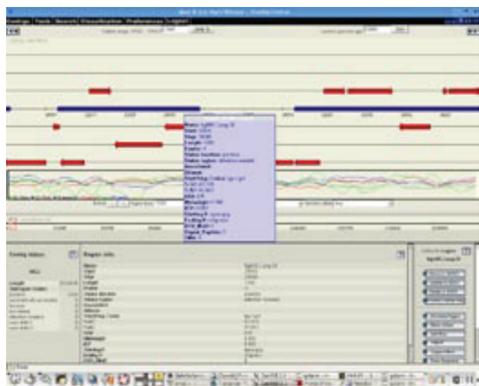


**Computational biology – including bioinformatics, high-throughput genome analysis, and metabolic reconstruction – has emerged as a major discipline in the 21st century, complementing experimentation. At Argonne, we are pursuing several avenues in computational biology. The centerpiece of our efforts is the SEED, an open source system created by the Fellowship for Interpretation of Genomes, Argonne, and the University of Chicago.**



### SUBSYSTEM-BASED METHODOLOGY FOR COMPARATIVE ANALYSIS

**Building effective computing systems that support petascale computing is a daunting challenge:**

The SEED designers believe that key to the development of high-throughput annotation technology is to have experts annotate single subsystems over the complete collection of genomes. In contrast, most annotation teams analyze a whole genome at a time. SEED researchers have developed detailed encodings of subsystems that make up the core cellular machinery.

These initial subsystems can be used to enhance sets of curated protein families and develop a consistent set of annotations.

The objective is to understand the evolutionary history of the genes within the subsystem. The encoded subsystems offer a detailed picture of what components have been identified and are present in each genome. Equally significant, they display exactly what is missing or ambiguous.

< The SEED has been integrated with GenDB, an annotation system that supports analysis of DNA. The SEED/GenDB data has been released, including an initial set of 100-150 subsystems.

### GRID-BASED GENOME ANNOTATION

Almost 300 genomes have been sequenced, and genomes of more than 1,600 organisms are at various levels of completion. To exploit the enormous scientific value of this information, we have developed GADU – the Genome Analysis and Database Update system.

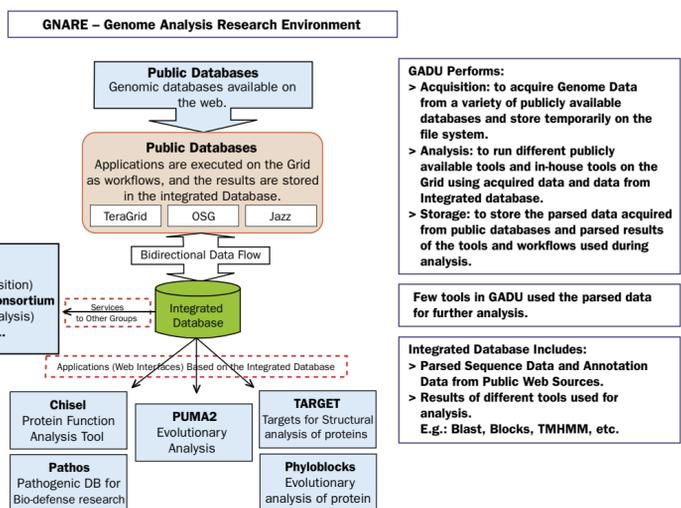
GADU provides an integrated database and a computational backend for data-driven bioinformatics applications.

It automates the major steps of genome analysis: acquisition of data from a variety of genomic databases, analysis by a diverse set of tools and algorithms, and storage of the results of analyses and annotations.

< The GADU/Gnare system is a high-performance, scalable computational pipeline for Grid-based genome annotation. The system, created to support the PUMA2 project, has been enhanced to service a variety of systems such as the SEED, PathosD, and MetaGenome.

High-throughput computations for genome analysis are performed over a large collection of distributed heterogeneous computing resources such as Grid3, TeraGrid, and DOE Science Grid.

[www.mcs.anl.gov/computational-bio/](http://www.mcs.anl.gov/computational-bio/)



# COMPUTATIONAL BIOLOGY: THE SCIENCE OF THE 21ST CENTURY

### SEMI-AUTOMATED TRANSFORMATION OF RECONSTRUCTIONS

One objective of the SEED project is to enable semi-automated reconstruction of prokaryotic biochemistry. An informal metabolic reconstruction comprises a set of encoded subsystems with a set of assignments for an organism.

Argonne is exploring two activities: extending the MONERA/KAH framework so that one can compare manually produced reaction networks with models produced with semi-automatic reaction-building tools; and producing ChemDB, a group-curated nonredundant chemistry database to support the semi-automated generation of stoichiometric matrices.

Moving from such an informal to a formal metabolic reconstruction is time-consuming, often taking an expert a year or more. To minimize this effort,

> Large metabolic reconstructions can be displayed on Argonne's μMural.



### SUPPORTING RESEARCH IN BIODEFENSE

The National Microbial Pathogen Data Resource Center and the SEED environments have a shared heritage and share many tools. In general, however, the SEED has more genomes and more tools for comparative analysis, whereas the NMPDR has a more refined set of genomes and more thorough curation. The NMPDR supports research in biodefense, emerging infectious diseases, and re-emerging pathogens.

NMPDR researchers are developing an integrated database providing a single Web-based entry point to all relevant organism-related data. The database will be used, for example, for identifying potential targets for the development of vaccines, therapeutics, and diagnostics. Argonne is providing the advanced bioinformatics environment for the NMPDR. This environment is used to explicate the physiology of the pathogens, to clarify the detailed variations that determine phenotype, and to develop consistent interpretations of functional data.

< Streptococcus pneumoniae, which causes pneumonia, meningitis, and osteomyelitis (among other diseases), is one of the pathogens targeted by the NMPDR.

### EVOLUTIONARY ANALYSIS OF ENZYMATIC FUNCTIONS

The availability of large volumes of sequence and enzymatic data for taxonomically and phenotypically diverse organisms now allows for systematic exploration of adaptive mechanisms that led to diversification of enzymatic functions. The system Chisel, developed at Argonne, provides the framework for such analyses.

of analysis of evolutionary versions of enzymatic functions for 1,157 organisms; (2) a hierarchical clustering algorithm for classifying enzymatic sequences in functional categories and a library of corresponding HMM profiles; and (3) an interactive tool based on a library of Chisel HMM profiles for classifying unannotated sequences.

Chisel includes (1) an integrated Enzymatic Knowledge Base containing sequence data and annotations, enzymatic and metabolic data from the public resources, as well as the results

Chisel is currently used in several large-scale projects including those with applications in metagenomes, biodefense, and bioremediation.

