Update on Argonne/Chicago BioGrid Developments

Rick Stevens

Argonne National Laboratory

The University of Chicago

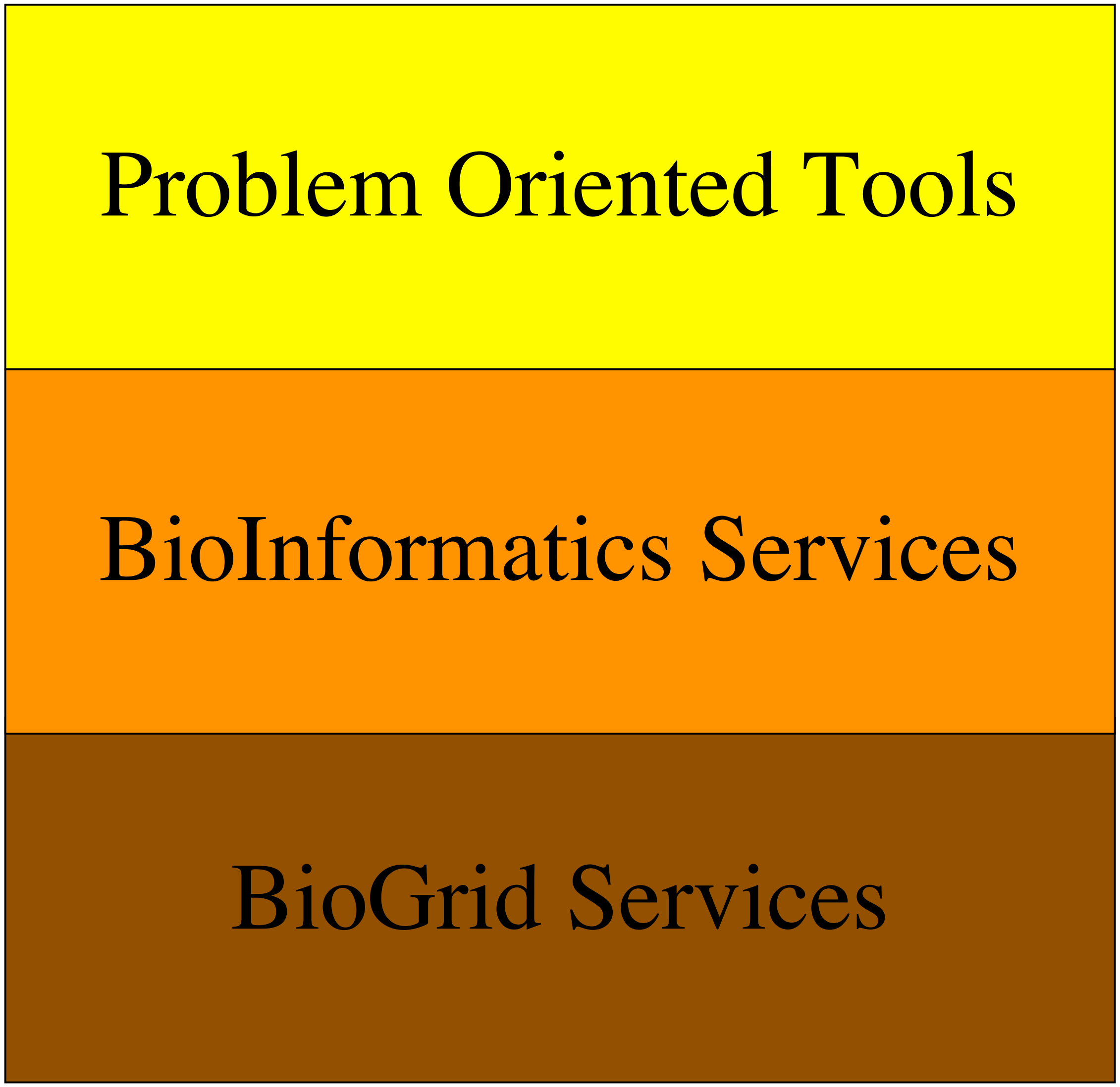
Stevens@cs.uchicago.edu

Acknowledgements

Many people have contributed to this work

- Ross Overbeek
- Natalia Maltsev
- Ed Frank
- Alex Rodrigez
- Dina Sulakhe
- Veronkia Vonstein
- Terry Disz
- Bob Olson
- Mark Hereld
- Mike Papka
- Mike Wilde
- Miron Livny
- Ian Foster
- Rick Stevens

An Example BioGrid Services Model



Problem Oriented Tools

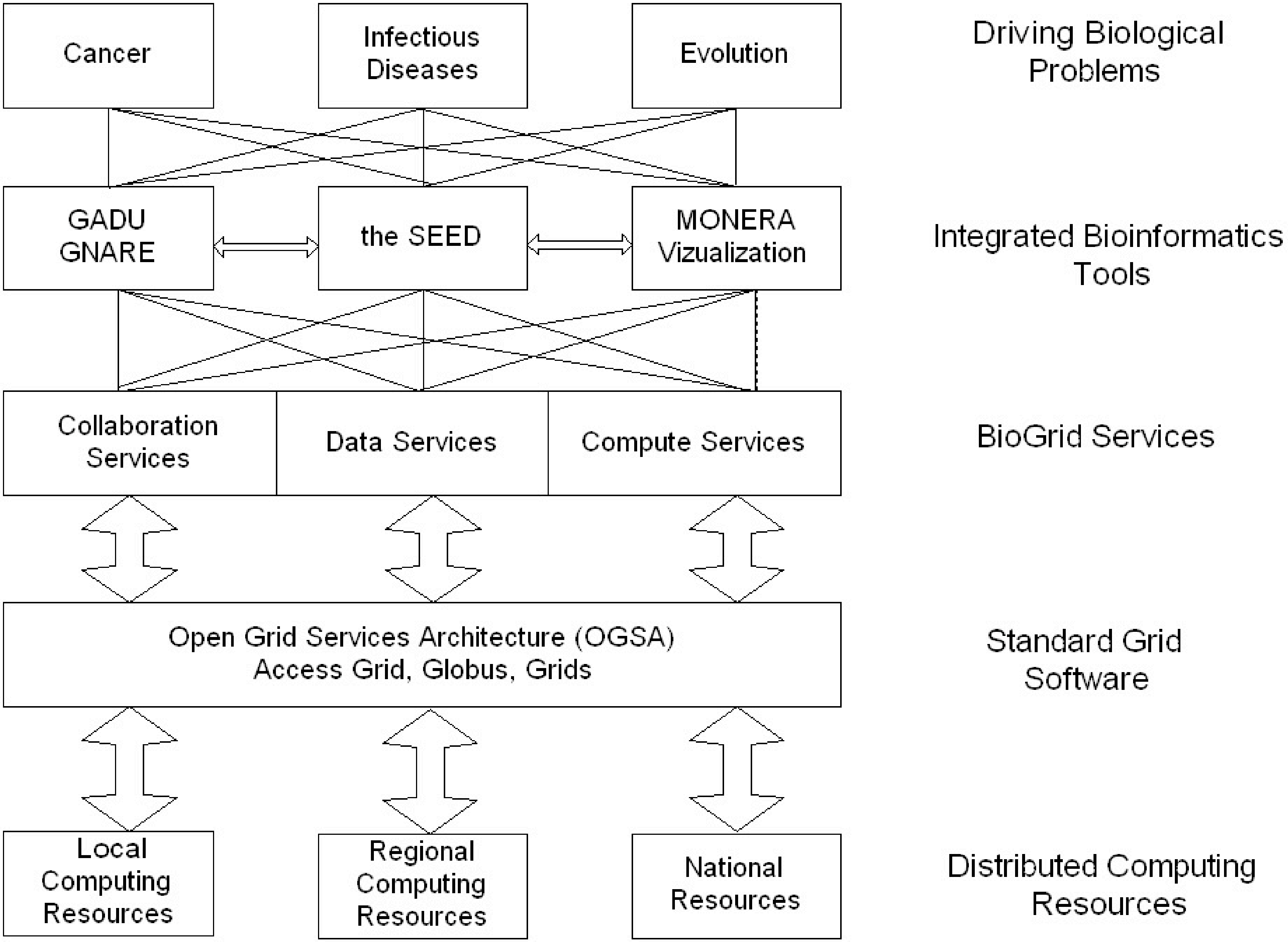
- Drug Discovery
- Microbial Engineering
- Molecular Ecology
- Oncology Research

BioInformatics Services

- Integrated Databases
- Sequence Analysis
- Protein Interactions
- Cell Simulation

BioGrid Services

- Compute Services
- Workflow Services
- Data Service
- Collaboration Services



Cancer

Infectious Diseases

Evolution

Driving Biological Problems

GADU GNARE

the SEED

MONERA Vizualization

Integrated Bioinformatics Tools

Collaboration Services

Data Services

Compute Services

BioGrid Services

Open Grid Services Architecture (OGSA)
Access Grid, Globus, Grids

Standard Grid Software

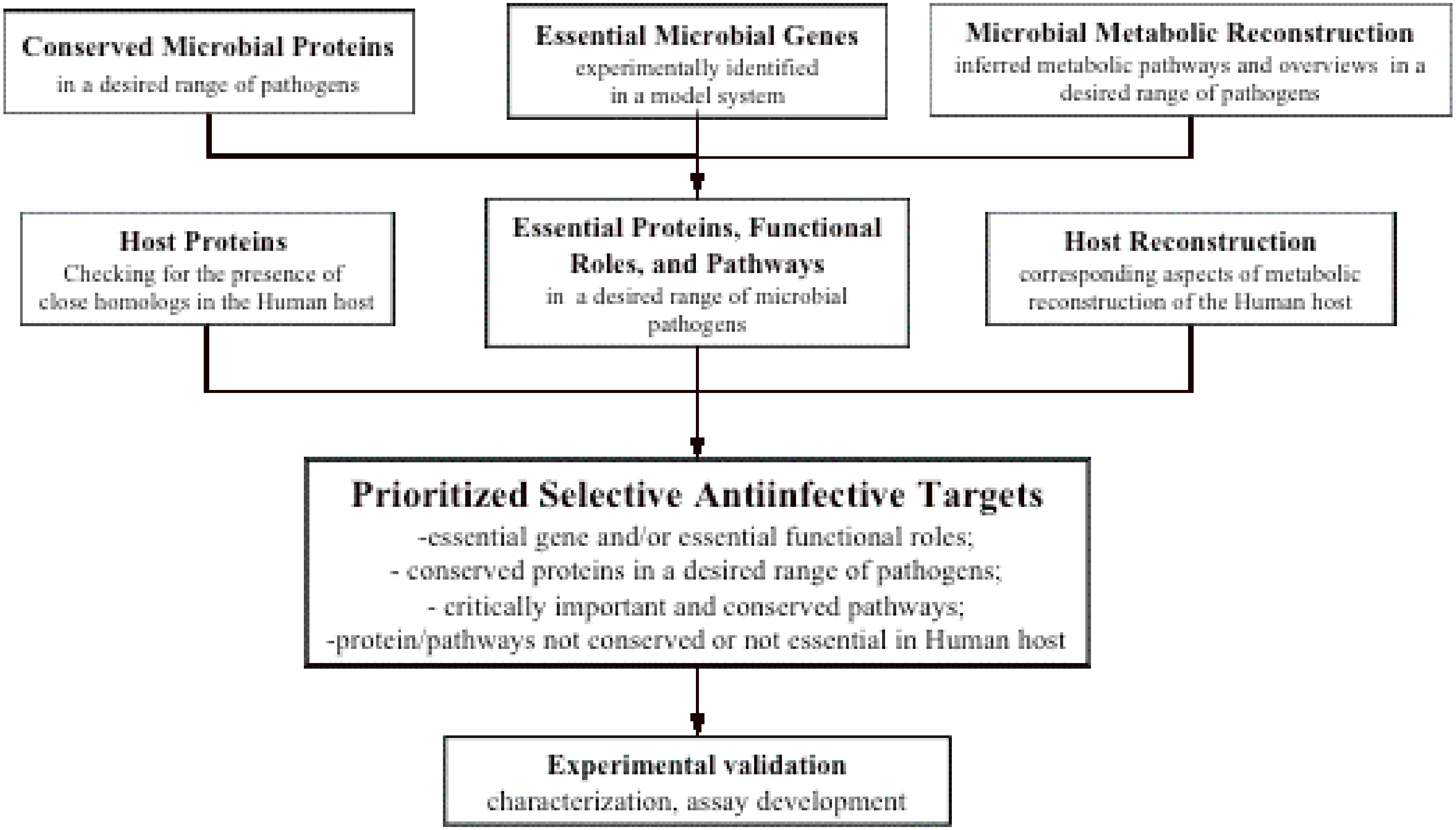
Local Computing Resources

Regional Computing Resources

National Resources

Distributed Computing Resources

Functional Genomics Approach to Anti-microbial Agent Development



Conserved Microbial Proteins
in a desired range of pathogens

Essential Microbial Genes
experimentally identified
in a model system

Microbial Metabolic Reconstruction
inferred metabolic pathways and overviews in a desired range of pathogens

Host Proteins
Checking for the presence of close homologs in the Human host

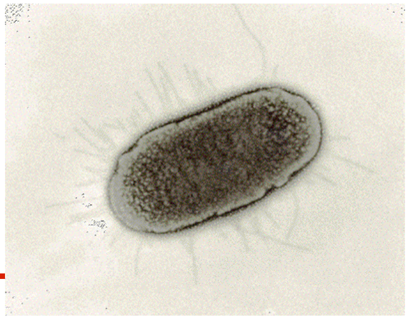
Essential Proteins, Functional Roles, and Pathways
in a desired range of microbial pathogens

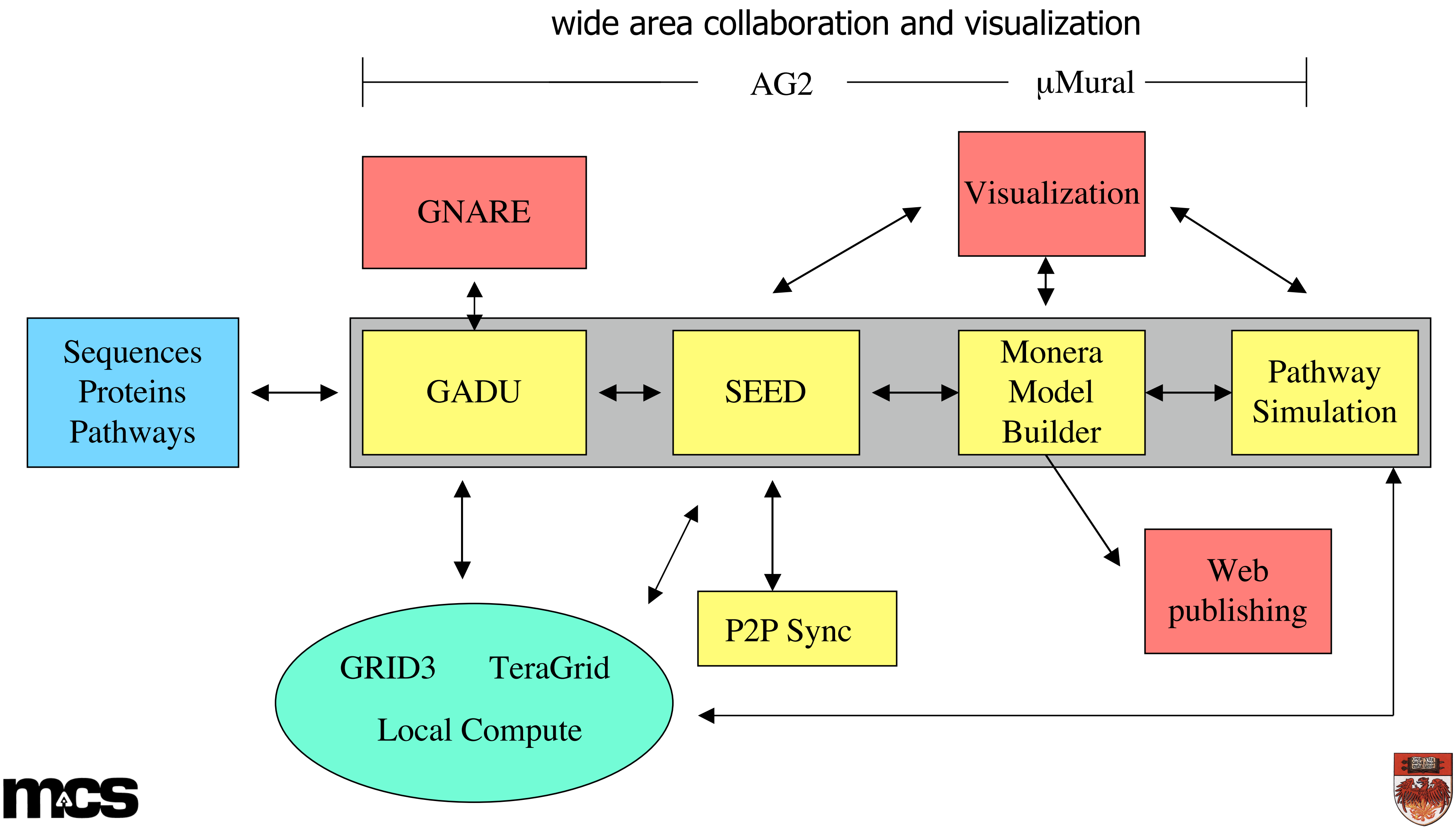
Host Reconstruction
corresponding aspects of metabolic reconstruction of the Human host

Prioritized Selective Antiinfective Targets
-essential gene and/or essential functional roles;
- conserved proteins in a desired range of pathogens;
- critically important and conserved pathways;
-protein/pathways not conserved or not essential in Human host

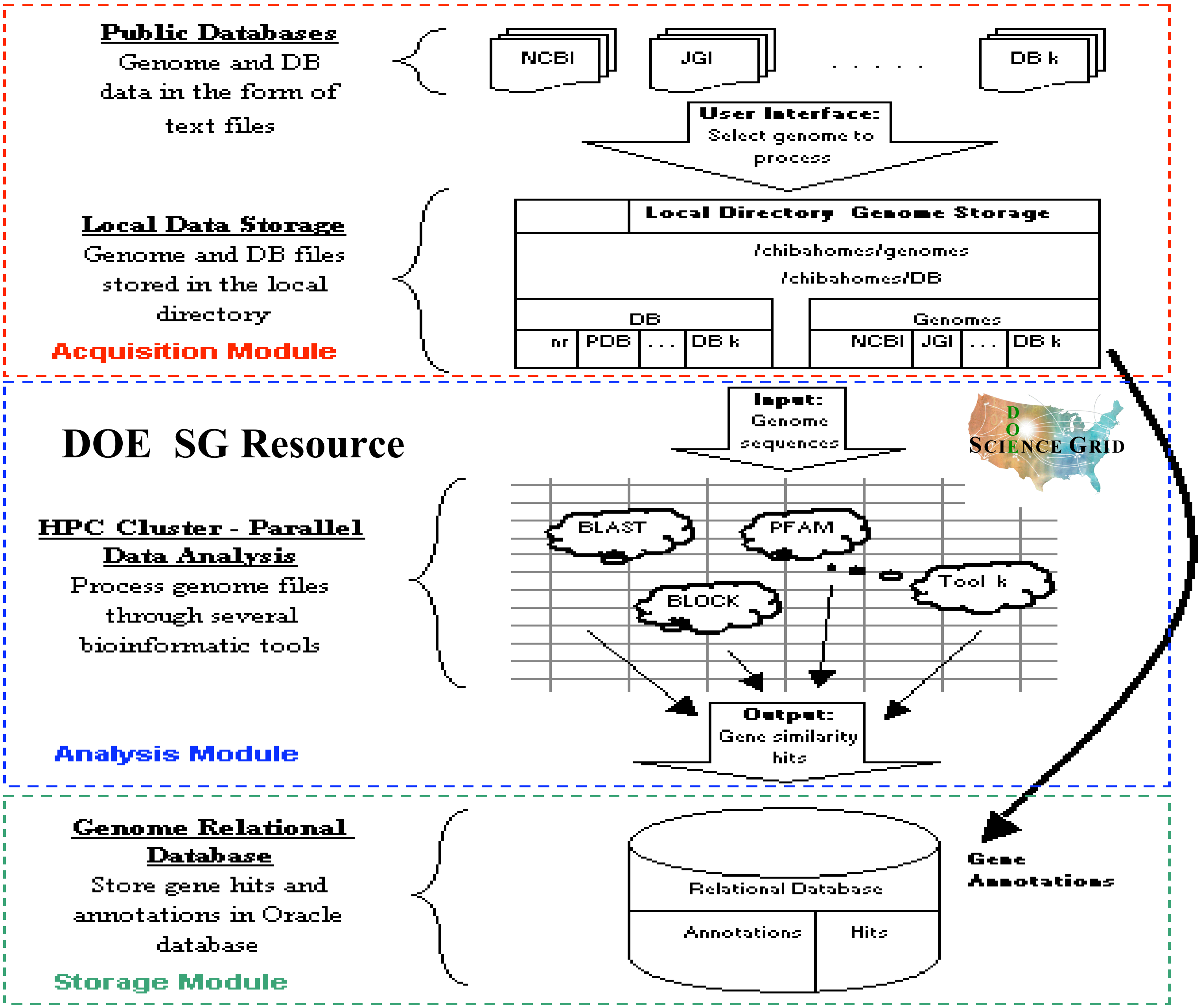
Experimental validation
characterization, assay development

Argonne Systems Biology Workflow



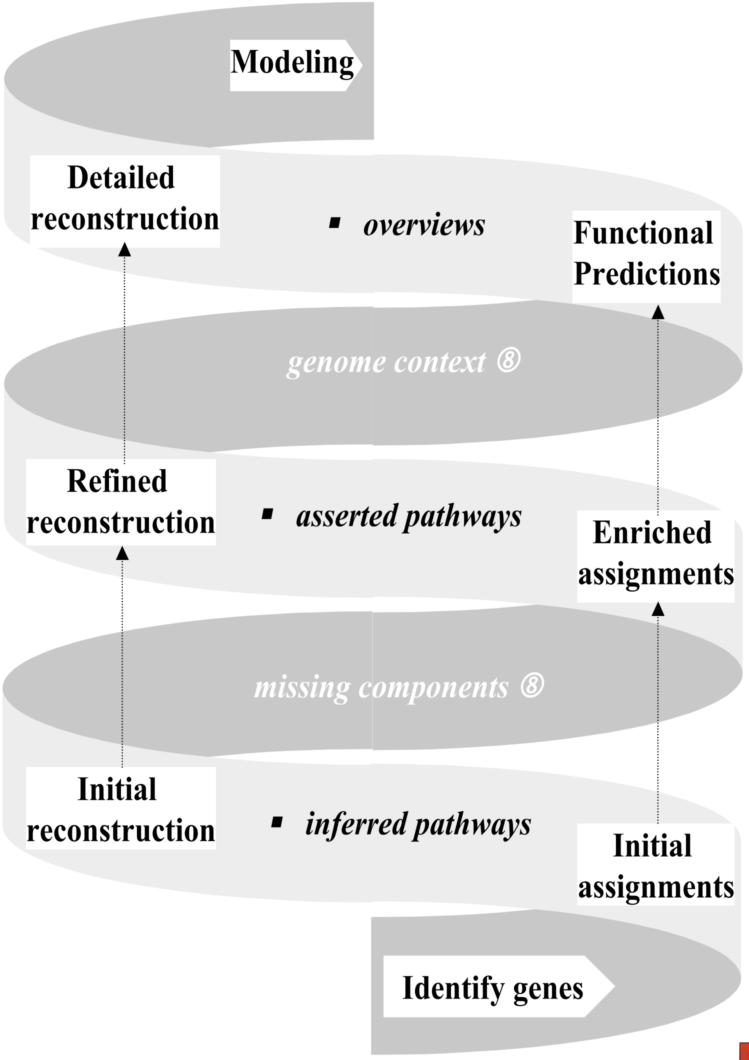


GADU Data Flow


Public Databases
Genome and DB data in the form of text files
NCBI
JGI
. . . . . .
DB k
User Interface: Select genome to process
Local Data Storage
Genome and DB files stored in the local directory
Local Directory Genome Storage
/chibahomes/genomes
/chibahomes/DB
DB
nr
PDB
. . .
DB k
Genomes
NCBI
JGI
. . .
DB k
Acquisition Module
DOE SG Resource
Input: Genome sequences
DOE SCIENCE GRID
HPC Cluster - Parallel Data Analysis
Process genome files through several bioinformatic tools
BLAST
PFAM
BLOCK
Tool k
Output: Gene similarity hits
Analysis Module
Genome Relational Database
Store gene hits and annotations in Oracle database
Relational Database
Annotations
Hits
Gene Annotations
Storage Module

What Is Annotation All About?

- We identify the genes
- We identify clear functions and make tentative guesses
- We build an initial metabolic reconstruction
- We refine all estimates iteratively, seeking maximal consistency
- We identify "missing genes"
- We build "portfolios" to support identification of missing genes
- We confirm these assertions.
- We project the results



Modeling

Detailed reconstruction

▪ *overviews*

Functional Predictions

genome context ⑧

Refined reconstruction

▪ *asserted pathways*

Enriched assignments

missing components ⑧

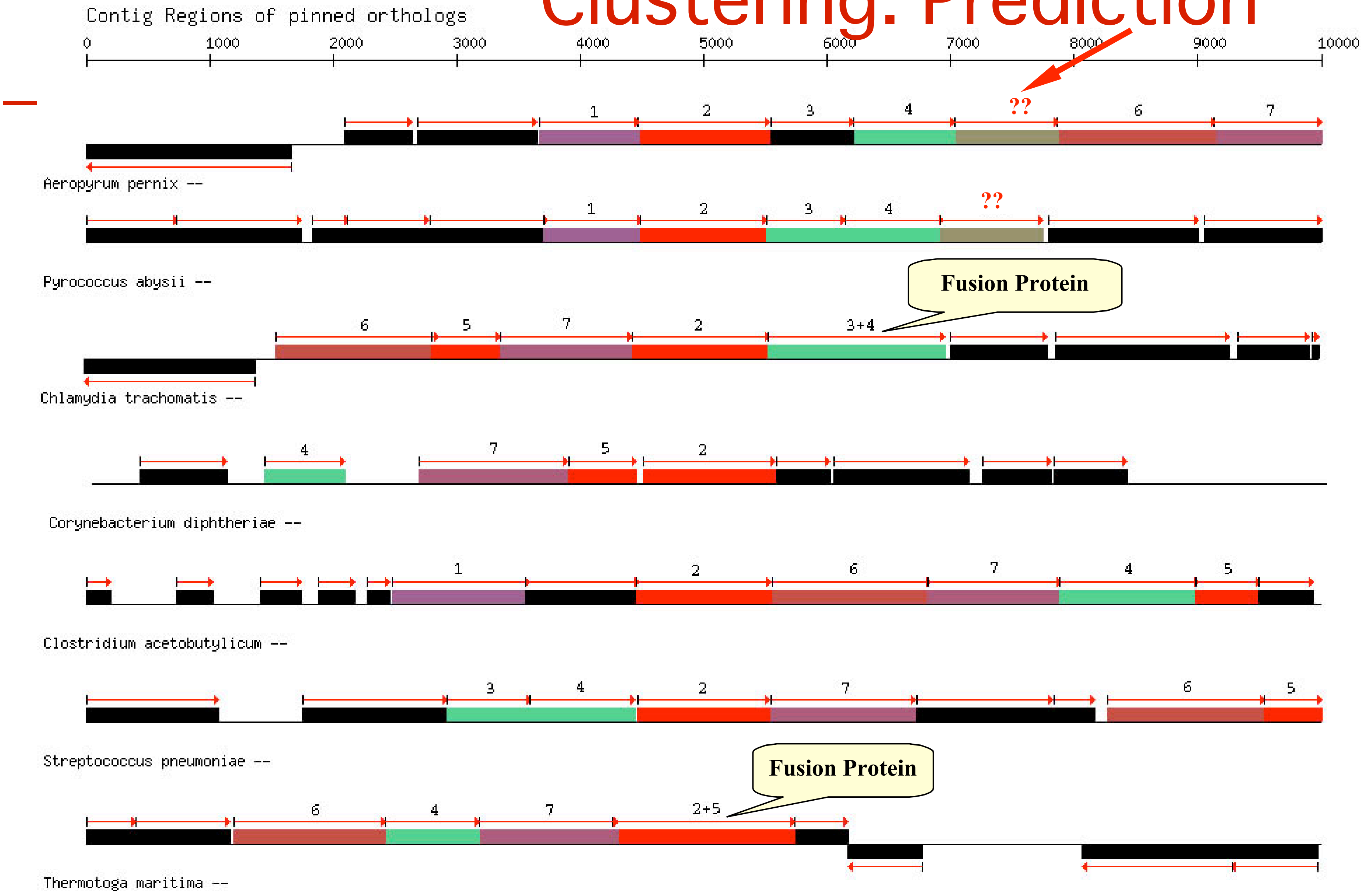
Initial reconstruction

▪ *inferred pathways*

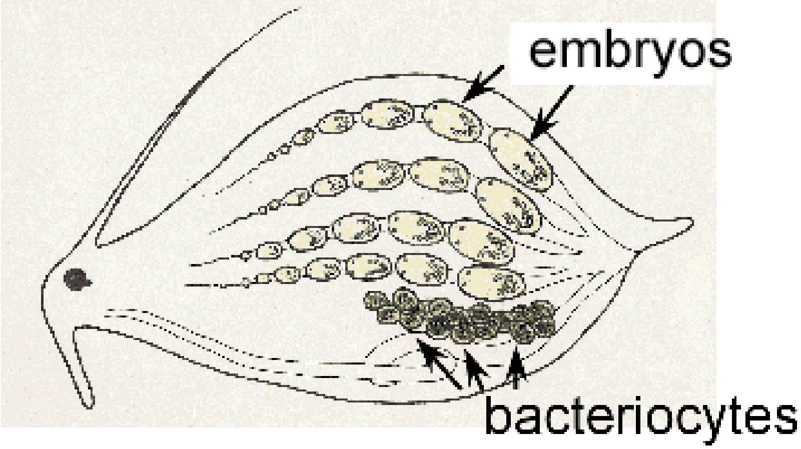
Initial assignments

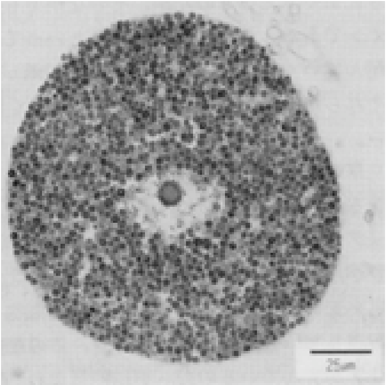
Identify genes

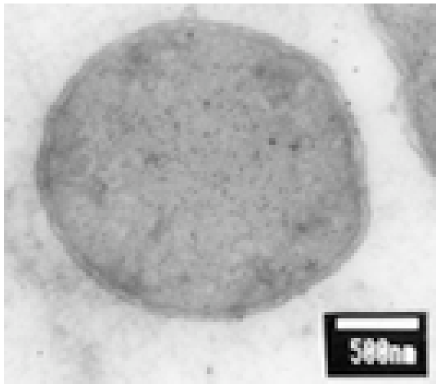
Chromosomal Clustering: Prediction

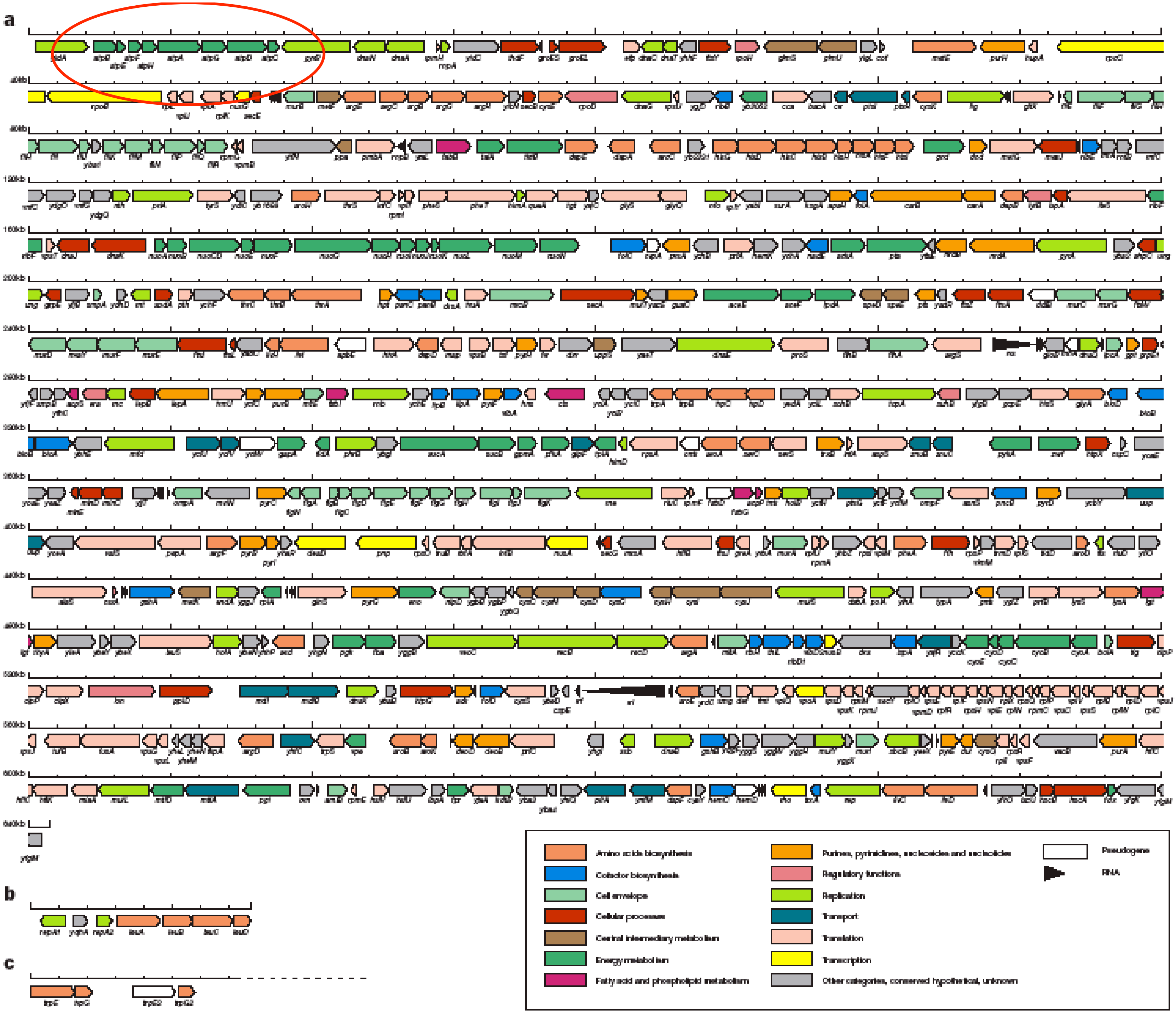


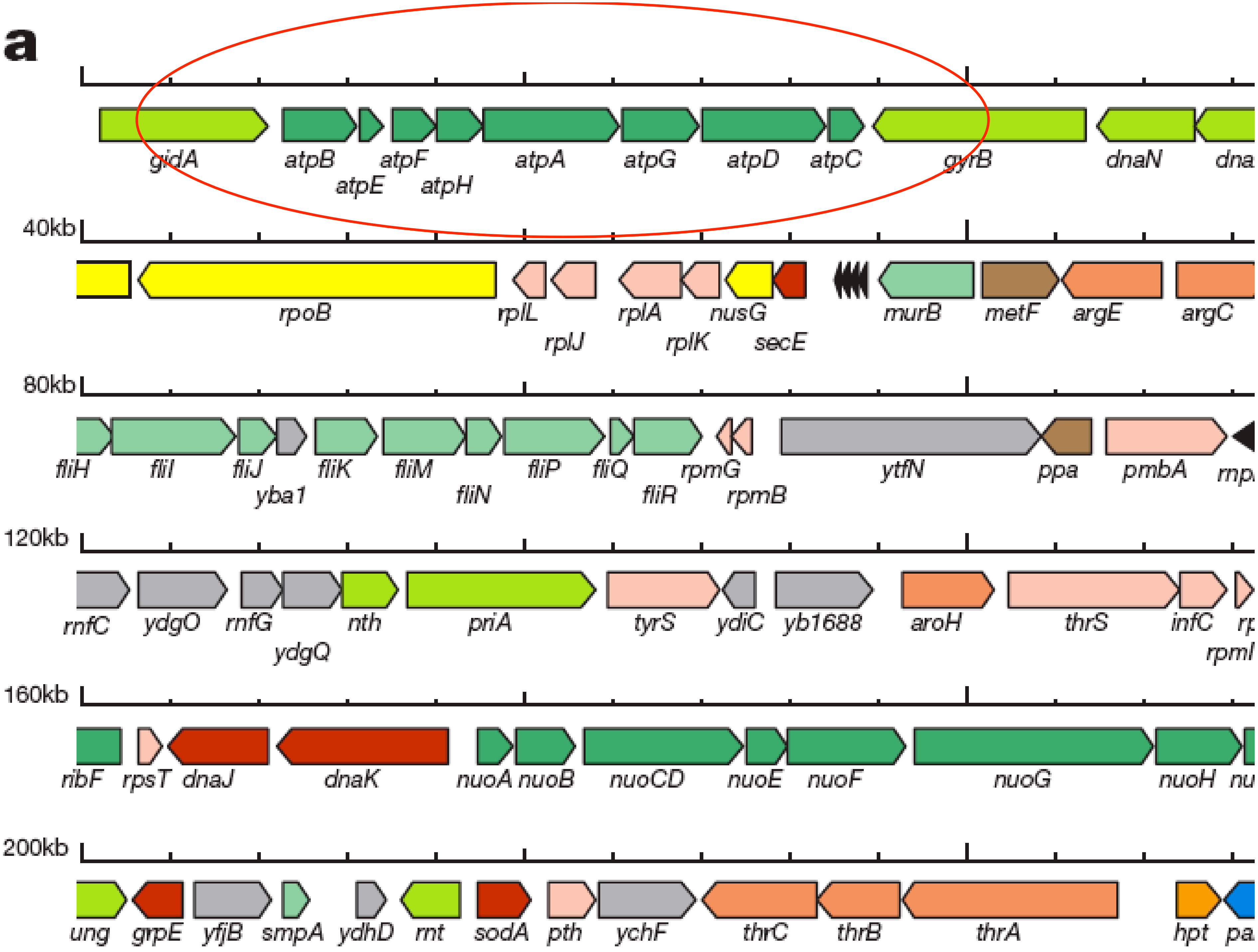
Buchnera sp.



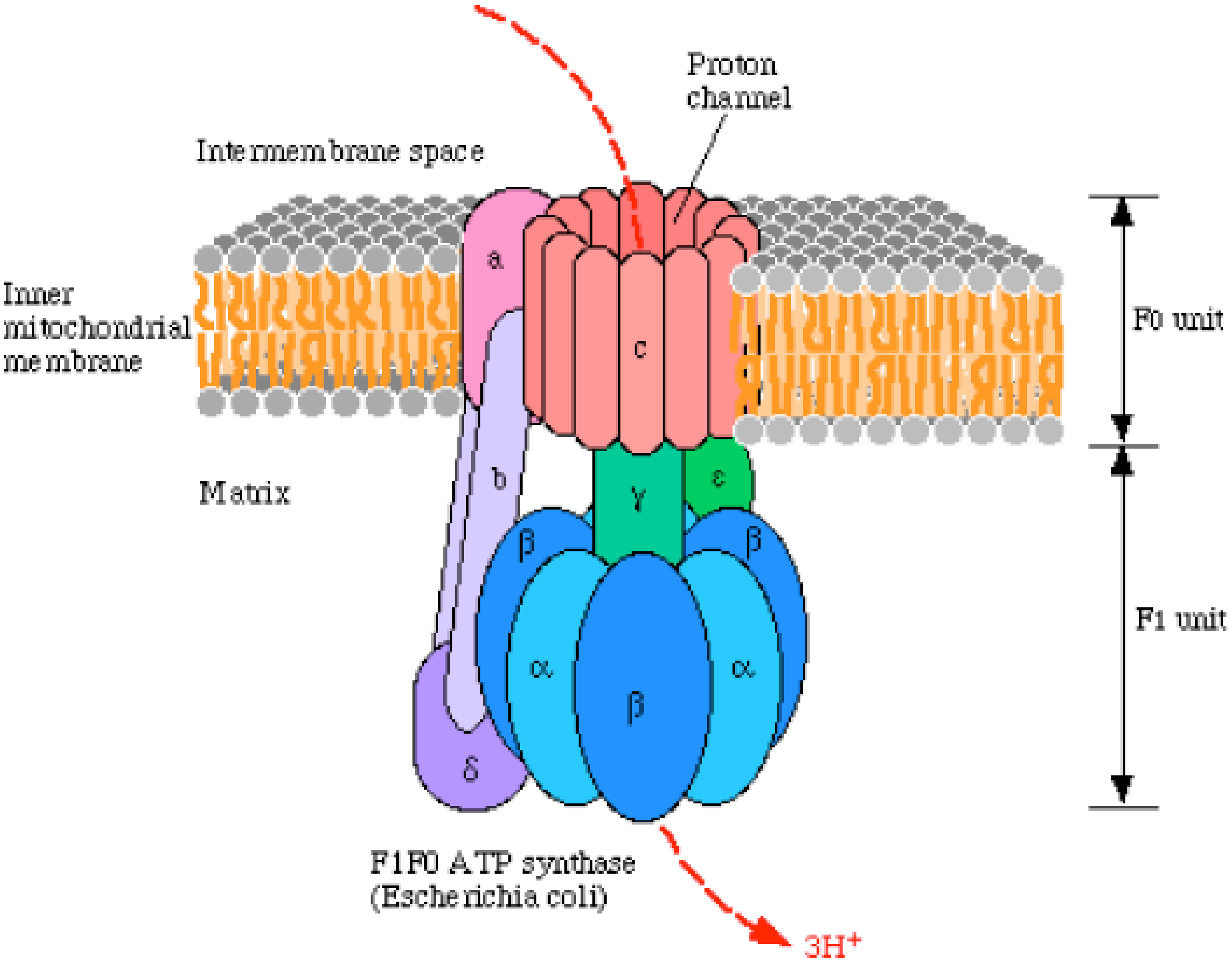





a
atpB
atpE
atpF
atpH
atpA
atpG
atpD
atpC
gyrB
dnaN
dnaA
rpmH
rnpA
yidC
thdF
groES
groEL
rpoC
rpoB
ygbM
b
repA1
yqhA
repA2
leuA
leuB
leuC
leuD
c
trpE
trpG
trpE2
trpG2
Amino acids biosynthesis
Cofactor biosynthesis
Cell envelope
Cellular processes
Central intermediary metabolism
Energy metabolism
Fatty acid and phospholipid metabolism
Purines, pyrimidines, nucleosides and nucleotides
Regulatory functions
Replication
Transport
Translation
Transcription
Other categories, conserved hypothetical, unknown
Pseudogene
RNA


a
gidA
atpB
atpE
atpF
atpH
atpA
atpG
atpD
atpC
gyrB
dnaN
dna
40kb
rpoB
rplL
rplJ
rplA
rplK
nusG
secE
murB
metF
argE
argC
80kb
fliH
fliI
fliJ
yba1
fliK
fliM
fliN
fliP
fliQ
fliR
rpmG
rpmB
ytfN
ppa
pmbA
rnp
120kb
rnfC
ydgO
rnfG
ydgQ
nth
priA
tyrS
ydiC
yb1688
aroH
thrS
infC
rp
rpml
160kb
ribF
rpsT
dnaJ
dnaK
nuoA
nuoB
nuoCD
nuoE
nuoF
nuoG
nuoH
nu
200kb
ung
grpE
yfjB
smpA
ydhD
rnt
sodA
pth
ychF
thrC
thrB
thrA
hpt
pa

ATP SYNTHESIS



F1F0 ATP synthase (Escherichia coli)

F-type ATPase (Bacteria)

beta	alpha	gamma	delta	epsilon	c	a	b

Nanoarchaeum equitans

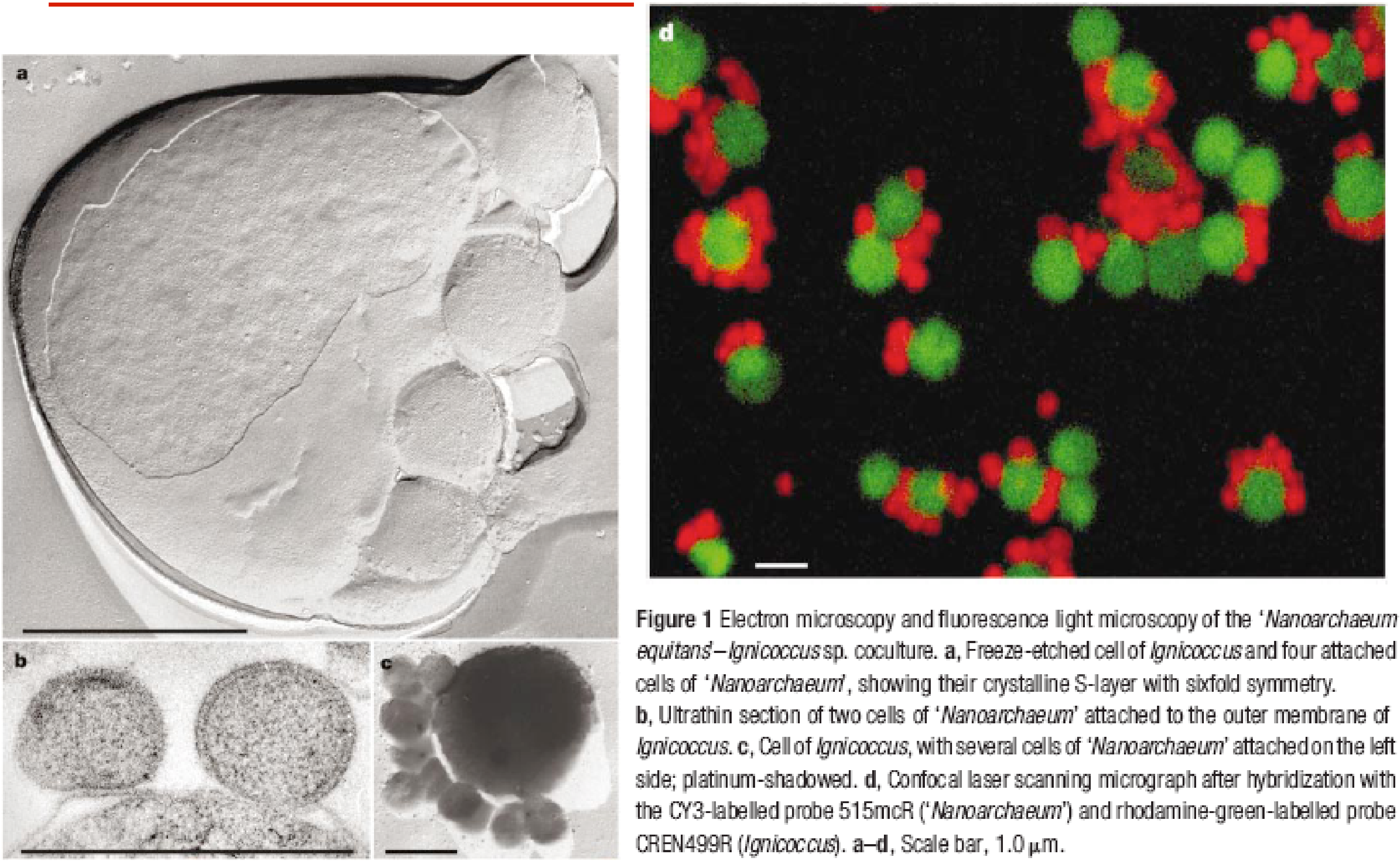
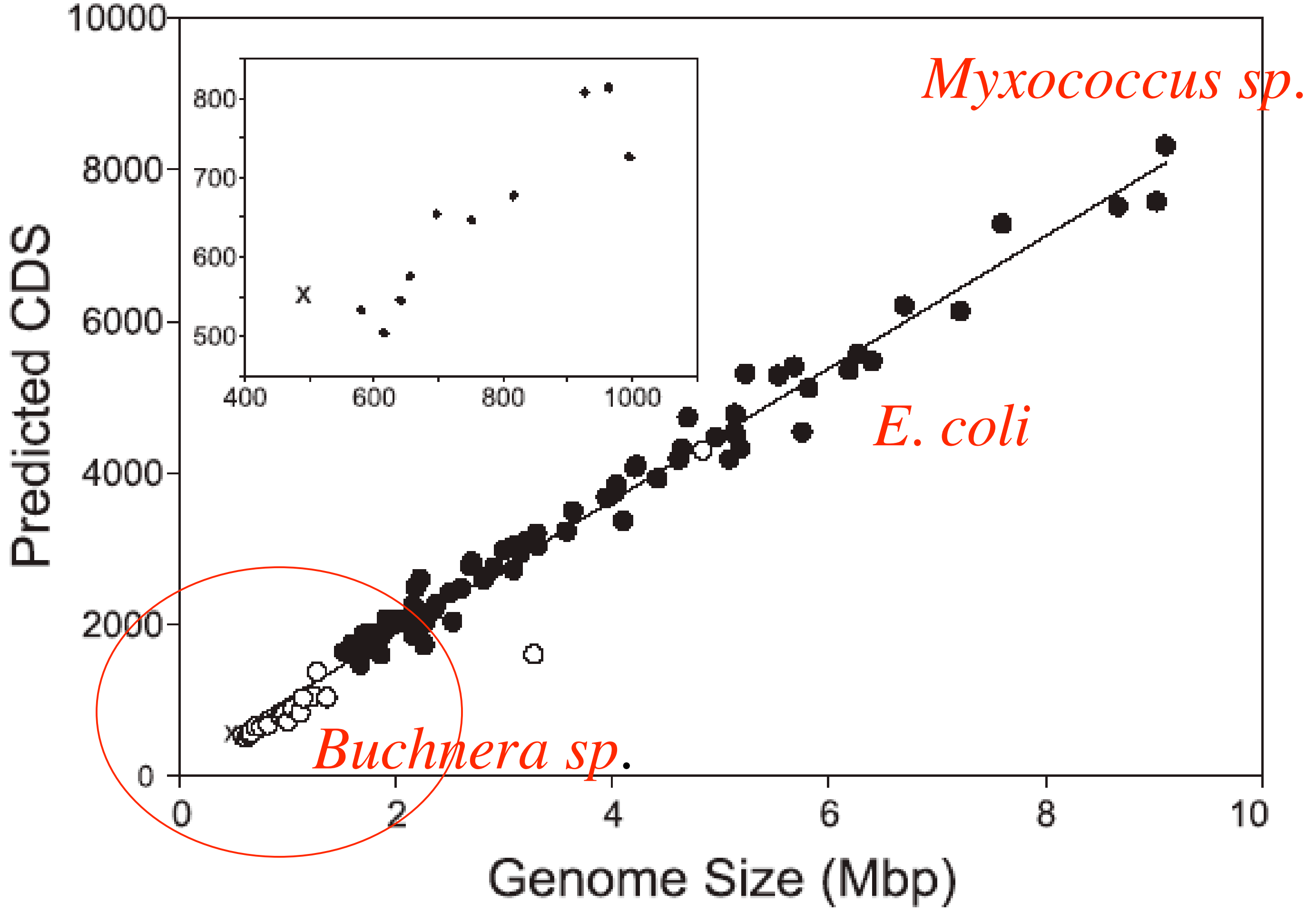


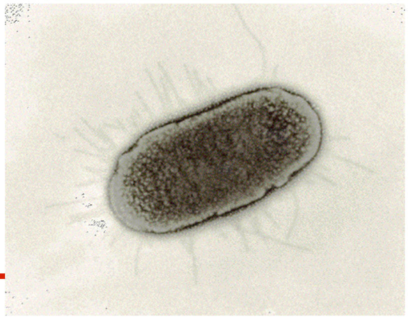
Figure 1 Electron microscopy and fluorescence light microscopy of the '*Nanoarchaeum equitans*'–*Ignicoccus* sp. coculture. **a,** Freeze-etched cell of *Ignicoccus* and four attached cells of '*Nanoarchaeum*', showing their crystalline S-layer with sixfold symmetry. **b,** Ultrathin section of two cells of '*Nanoarchaeum*' attached to the outer membrane of *Ignicoccus*. **c,** Cell of *Ignicoccus*, with several cells of '*Nanoarchaeum*' attached on the left side; platinum-shadowed. **d,** Confocal laser scanning micrograph after hybridization with the CY3-labelled probe 515mcR ('*Nanoarchaeum*') and rhodamine-green-labelled probe CREN499R (*Ignicoccus*). **a–d,** Scale bar, 1.0 μm.

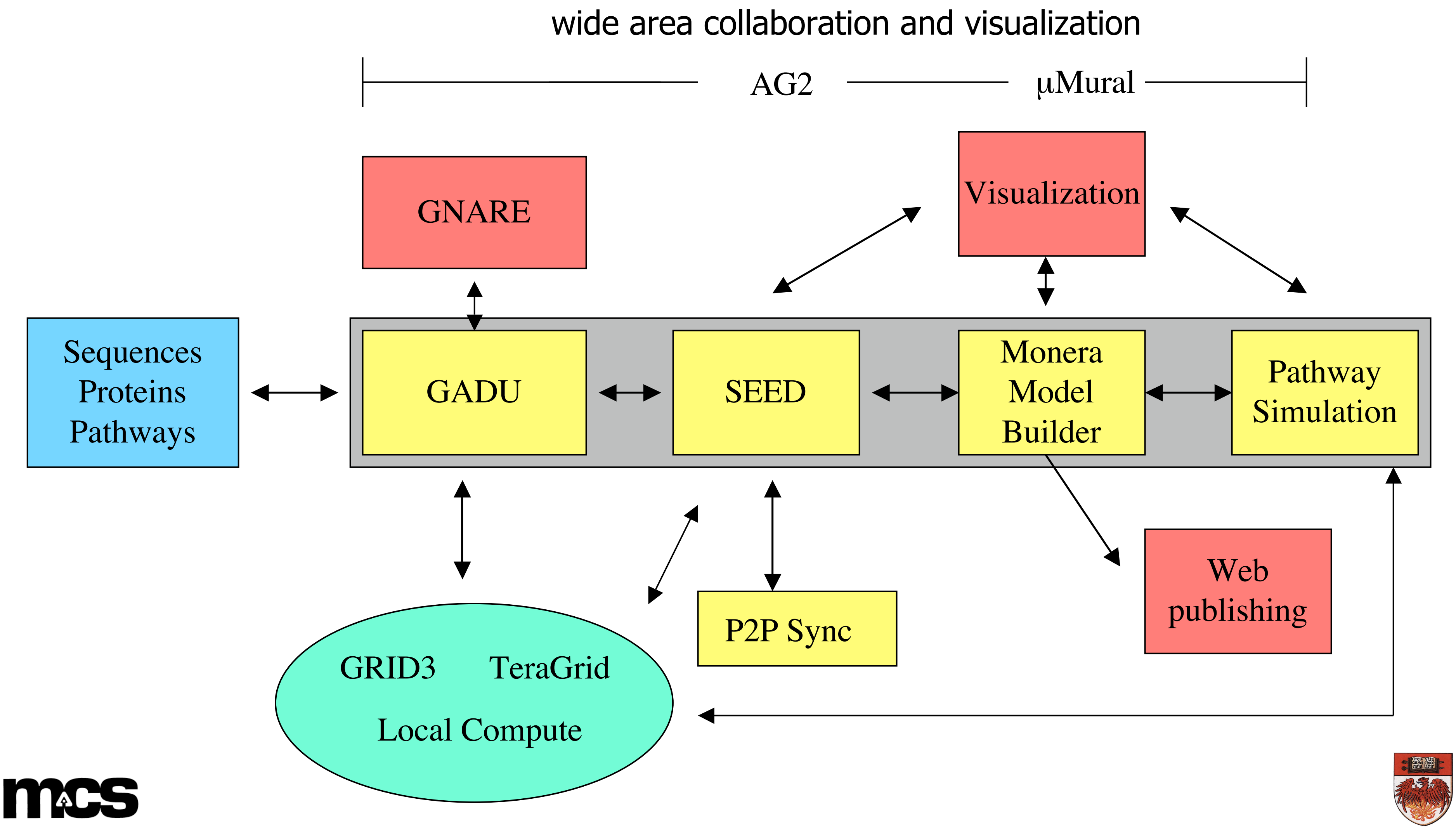


A Little Mystery?

Fig. 1. Correlation between microbial genome size and the number of predicted coding DNA sequences CDS. Bacterial genomes predicted to be undergoing reductive evolution are indicated by open circles, whereas other genomes are indicated by filled circles. The *N. equitans* genome is marked by "x". (*Inset*) An expansion of the data from small microbial genomes with the abscissa shown in genome size units of kbp.

Argonne Systems Biology Workflow





wide area collaboration and visualization

AG2

μMural

GNARE

Visualization

Sequences
Proteins
Pathways

GADU

SEED

Monera
Model
Builder

Pathway
Simulation

P2P Sync

Web
publishing

GRID3 TeraGrid
Local Compute

The SEED: Peer-to-Peer Software for Distributed Curation of Biological Data

- Support community-wide annotations and analysis of biological data
- Maintain an up-to-date collection of publicly-available datasets within the SEED framework
- Peer-to-peer synchronization is ideally suited for community-wide annotation of data collections
- Enable the SEED to be gracefully extended both via plug-in modules, but also through new forms of data integration

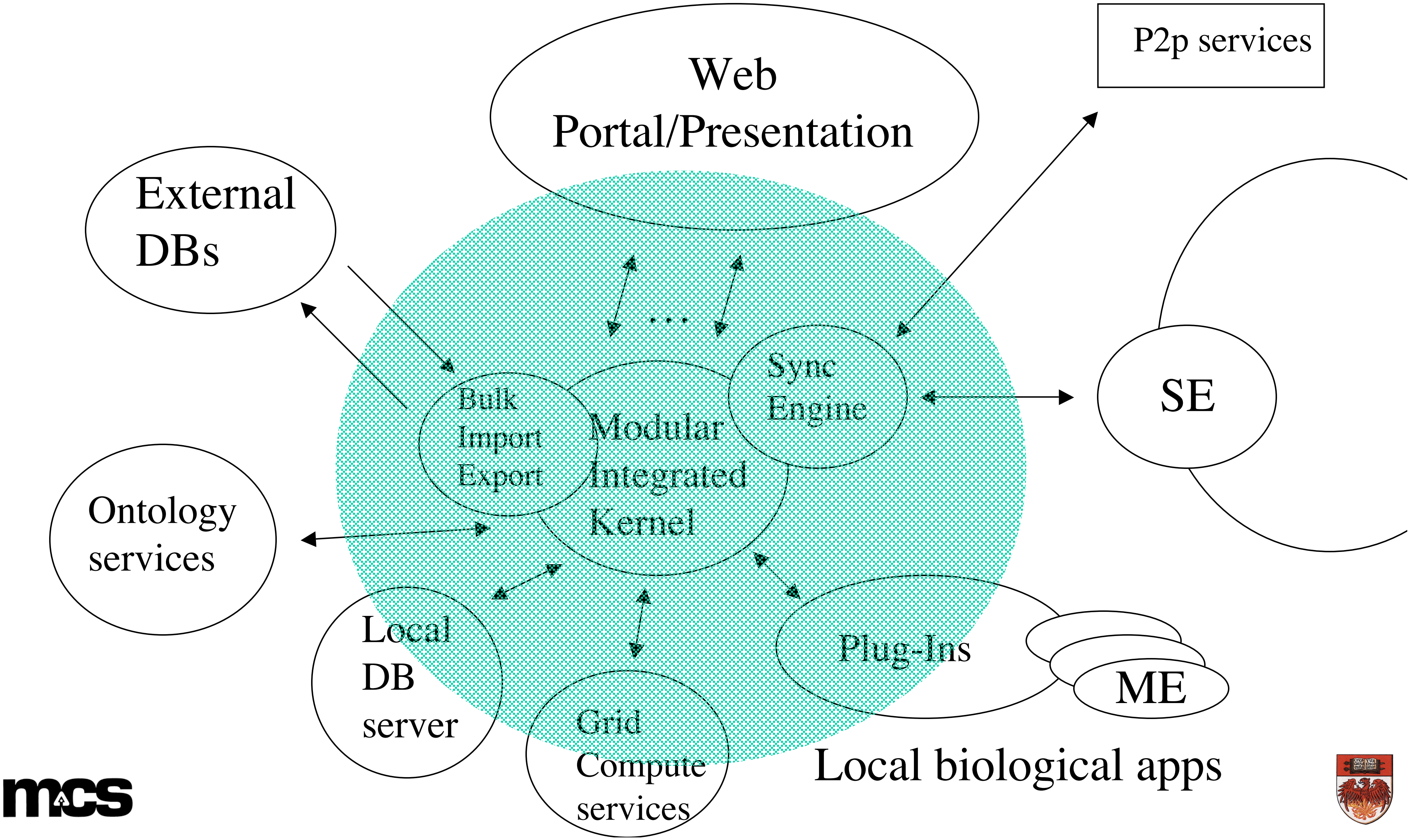
What do users need to be able to do?

- Install and share new genomes
  - Publish to a limited set of colleagues
- Share gene function assignments
  - Locally in a collaboration and remotely
- Share gene annotations (notes)
  - Can be arbitrary annoations (xml, html,etc.)
- Share naming rules (translations)
  - Publish dictionaries (ontologies) as views
- Lightweight code update
  - Propogate code to peers to continously update the algorithms

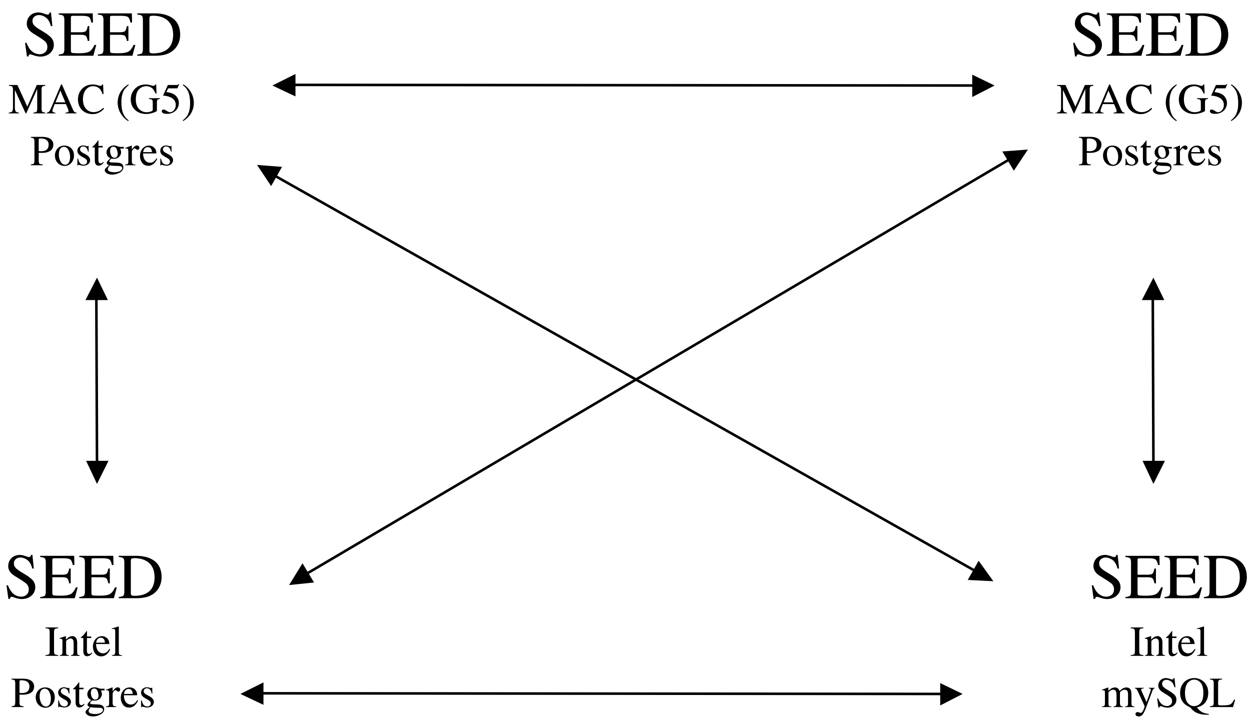
- Environment update (heavyweight, tools etc.)
- Clone the system for teams and peers

Peer-to-Peer Open Life Sciences Grid

the prototype SEED



Peer-to-Peer Demonstration at SC'03



SEED
MAC (G5)
Postgres

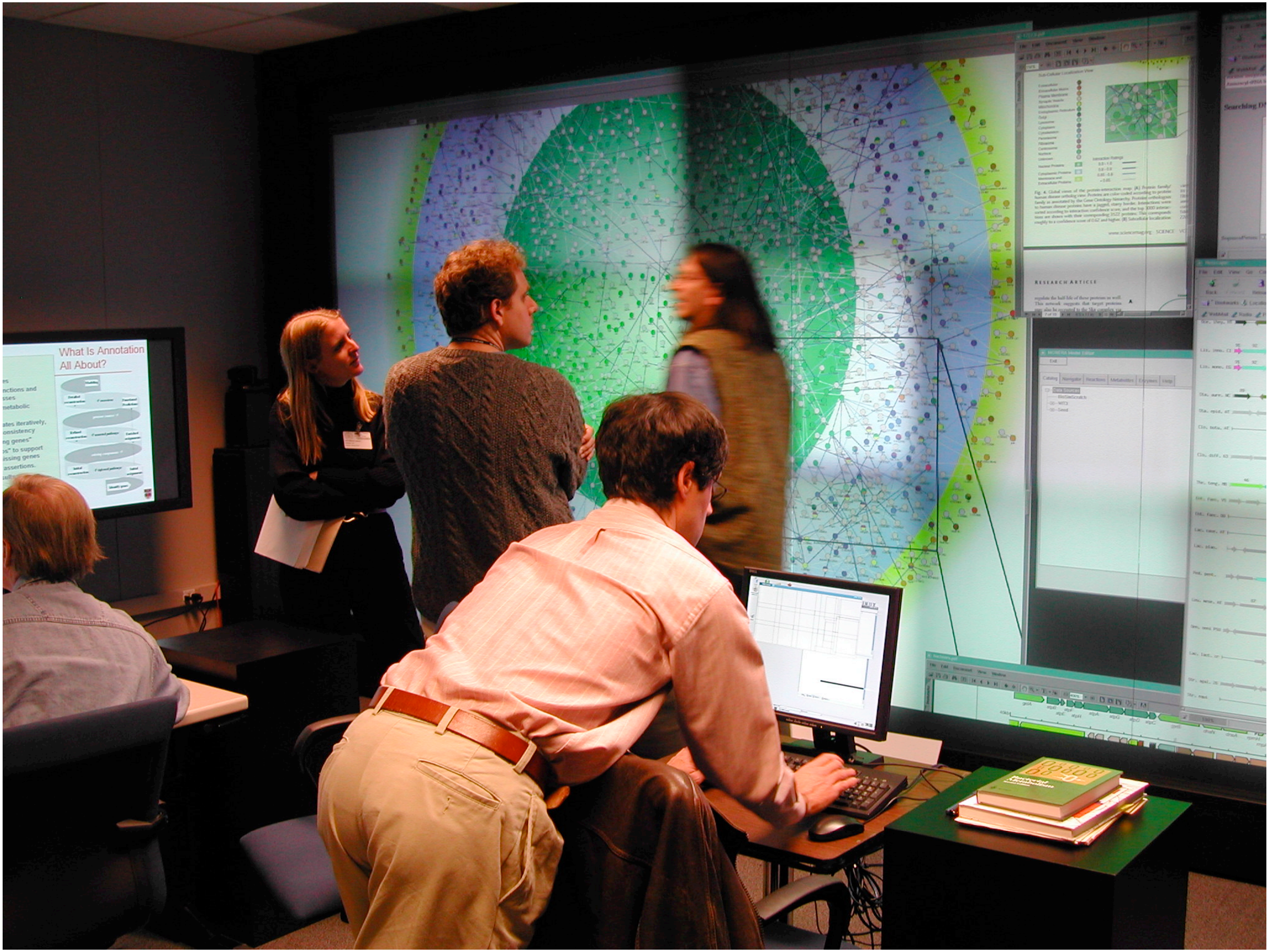
SEED
MAC (G5)
Postgres

SEED
Intel
Postgres

SEED
Intel
mySQL

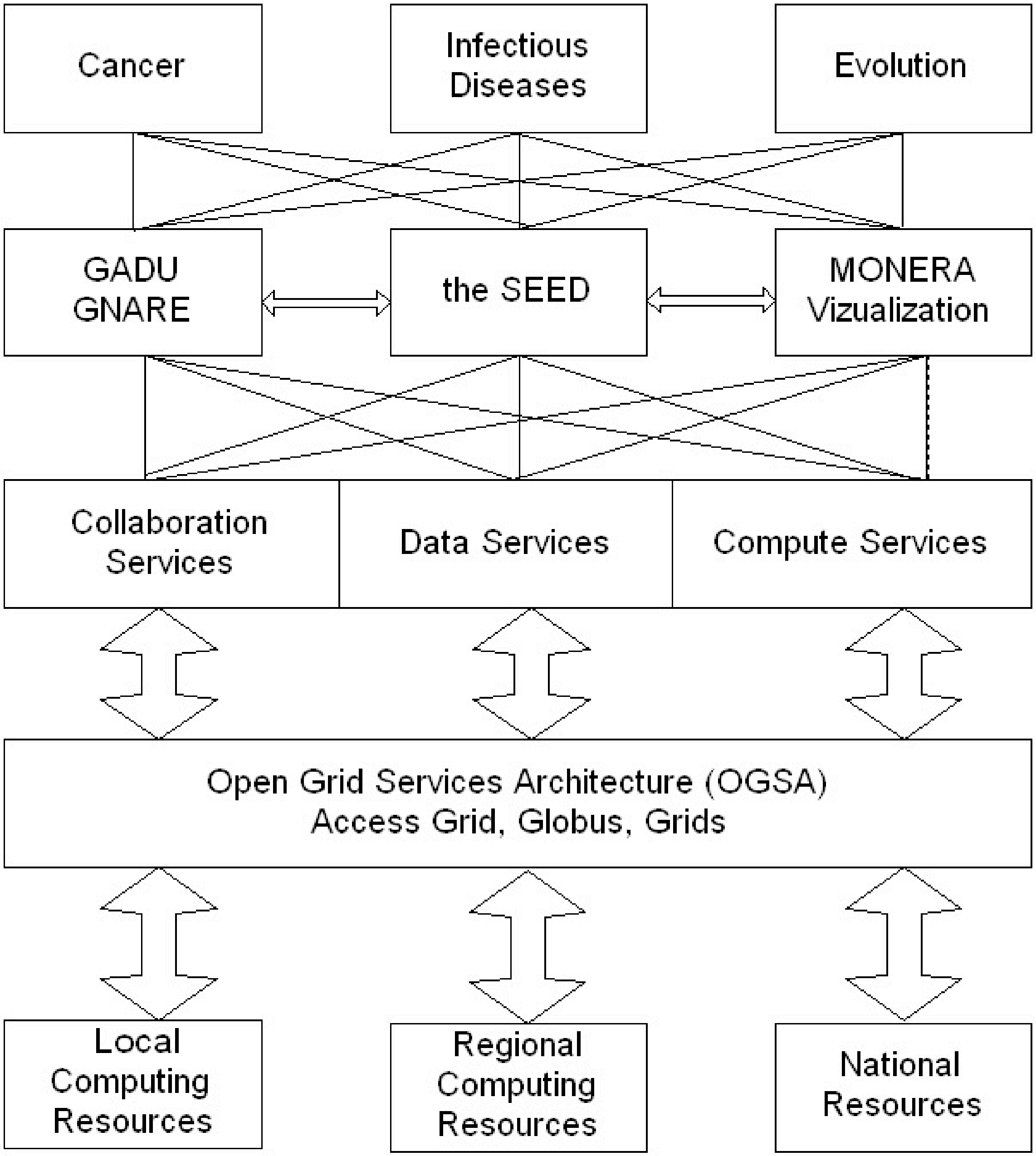
Example Biological Use Cases

- Comparative analysis of gene clusters
- Looking for missing genes
- Comparing pathways between organisms
- Reconstructing core cellular machinery
- Extracting rules for model development
- Visualization of whole cell networks
- Studying horizontal gene transfer
- Studying evolution of metabolism


What Is Annotation All About?
Research Article

Current SEED release

- Contains 33 archaeal, 391 bacterial, 480 eukaryotic, and 1177 viral genomes
- Of these, 24 archaeal, 189 bacterial, and 8 eukaryotic genomes are more-or-less complete
- 1.9 million entries in our non redundant database
- We are just now adding the ~1M environmental sequences recently deposited in Genbank
- In use in several large-scale annotation projects
- Basis for a system for curating microbial pathogens funded by NIH
- First SEED collaboration meeting was held in La Jolla at the end of February.
- Next SEED collaboration meeting will be held in Chicago in May, all serious groups are welcome to attend and get involved.



Cancer | Infectious Diseases | Evolution

Driving Biological Problems

GADU GNARE | the SEED | MONERA Vizualization

Integrated Bioinformatics Tools

Collaboration Services | Data Services | Compute Services

BioGrid Services

Open Grid Services Architecture (OGSA)
Access Grid, Globus, Grids

Standard Grid Software

Local Computing Resources | Regional Computing Resources | National Resources

Distributed Computing Resources

Near Term Futures for SEED Project

- We are starting up a project aimed at gearing up to annotate a 1000 genomes within 3 years
- We expect most of these will be microbial, but we are adding capabilities for Eukaroyotes as fast as we can
- We now have about ~20 installations up and running with p2p synchronization working at a basic level
- We have launched an effort to support detailed annotations of gene clusters (core metabolism, conserved translation core, replication, motility, etc.)
- We are actively seeking development collaborators

Core BioGrid Services

- Collaboration
- Data
- Compute

BioGrid Services - Collaboration

- Based on our 10+ years of work on collaboration tools and collaborative environments
- Core services are:
  - Event channels to provide application synchronization
  - Venues for sharing and storing state
  - Real-time media (audio, video, text)
  - Access to Grid services

BioGrid Services - Data

- Two primary needs
  - Peer-to-Peer data updates and synchronization
    - Updates to local databases (annotations, genomes, etc.)
    - Updates to local codebases (new functionality)
    - Support of data access restrictions/rights
  - Access to large-scale shared data resources
    - Databases too large for large-scale replication
      - Microarray data, similarity matrices, imaging, mass-spec
    - Real-time data feeds from instruments
      - High-throughput data
- Proposed data model for large-scale access
  - Attribute-value pairs (via scripting interfaces PERL, Python)

BioGrid Services - Compute

- Two Primary needs
  - Simple scripting interface to Grid resources
    - "Grid shell" enabling users to move existing pipelines to Grids
    - Prototype grid shell
      - Map grid resources into hiearchical tree structures
  - Tools for workflow management
    - Enabling complex workflows to be scheduled against Grid resources
      - Moving from simple pipes to complex flows with human-in-loop
    - Virtualization of compute, data and collaboration (UI) resources
      - Enabling access to grid resources via simple scripting tools

Some Comments on Grid Futures!

- Grid Services ≥ Web Services
  - at least for large-scale science applications
- Two big open problems
  - Simple Grid development environment for scientific end users (e.g. Grid version of csh)
  - Peer-to-Peer environments as Grid environments
    - Enable ad hoc deployment by end users
- Metric for Grid progress needs to be # of deployed production applications
  - Compare to other non-grid platforms (linux, mac, windows)