

# GNARE—a grid-based server for the analysis of user submitted genomes

Dinanath Sulakhe<sup>1</sup>, Mark D'Souza<sup>1</sup>, Mustafa Syed<sup>2</sup>, Alexis Rodriguez<sup>2</sup>,  
Yi Zhang<sup>3</sup>, Elizabeth M. Glass<sup>1,2</sup>, Margaret F. Romine<sup>4</sup> and Natalia Maltsev<sup>1,2,\*</sup>

<sup>1</sup>Computation Institute, University of Chicago, Chicago, IL 60637, <sup>2</sup>Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, <sup>3</sup>University of Illinois at Chicago, Chicago, IL 60607 and <sup>4</sup>Pacific Northwest National Lab, Richland, WA 99352, USA

Received January 31, 2007; Revised April 17, 2007; Accepted April 25, 2007

## ABSTRACT

**GeName Analysis Research Environment (GNARE) is a bioinformatics server that supports both automated and interactive expert-driven analysis of user-submitted genomes and metagenomes. These analyses include gene function prediction and development of organism-specific metabolic reconstructions from sequence data. GNARE provides a framework for comparative and evolutionary analysis as well as annotation of genomes and metabolic networks in the context of phenotypic and taxonomic information. Results of analyses and metabolic models are visualized and extensively annotated with information from public databases. GNARE uses automated workflows and a Grid-based computational backend to perform high-throughput analysis of genomes. This use of distributed computing allows the analysis of an average-sized prokaryotic genome in less than 5h. GNARE is available at <http://compbio.mcs.anl.gov/gnare/>.**

## INTRODUCTION

In the past 10 years the amount of data in genomic databases has doubled or tripled every 2 years. In order for researchers to take advantage of the vast scientific value of this information for understanding biological systems, the information must be integrated, analyzed and modeled computationally in a timely fashion. The development of predictive computational models of an organism's functionality is essential for the progress of such fields as medicine, biotechnology and bioremediation. These models allow for the prediction of protein functions in newly sequenced genomes, as well as the existence of particular metabolic pathways and regulatory networks. Conjectures developed during genome analysis provide

invaluable assistance to researchers in experimental planning and help conserve time and resources required for characterizing an organism's biochemical and physiological properties. Essential for fulfilling this task is the development of high-throughput computational environments that integrate (i) large amounts of genomic and experimental data; (ii) comprehensive tools for knowledge discovery and data mining and (iii) comprehensive user interfaces that provide tools for easy access, navigation, visualization and annotation of biological information.

## Motivation

A significant number of sequencing projects (often for sequencing a single genome) and the initial interpretation of genomes are conducted by universities and small sequencing facilities. These may not have the desire or sufficient resources to develop the highly integrated and scalable bioinformatics system required for the interpretation of newly sequenced genomes. This time and resource consuming task is best performed by large bioinformatics centers. To address the needs of these groups, as well as groups interested in analysis of a particular organism or organisms (e.g. *Shewanella Federation*, and various biodefense and metagenome projects), we have developed a Web-based public server, GeName Analysis Research Environment (GNARE), which allows scientific groups and individual users to analyze genomic data using an integrated and automated bioinformatics environment based on advanced computational technologies.

GNARE leverages the following bioinformatics systems and analytical tools being developed by our team:

- (i) The PUMA2 system (1) for high-throughput genetic sequence and evolutionary analyses of genomes and metabolic reconstructions from sequence data. This system contains precomputed analyses of over 1000 publicly available genomes and automated metabolic reconstructions for over 330 organisms.

\*To whom correspondence should be addressed: Email: [sulakhe@mcs.anl.gov](mailto:sulakhe@mcs.anl.gov)

PUMA2 has been used by the scientific community worldwide.

- (ii) Tools for high-resolution comparative and evolutionary analysis of genomes developed by our group [e.g. PUMA2 function prediction tool, Chisel (<http://compbio.mcs.anl.gov/CHISEL>), a workbench for identification of taxonomic and phenotypic variations of enzymes, tools for comparative analysis of metabolic pathways].
- (iii) Genome Analysis and Database Update system (GADU) (2,3), an automated scalable computational pipeline for data acquisition and analysis by a variety of bioinformatics tools. GADU utilizes Grid resources for high-throughput computations.

## ANALYSIS OF GENOMES IN GNARE

The major steps of automated analysis of genomes in GNARE include: (a) assignment of functions to gene products, (b) development of metabolic reconstructions from sequence data and (c) comparative analysis by a variety of bioinformatics tools in the framework of phenotypic and taxonomic information. The results of these automated analyses may be further refined interactively by users. GNARE supports both public and private curation of genomes (prokaryotic as well as eukaryotic) and metabolic models by individual users as well as groups of experts.

### Steps of genome analysis in GNARE

*Step 1. Submission of genomes.* Users can submit sequence data for analysis via the Web interface or FTP. Currently GNARE accepts translated ORFs in FASTA format from genomes of all taxonomies. Prediction of ORFs is a non-trivial problem, especially for eukaryotic genomes, and a number of tools are publicly available for this purpose [e.g. Critica (4), Glimmer3 (5), GeneMarkS (6), GeneMark.hmm (7)]. The next release of GNARE will provide a choice of ORF prediction tools to the users. Predicted ORFs will be translated and analyzed as described later.

*Step 2. Assignment of function to gene products.* GNARE uses a voting algorithm and Chisel developed by our group to assign potential protein functions. These tools utilize the results of bioinformatics tools such as BLAST (8), Blocks (9) and InterPro (10,11) that are computed with distributed Grid resources using GADU.

The results of gene function prediction and results of the tools are presented in an interactive interface that supports user curation of every protein sequence annotation. Users can also perform further analysis of the sequence with over 30 bioinformatics tools to enhance their ability to produce accurate gene function predictions.

*Step 3. Metabolic reconstructions from sequence data.* Identification of gene functions allows reconstruction of metabolic networks that potentially exist in the organism. These metabolic reconstructions [EMP (12), WIT2 (13), KEGG (14), Ecocyc (15) and PUMA2]

have proved useful for developing organism and process-specific functional models of metabolic networks.

GNARE leverages the PUMA2 knowledge base for the development of automated metabolic reconstructions from sequence data provided by the user. GNARE also supports interactive development and annotation of metabolic models. The developed metabolic reconstructions are based on pathway data from the EMP collection of enzymes and metabolic pathways as well as the KEGG database.

*Step 4: Comparative analysis of genomes in the framework of taxonomic and phenotypic information.* The comparative and evolutionary analysis of genomes and metabolic networks in the framework of phenotypic and taxonomic information provides an organism-centric, systems-level view of genomic data. It allows the user to trace the evolutionary history associated with entire biochemical pathways and biological processes, rather than of individual genes, and to reconstruct the evolutionary progression of organisms that possess these pathways. It also enables identification, analysis and characterization of evolutionary patterns associated with particular phenotypes or phylogenetic neighborhoods. GNARE uses phenotypic data (e.g. habitat, carbon and energy source, pathogenicity and temperature optimum) from NCBI and manually extracted from the literature.

GNARE uses the following technologies developed by our group to support comparative analysis of user-submitted genomes:

- (i) Chisel, an integrated computational workbench for identification and characterization of enzymatic sequences.
- (ii) PhyloBlocks (<http://compbio.mcs.anl.gov/ulrich/phyloblock/>), a tool for phylogenetic analysis and interactive development of consensus sequences and HMM profiles from the sets of homologs selected by the user.
- (iii) Tools for comparative analysis of metabolic networks. These allow users to compare pathways of interest in their organism with those of over 397 other organisms from PUMA2. The user is presented with a graphical representation of the pathway as well as a spreadsheet showing the enzymatic content. Users can then identify missing enzymes and search for possible candidates.

## SYSTEM ARCHITECTURE

### Interface for user interactions

All the transactions in GNARE are user specific and can be accessed only by the user through login. A login-based authentication model is used for validating all the users and their transactions on the GNARE server. The user-submitted input data, the analysis performed by the user and the results are stored in a user-specific space. In addition, some users have opened their genomes for public access and annotation, which are available with a guest login.

Users can upload protein sequence data in the form of a FASTA file through the Web interface, or they can provide an FTP path to the input data. The interface is in the form of a Web form from where the user can select the bioinformatics tools to be run for the analysis of the uploaded sequence data. Currently GNARE supports Grid-based analysis using BLAST, Blocks and Chisel, this set of tools will be expanded in the near future. The user can monitor the progress of the analysis as these tools are being executed by the Workflow Executor. After Grid-based analysis of the data is completed, it is further analyzed by the PUMA2 gene function prediction algorithm. The predicted functions, Chisel results and user annotations are used for the development of an automated metabolic reconstruction. The results of all analyses are visualized and presented to the user through the Web interface. GNARE uses features of the PUMA2 system for representation of the analysis results. Various templates are used from the framework of PUMA2, such as the protein page representing BLAST and Blocks results, metabolic reconstruction and comparison of metabolic pathways.

### Integrated user database

Efficient comparative analysis in GNARE is supported by integrated knowledge base consisting of two parts:

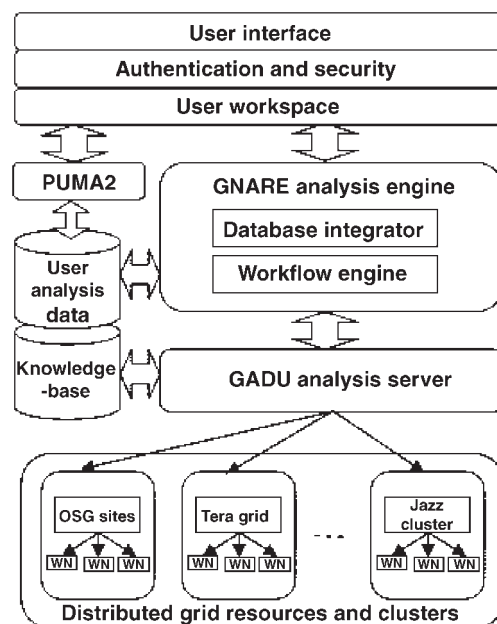
- (i) A relational integrated PUMA2 database containing vast amounts of genomic, metabolic, enzymatic, taxonomic and phenotypic information obtained from over 20 public databases and the precomputed results of analyses by various bioinformatics analysis tools (e.g. BLAST, Blocks and InterProScan) for over 1000 genomes.
- (ii) A relational database for user-submitted sequence data, results of analyses by the bioinformatics tools, metabolic reconstruction, functional predictions and user annotations.

### GNARE analysis engine

The GNARE analysis engine executes the analysis steps previously described for the interpretation of user-submitted genomes (Figure 1). The analysis engine has two modules: a database integrator and a workflow engine for user-submitted data. The database integrator stores the sequence data submitted by the user and the data generated by the tools at various levels of analysis into the relational user database.

The workflow engine, GADU, executes the various analysis tools in the specified order, taking care of dependencies. GADU is an automated, scalable, high-throughput computational workflow engine that enables the Grid execution of bioinformatics tools. It acts as a gateway to the Grid, handling all computational analyses for the GNARE system.

GADU has access to thousands of CPUs from various large-scale Grid resources such as the Open Science Grid (OSG) (<http://www.opensciencegrid.org>) and TeraGrid (<http://www.teragrid.org>). GADU's flexible architecture makes it simple to use Grid resources having different



**Figure 1.** Components of the GNARE architecture and their interactions. Users input sequence data via the secure web interface and a workspace for that project is created. Sequences are submitted to the GNARE Analysis Engine, which utilizes GADU for access to high-throughput computational resources. These resources include Open Science Grid, TeraGrid and local clusters. Data and results derived from computational analysis in GNARE is then integrated into the knowledge-base and displayed to the user. Users can then view, analyze and annotate their genome.

architectures (3), such as 32-bit processors of OSG and 64-bit processors of TeraGrid, and can easily add new resources to its pool. GADU can run its jobs on stand-alone clusters as well, such as the Jazz cluster (<http://www.lcrcl.anl.gov/jazz/index.php>) at Argonne National Laboratory. GADU uses the Grid resources on an opportunistic basis with no reservations for resources, whereas using shared local clusters requires reservations for faster results.

## RESULTS

GNARE has been used for analysis of 11 *Shewanella* strains for the needs of Shewanella Federation, the Hanford site microbial community metagenome and genomes of pathogenic organisms for the NIH GLRCE for Biodefense and Emerging Infectious Disease Research Consortium.

The *Shewanella* Federation (DOE OBER) aims to characterize and model the biology of *Shewanella oneidensis* MR-1 as well as other members of this Genus. Genomes of 11 strains and an MR-1 plasmid of *Shewanella* (total 43839 protein sequences) were submitted to GNARE via the Web interface. This data was analyzed using the GNARE automated pipeline from which the genomes were annotated and automated metabolic reconstructions were developed (Figure 2). The analysis was performed on 1317 CPUs of distributed

Analysis of : *Shewanella\_oneidensis\_MR1*

From this page you can explore:

- the automated assignments of functions using the PUMA2 function prediction.
- algorithm analyses of sequences using CHISEL workbench (see the table below).

Follow the links to see the pre-computed results of analyses of individual sequences in PUMA2 framework

Metabolic Reconstructions	
Using Kegg Data	Using EMP Data

Prioritize Metabolic Reconstructions	
Using KEGG data	Using EMP data

Tools Selected		Status
► BLAST	-- PUMA2_FP	COMPLETED
	-- META_RECON	COMPLETED
► BLOCKS		COMPLETED
► CHISEL		COMPLETED

Download Annotations [Excel]	
Download All Annotations	
Download User Annotations	

**A**

Protein ID	User Annotation	PUMA2 Annotation	User EC prediction	PUMA2 EC prediction	Chisel EC prediction
SO_0001	flavodoxin protein, MioC	flavodoxin			1.6.5.5
SO_0011	DNA gyrase, B subunit, GyrB	DNA gyrase, subunit B	5.99.1.3	5.99.1.3	5.99.1.3
SO_0014	glycyl-tRNA synthetase, beta subunit, GlyS	glycyl-tRNA synthetase, beta subunit	6.1.1.14	6.1.1.14	6.1.1.14

**B**

Protein sequence from user-submitted genome: *Shewanella\_oneidensis\_MR1*, SO\_0001

User: Dina

Organism name assigned by user: *Shewanella\_oneidensis\_MR1*

puma2 suggested function: flavodoxin

Chromosomal Comparison

STRING

Similarity -- Global

BLAST vs. nr

Fasta3 vs. UniProt

Blocks-Blast

PhyloBlast

TCDB

Similarity -- Local

InterPro

Blocks

Protein families

E-value score legend < 1 e^-100

Sequence length (146 aa)

Transmembrane regions

INTERPRO

IPR001094

IPR008254

BLOCKS

IPB001094

IPB003097

BLAST vs. nr

24371601

113972264

78667834

111016513

**C**

Enter new annotation for SO\_0001

CONTIG	
LOCATION	complement(334..774)
TYPE	
GENE	SO_0001
GENE NAME	mioC
PRODUCT	flavodoxin protein, MioC
FUNCTION	
EC	1.6.5.5
LOCALIZATION	periplasma

**Figure 2.** Screenshots from GNARE showing sequence analysis results of *Shewanella oneidensis* MRI. GNARE provides users with: (A) suggested functional annotations, (B) a suite of tools for interactive analysis and annotation by the user. The annotation fields (C) are comprehensive and enable users to compile information and predictions for various protein features, besides that of protein function.

computational resources from OSG and TeraGrid which took about 15h.

Similar analysis was performed for the Hanford site metagenome. An essential part of this analysis was identification of taxonomic variations of enzymes using Chisel. Such analysis allows developing or refining suggestions regarding the taxonomy of the organisms found in the metagenome and attributing particular enzymatic steps to specific taxonomic groups.

Functional and metabolic models were created to support research projects for the NIH GLRCE for Biodefense and Emerging Infectious Disease Research consortium. *Bacillus anthracis* strains were analyzed using GNARE for the identification and characterization of essential genes. These essential gene

candidates were then used in therapeutic inhibition studies.

## CONCLUSIONS AND FUTURE PLANS

The availability of new genomic DNA sequences and tools for identifying and predicting functions encoded therein are growing rapidly (16). High-throughput computing and data integration technologies are required to accommodate the growth of the data and increasing requirements for its analysis. The value of this information can then be effectively utilized for developing a systems level understanding of individual organisms and microbial communities.

A number of groups have developed servers for genome analysis such as: (a) The BaSys system (17) has the capability to annotate bacterial genomic DNA sequences. The annotation is automated, without any input from the user, and takes about 24h for an average bacterial genome; (b) IMG (18) allows analysis of public genomes and genomes available through the JGI sequencing facility as well as set of genes provided by the users; (c) The NMPDR (19) has developed a prototype of the 48h genome annotation server that performs integration of the user-submitted genomes into the SEED framework (20). GNARE contains a number of features that distinguish it from the other systems. Such features include unique set of interactive tools for the comparative analysis of sequences and metabolic networks in the framework of taxonomic and phenotypic information and the development of hierarchical metabolic reconstructions for the metagenomic data. It provides a platform where the user can perform annotation based on expert knowledge.

GNARE provides users with automated annotations of protein sequences and metabolic reconstructions for an average-sized prokaryotic genome in less than 5 h. It uses multiple grid resources simultaneously for faster analysis of the data. In the future, GNARE will also accept nucleotide sequences as input for the analysis. We plan to allow annotation of genomes with additional classes of information submitted by the users (e.g. gene expression data, biochemical and enzyme kinetic data), increase the number of Grid-supported bioinformatics tools for high-throughput analysis of data, implement Web-services to access additional tools and a Web portal to provide efficient collaborative environment.

## AVAILABILITY

GNARE is a public system available at <http://compbio.mcs.anl.gov/gnare>. Users can access GNARE with a guest login to browse the precomputed results. In order to perform high-throughput analysis user registration is required. Contact the administrator at [gnare@mcs.anl.gov](mailto:gnare@mcs.anl.gov) for further details.

## ACKNOWLEDGEMENTS

All authors acknowledge support by UChicago Argonne, LLC, as Operator of Argonne National Laboratory ('Argonne'), Argonne, a U.S. Department of Energy Office of Science Laboratory, is operated under Contract No. DE-AC02-06CH11357. N. Maltsev, E. M. Glass and M. Syed acknowledge membership within and support in part and D. Sulakhe in full from the Region V 'Great Lakes' Regional Center of Excellence in Biodefense and Emerging Infectious Diseases Consortium (GLRCE, National Institute of Allergy and Infectious Diseases Award 1-U54-AI-057153). M. D'Souza acknowledges membership and support to NMPDR Bioinformatics Resource Center NIH/NIAID (Award NNSN 266200400042C).

The GNARE development team is grateful for the help, advice and invaluable contributions from the ANL Globus group, especially Mike Wilde, Veronika Nefedova and Ian Foster; Eduard Dragut; former members of the team, Luke Ulrich, Jason Ting and Tanuja Bompada. Funding to pay the Open Access publication charges for this article was provided by XXXX.

*Conflict of interest statement.* None declared.

## REFERENCES

- Maltsev,N., Glass,E., Sulakhe,D., Rodriguez,A., Syed,M.H., Bompada,T., Zhang,Y. and D'Souza,M. (2006) PUMA2 – Grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res.*, **34**, D369–372.
- Rodriguez,A., Sulakhe,D., Marland,E., Nefedova,V., Yu,G.X. and Maltsev,N. (2003) GADU - Genome Analysis and Database Update Pipeline. *ANL/MCS-P1029-0203*.
- Sulakhe,D., Rodriguez,A., D'Souza,M., Wilde,M., Nefedova,V., Foster,I. and Maltsev,N. (2005) Gnare: automated system for high-throughput genome analysis with grid computational backend. *J. Clin. Monit. Comput.*, **19**, 4-5, 361–369.
- Badger,J.H. and Olsen,G.J. (1999) CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.*, **16**, 512–524.
- Delcher,A.L., Bratke,K.A., Powers,E.C. and Salzberg,S.L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679.
- Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.
- Lomsadze,A., Ter-Hovhannisyanyan,V., Chernoff,Y. and Borodovsky,M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.*, **33**, 6494–6506.
- Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the blocks database servers. *Nucl. Acids Res.*, **28**, 228–230.
- Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–205.
- Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.* Jul 1, 33(Web Server issue), W116–120.
- Selkov,E., Basmanova,S., Gaasterland,T., Goryanin,I., Gretchkin,Y., Maltsev,N., Nenashev,V., Overbeek,R., Panyushkina,E. *et al.* (1996) The metabolic pathway collection from EMP: the enzymes and metabolic pathways database. *Nucleic Acids Res.*, **24**, 26–28.
- Overbeek,R., Larsen,N., Pusch,G.D., D'Souza,M., Selkov,E.Jr, Kyrpides,N., Fonstein,M., Maltsev,N. and Selkov,E. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, **28**, 123–125.
- Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–280.
- Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,K. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.*, **33**, D334–337.
- Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2005) Genbank. *Nucleic Acids Res.*, **33**, 34–38.

17. Van Domselaar,G.H., Stothard,P., Shrivastava,S., Cruz,J.A., Guo,A., Dong,X., Lu,P., Szafron,D., Greiner,R. *et al.* (2005) BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res.*, **1**, **33**(Web Server issue), W455–459.
18. Markowitz,V.M., Korzeniewski,F., Palaniappan,K., Szeto,E., Werner,G., Padki,A., Zhao,X., Dubchak,I., Hugenholtz,P. *et al.* (2006) The integrated microbial genomes (IMG) system. *Nucleic Acids Res.*, **34**(Database issue), D344–D348.
19. McNeil,L.K., Reich,C., Aziz,R.K., Bartels,D., Cohoon,M., Disz,T., Edwards,R.A., Gerdes,S., Hwang,K. *et al.* (2007) The National Microbial Pathogen Database Resource (NMPDR): a genomics platform based on subsystem annotation. *Nucleic Acids Res.*, **35**(Database issue), D347–353.
20. Overbeek,R., Begley,T., Butler,R.M., Choudhuri,J.V. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**, 5691–5702.