		√
gene β	√	√	√	×	×		×
D_g	1/1	1/1	1/2				1/3

$$\log \frac{\sum_K \frac{1}{1 + D_g\left(g_i, g_j, K\right)}}{D_H\left(g_i, g_j\right)}$$