Nucleic Acids Research, 2006, Vol. 34, Database issue ***D369–D372***
doi:10.1093/nar/gkj095

PUMA2—grid-based high-throughput analysis of genomes and metabolic pathways

Natalia Maltsev[1,2,]*, Elizabeth Glass[1,2], Dinanath Sulakhe[1], Alexis Rodriguez[1], Mustafa H. Syed[1], Tanuja Bompada[1], Yi Zhang[3] and Mark D'Souza[2]

[1]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA, [2]Computation Institute, University of Chicago, Chicago, IL 60637, USA and [3]University of Illinois at Chicago, Chicago, IL 60607, USA

Received August 15, 2005; Revised and Accepted October 16, 2005

ABSTRACT

The PUMA2 system (available at http://compbio.mcs.anl.gov/puma2) is an interactive, integrated bioinformatics environment for high-throughput genetic sequence analysis and metabolic reconstructions from sequence data. PUMA2 provides a framework for comparative and evolutionary analysis of genomic data and metabolic networks in the context of taxonomic and phenotypic information. Grid infrastructure is used to perform computationally intensive tasks. PUMA2 currently contains precomputed analysis of 213 prokaryotic, 22 eukaryotic, 650 mitochondrial and 1493 viral genomes and automated metabolic reconstructions for >200 organisms. Genomic data is annotated with information integrated from >20 sequence, structural and metabolic databases and ontologies. PUMA2 supports both automated and interactive expert-driven annotation of genomes, using a variety of publicly available bioinformatics tools. It also contains a suite of unique PUMA2 tools for automated assignment of gene function, evolutionary analysis of protein families and comparative analysis of metabolic pathways. PUMA2 allows users to submit batch sequence data for automated functional analysis and construction of metabolic models. The results of these analyses are made available to the users in the PUMA2 environment for further interactive sequence analysis and annotation.

INTRODUCTION

Evolutionary analysis of a wide spectrum of diverse organisms is essential for understanding how they adapt to environments. Common ancestry of eukaryotes and prokaryotes leads to similarity of many molecular functions. However, differences in organisms' structural complexity, physiology and lifestyle result in divergent evolution and emergence of variants of molecular function, metabolic organization and phenotypic features. Recent progress in genomics, bioinformatics and physiological studies allows for systematic exploration of adaptive mechanisms that led to diversification of biological systems. Such adaptive changes usually are not limited to one component of the system; on the contrary, in the process of adaptation, organisms undergo co-adaptive changes, such as the complementary changes of protein sequences to accommodate changes in an enzyme's active site or co-evolution of properties of different steps in metabolic pathways.

PUMA2, available at http://compbio.mcs.anl.gov/puma2, provides an environment for studying co-evolution of genomes and metabolic pathways and enzymes. It also supports the comparative analysis of sequence data and metabolic pathways for identification, analysis and characterization of evolutionary patterns associated with particular phylogenetic neighborhoods or phenotypes. The system enables high-throughput automated analysis of genomes, development of metabolic reconstructions from sequence data and community curation of genomic and metabolic data (Figure 1).

A number of excellent resources such as KEGG (1), MetaCyc (2) and IMG (3) support high-throughput analysis of genomes and metabolic reconstructions. Although PUMA2 has numerous commonalities with these systems, it offers a number of unique features. In brief PUMA2 (i) supports the interactive development of user models for public and user-submitted genomes; (ii) utilizes Grid technology for computationally intensive tasks; and (iii) provides new tools for comparative analysis of genomes and metabolic networks in the framework of taxonomic and phenotypic information. These last two features represent major differences between PUMA2 and its predecessor, the WIT2 system (4).

In addition to genome analysis, PUMA2 supports comprehensive annotation of genomes. The PUMA2 knowledge base integrates information from >20 public databases, including

*To whom correspondence should be addressed: Tel: +1 630 252 5195; Fax: +1 630 252 5986; Email: maltsev@mcs.anl.gov
Correspondence may also be addressed to Mark D'Souza. Tel: +1 630 252 7137; Fax: +1 630 252 5986; Email: dsouza@mcs.anl.gov

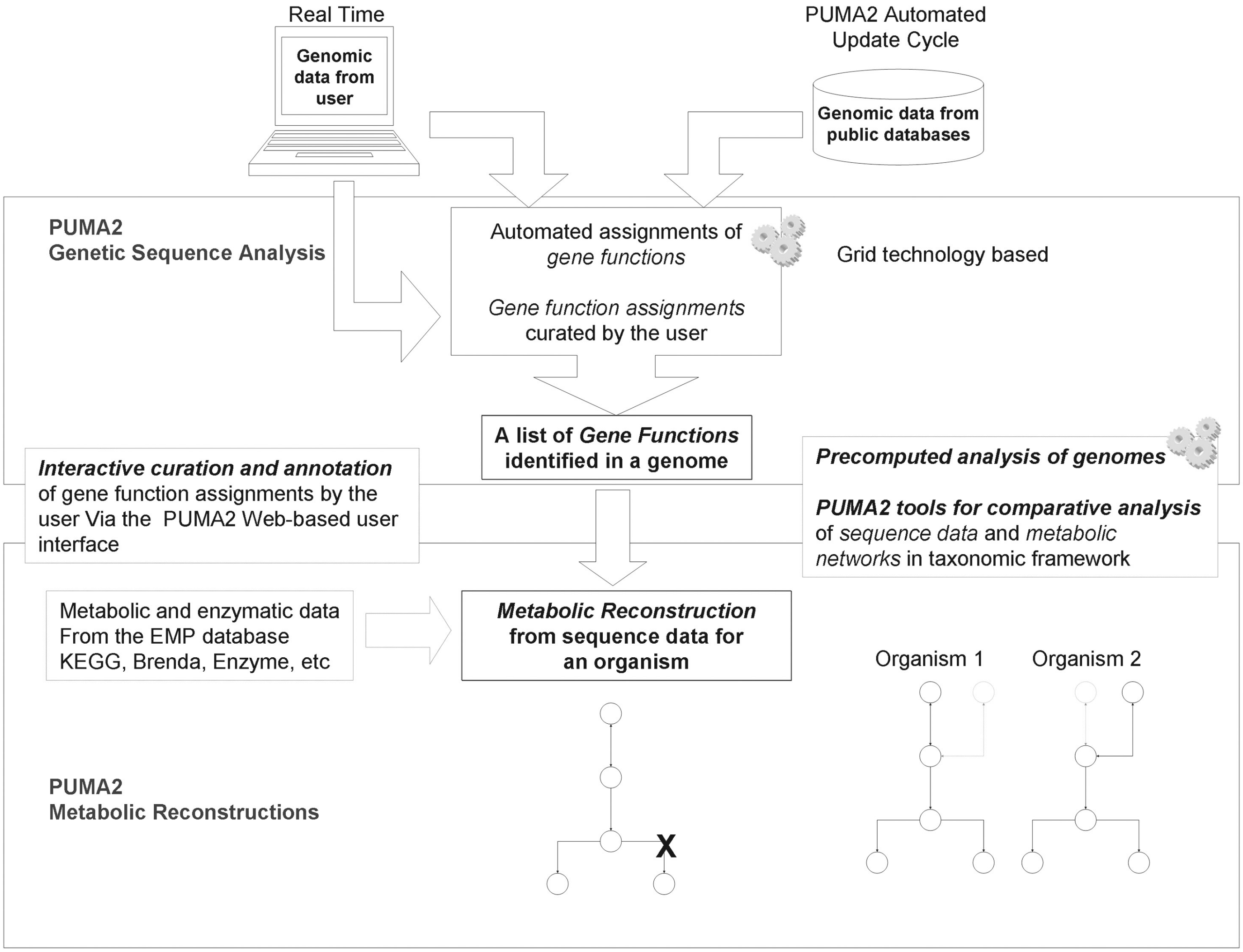
Real Time
Genomic data from user
PUMA2 Automated Update Cycle
Genomic data from public databases
PUMA2 Genetic Sequence Analysis
Automated assignments of *gene functions*
Gene function assignments curated by the user
Grid technology based
A list of *Gene Functions* identified in a genome
Interactive curation and annotation of gene function assignments by the user Via the PUMA2 Web-based user interface
Precomputed analysis of genomes
PUMA2 tools for comparative analysis of *sequence data* and *metabolic networks* in taxonomic framework
Metabolic and enzymatic data From the EMP database KEGG, Brenda, Enzyme, etc
Metabolic Reconstruction from sequence data for an organism
Organism 1
Organism 2
X
PUMA2 Metabolic Reconstructions

Figure 1. Representation of the stages of analyses of genomes in PUMA2. Users can browse data that is integrated and analyzed in PUMA2 or submit their own sequences. Precomputed results for homology, domain architecture, functional analysis and metabolic reconstructions are provided in an interactive framework.

sequence information and annotations [e.g. NCBI (5), PIR (6) and UniProt (7)], structural information [e.g. PDB (8), CATH (9) and SCOP (10)], metabolic information [e.g. EMP (11), KEGG and ENZYME (12)], taxonomic information from NCBI, gene ontologies from Gene Ontology (GO) (13) and physiological information [e.g. NCBI, TIGR (14) and the literature]. Extensive cross-referencing facilitates easy navigation of the data in PUMA2. The sections below describe major capabilities of the PUMA2 system in more detail.

AUTOMATED HIGH-THROUGHPUT ANALYSIS OF GENOMES IN PUMA2

Currently, the PUMA2 system contains automated precomputed analysis of 213 prokaryotic, 22 eukaryotic, 650 mitochondrial and 1493 viral genomes and automated metabolic reconstructions for >200 organisms. All data in PUMA2 are periodically updated. Sequence data in PUMA2 are obtained from the public sequence databases or provided by the users. It is analyzed by a variety of bioinformatics tools [e.g. BLAST (15), Blocks (16), Pfam (17), PepStats (18) and TMHMM (19)], as well as PUMA2 tools for prediction of gene functions and evolutionary analysis of enzymatic functions [e.g. Chisel and PhyloBlocks]. Grid technology is used to perform computationally intensive tasks. The results of these precomputed analyses are integrated into the database and presented to the users. PUMA2 tools for gene function prediction utilize the results of precomputed analyses of genomic data using BLAST, as well as InterPro, Blocks and TMHMM, to perform rules-based classification of the un-annotated sequences. PUMA2 also supports interactive analysis of sequences by the users, providing access to >40 publicly available bioinformatics tools.

PUMA2 environment for data curation by users

Although automated analysis may provide useful initial annotation of the genomes, these results require rigorous curation by expert biologists. PUMA2 provides registered users with tools for user-driven metabolic model development, reassignment of functions to genes and addition of comments. The user annotations in PUMA2 are persistent and are presented to the user when logged into the system.

PUMA2 sequence analysis tools

PUMA2 contains a suite of tools to facilitate the identification of evolutionary patterns and motifs characteristic of particular biological functions and their variations. The suite includes the following:

(i) Chisel, a web-based computational workbench for evolutionary analysis of enzymatic sequences (http://compbio.mcs.anl.gov/CHISEL). Chisel utilizes information from the PUMA2 knowledge base to perform rules-based clustering and classification of annotated enzymatic sequences into functional categories. The resulting clusters are used for developing a library of Hidden Markov Models (HMM profiles) for particular enzymatic functions and (when possible) their taxonomic and phenotypic variations. These profiles are used by the classification tool for prediction of functions of hypothetical proteins.

(ii) PhyloBlocks, a tool that allows a user to develop high-resolution HMM profiles for particular protein family interactively (http://compbio.mcs.anl.gov/ulrich/phyloblock).

(iii) Tools for comparative analysis of enzymes and metabolic networks in phenotypic and taxonomic framework.

PUMA2 allows to perform analyses of the data that allow users to ask such questions as ‘What metabolic pathways are common between hyperthermophilic organisms that live in aquatic environment?’ and ‘What variations of L-lactate dehydrogenase is characteristic for Firmicutes?’.

Metabolic reconstructions from sequence data in PUMA2

The technology of metabolic reconstructions from sequence data (1,4,20,21) proved useful for developing organism and process-specific functional models. PUMA2 currently contains automated metabolic reconstructions from the sequence data for >200 completely sequenced organisms. The system supports the development of automated metabolic reconstructions, which provide an initial basis for the development of expert-curated models. The developed metabolic reconstructions are based on pathway data from the EMP collection of enzymes and metabolic pathways, being developed by the EMP Project Inc., containing enzymatic information and metabolic diagrams accumulated from the literature describing >3000 metabolic pathways in a structured, indexed and searchable form. PUMA2 provides tools for comparative analysis of metabolic networks that allow identification of variations of the metabolic pathways characteristic for particular organisms or taxonomic groups of organisms, identification of ‘missing’ enzymes and viewing of the pathway data in a larger context of hierarchy of biological processes. PUMA2 metabolic reconstructions provide links to sequence data and enzymatic data. Genomic data and metabolic models in PUMA2 are annotated with the GO terms and cross-referenced with BioPAX format (22). Such representation simplifies navigation and comparative analyses of metabolic networks.

PUMA2 use of grid technology for high-throughput analysis

High-throughput computations in PUMA2 are performed by an automated Genome Analysis and Database Update (GADU) server with a grid-based distributed computational backend (23). GADU was developed as a collaboration between the ANL Globus group and the bioinformatics group in the framework of the NSF NCSA alliance. It leverages experience and technology developed by the GriPhyn (Physics Grid) project. High-throughput computations in GADU are performed by using distributed heterogeneous grid computing resources such as Grid2003 (24), OSG and the DOE Science Grid. Periodic updates of the PUMA2 system and analysis of the sequence data using BLAST, Blocks, Pfam, Chisel and the like are performed in the form of scientific workflows expressed and controlled by a ‘virtual data’ model using the Chimera Virtual Data System (25), which transparently maps computational workflows to distributed grid resources.

Using the GADU system allows for automated genetic sequence analysis of an average bacterial genome (∼4000 protein sequences), development of automated metabolic reconstructions and integration of resulting models in the PUMA2 framework and its presentation to the user via a web interface in <3 h. Such analysis includes analysis of the genome by BLAST and Blocks, automated assignment of functions to genes and development of automated metabolic reconstructions.

PUMA2 support of genomes provided by users

Some users are interested in analyses of genomes that are not yet included in the PUMA2 framework. PUMA2 provides automated analyses of user-submitted sets of sequences. These sequence sets may be complete or incomplete genomes or sets of sequences of interest to the user. Limited amounts (up to 50 sequences) of user-submitted protein sequences may be analyzed via the PUMA2 website. Analysis of larger amounts of user-provided genomic data is available per request.

AVAILABILITY

PUMA2 is available for use via the web-based user interface at http://compbio.mcs.anl.gov/puma2. The following datasets are available on request: metabolic data from EMP database in BioPax (OWL) format and precomputed results from organism specific analyses by BLAST, Blocks and Chisel. Requests may be sent via email to puma2@mcs.anl.gov.

ACKNOWLEDGEMENTS

The PUMA2 development team is grateful for the help, advice and invaluable contributions from the ANL Globus group, especially Mike Wilde, Veronika Nefedova and Ian Foster; former members of the PUMA2 development team, Luke Ulrich, Jason Ting and Ramya Krishnamurthy; and Robert Petryszak from the European Bioinformatics Institute for making ClustR available for our use. The work was supported in part by the U.S. Department of Energy under Contract W-31-109-ENG-38.

Conflict of interest statement. None declared.

REFERENCES

1. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
2. Krieger,C.J., Zhang,P., Mueller,L.A., Wang,A., Paley,S., Arnaud,M., Pick,J., Rhee,S.Y. and Karp,P.D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **32**, D438–D442.
3. Markowitz,V., Korzeniewski,F., Palaniappan,K., Szeto,E., Werner,G., Padki,A., Zhao,X., Dubchak,I., Hugenholtz,P., Anderson,I. *et al.* (2006) The Integrated Microbial Genomes (IMG) System. *Nucleic Acids Res.*, in press.
4. 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.
5. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
6. Wu,C.H., Yeh,L.S., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al.* (2003) The Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347.
7. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
8. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
9. Pearl,F., Todd,A., Sillitoe,I., Dibley,M., Redfern,O., Lewis,T., Bennett,C., Marsden,R., Grant,A., Lee,D. *et al.* (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.*, **33**, D247–D251.
10. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
11. Selkov,E., Basmanova,S., Gaasterland,T., Goryanin,I., Gretchkin,Y., Maltsev,N., Nenashev,V., Overbeek,R., Panyushkina,E., Pronevitch,L. *et al.* (1996) The metabolic pathway collection from EMP: the enzymes and metabolic pathways database. *Nucleic Acids Res.*, **24**, 26–28.
12. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
13. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
14. Peterson,J.D., Umayam,L.A., Dickinson,T., Hickey,E.K. and White,O. (2001) The Comprehensive Microbial Resource. *Nucleic Acids Res.*, **29**, 123–125.
15. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
16. Henikoff,S., Henikoff,J.G. and Pietrokovski,S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471–479.
17. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
18. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
19. Moller,S., Croning,M.D.R. and Apweiler,R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–653.
20. Selkov,E., Maltsev,N., Olsen,G.J., Overbeek,R. and Whitman,W.B. (1997) A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data. *Gene*, **197**, GC11–GC26.
21. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.*, **33**, D334–D337.
22. Luciano,J.S. (2005) PAX of mind for pathway researchers. *Drug Discov. Today*, **10**, 937–942.
23. 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**, 77–168.
24. Foster,I., Gieraltowski,J., Gose,S., Maltsev,N., May,E., Rodriguez,A., Sulakhe,D., Vaniachine,A., Shank,J., Youssef,S. *et al.* (2004) The Grid 2003 Production Grid: Principles and Practice. HPDC, 13th IEEE International Symposium on High Performance Distributed Computing (HPDC-13 '04), 236–245.
25. Foster,I. (2005) Serivec-oriented science. *Science*, **308**, 814–817.