D390–D393 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj070

Megx.net—database resources for marine ecological genomics

Thierry Lombardot[1], Renzo Kottmann[1], Hauke Pfeffer[1], Michael Richter[1], Hanno Teeling[1], Christian Quast[1] and Frank Oliver Glöckner[1,2,*]

[1]Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany and [2]International University Bremen, D-28759 Bremen, Germany

Received August 8, 2005; Revised and Accepted October 8, 2005

ABSTRACT

Marine microbial genomics and metagenomics is an emerging field in environmental research. Since the completion of the first marine bacterial genome in 2003, the number of fully sequenced marine bacteria has grown rapidly. Concurrently, marine metagenomics studies are performed on a regular basis, and the resulting number of sequences is growing exponentially. To address environmentally relevant questions like organismal adaptations to oceanic provinces and regional differences in the microbial cycling of nutrients, it is necessary to couple sequence data with geographical information and supplement them with contextual information like physical, chemical and biological data. Therefore, new specialized databases are needed to organize and standardize data storage as well as centralize data access and interpretation. We introduce Megx.net, a set of databases and tools that handle genomic and metagenomic sequences in their environmental contexts. Megx.net includes (i) a geographic information system to systematically store and analyse marine genomic and metagenomic data in conjunction with contextual information; (ii) an environmental genome browser with fast search functionalities; (iii) a database with precomputed analyses for selected complete genomes; and (iv) a database and tool to classify metagenomic fragments based on oligonucleotide signatures. These integrative databases and webserver will help researchers to generate a better understanding of the functioning of marine ecosystems. All resources are freely accessible at http://www.megx.net.

INTRODUCTION

Over the last decade microbiology has undergone several changes. Robert Koch's invention of pure culture techniques at the end of the 19th century focussed microbiology on the isolation of bacteria for laboratory studies. In 1987 Carl Woese introduced the ribosomal RNA as a stable molecular marker for the classification and identification of microorganisms (1). The 'winds of change' blew in the field of microbiology (2) when the first cultivation-independent investigations reported an immense array of completely unexpected microbial diversity in the environment (3). The landmark publication of the first complete genome sequence of *Haemophilus influenzae* in 1995 (4) has transformed biology into a massively parallel and high throughput endeavour. This 'genomic revolution' finally reached the field of marine ecological genomics in the year 2000, defined as: 'The application of genomic sciences to understanding the structure and function of marine ecosystems' (5). Since 1995, >260 microbial genomes have been fully sequenced, and 600 more are well on their way (5). While most projects focus on microorganisms of medical or biotechnological interest, 22 complete marine genomes of environmental organisms are already available, and ~130 marine isolates are currently sequenced (Moore foundation http://www.moore.org). Recently, this cultivation-based approach has been complemented by a number of groundbreaking cultivation-independent—metagenomic—studies, the most prominent being the Venter Sargasso Sea expedition in 2004 (6), delivering >1.2 million new genes. This wealth of information caused a quantum leap in marine sciences and demands for different kinds of databases to transfer information into knowledge (7). The sequences, genomes, genes and predicted metabolic functions can not longer be regarded in an organism centric view but have to be handled in the context of the environment surrounding them. Therefore, it is necessary to link any environmental sequence information with its geographical location. This allows to correlate

*To whom correspondence should be addressed. Tel: +49 0421 2028938; Fax: +49 0421 2028580; Email: fog@mpi-bremen.de

© The Author 2006. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org

the genomic features found at a distinct sampling site with physical, chemical and biotic information to identify organism-specific adaptations and their role and impact on the environment. This new kind of integrative data resource opens the path to address questions like: Are there differences in the genetic repertoire when travelling from coastal marine sites to the open ocean? or Do habitat specific gene patterns with yet unknown functions exist? If the latter is true the correlation with site specific environmental parameters might allow predicting a potential function for them. Can these genetic properties in turn explain the distribution of the organisms?

Megx.net is designed to tackle these tasks linking marine genome and metagenome sequences not only with geography but providing additional information about annotation highlights, presence of environmentally relevant protein families and group-specific genes as well as a Geographic-BLAST server to trace genes across the marine environment.

SOURCES OF GENOMIC AND METAGENOMIC DATA

The genome sequences of all currently available marine microorganisms have been retrieved from the EMBL and GenBank databases (8,9). Twenty-two bacteria and archaea originating from the water column of the ocean and from marine sediments have been completely sequenced (October 2005). The sequences and associated gene annotation have been imported into a local relational database allowing fast data retrieval. The corresponding annotations originate from independent submissions to the EMBL or the GenBank databases, and are of variable quality owing to the following reasons: (i) the original annotations were performed at different times; (ii) no controlled vocabulary is used for gene product names; and (iii) the effort expended in assigning functions to genes is variable between genome projects. Ecologically relevant annotation highlights were selected from original genome publications for each organism.

Metagenomic fragments originating from marine systems have been selected according to semi-automatic literature screening. Seventy-eight original publications were found to deal with metagenomic fragment sequencing, corresponding to a total of 21 distinct marine geographic sampling sites (August 2005). The sequences and associated gene annotation were imported into a newly designed geographic database. New genomes or metagenomes will be integrated in the database and mapserver as soon as they become available. Precomputed searches will be updated every 2 months.

GENOME BROWSING

The genome browser allows easy and fast access to the sequences, their geographical location and the annotation highlights of each marine microorganism in the database. For example, the unexpected archaea-like C1 metabolism genes found in the genome of *Rhodopirellula baltica* can be accessed in their genomic context by a simple mouse click (Figure 1). Fast text search in the original annotations and BLAST searches are also available.

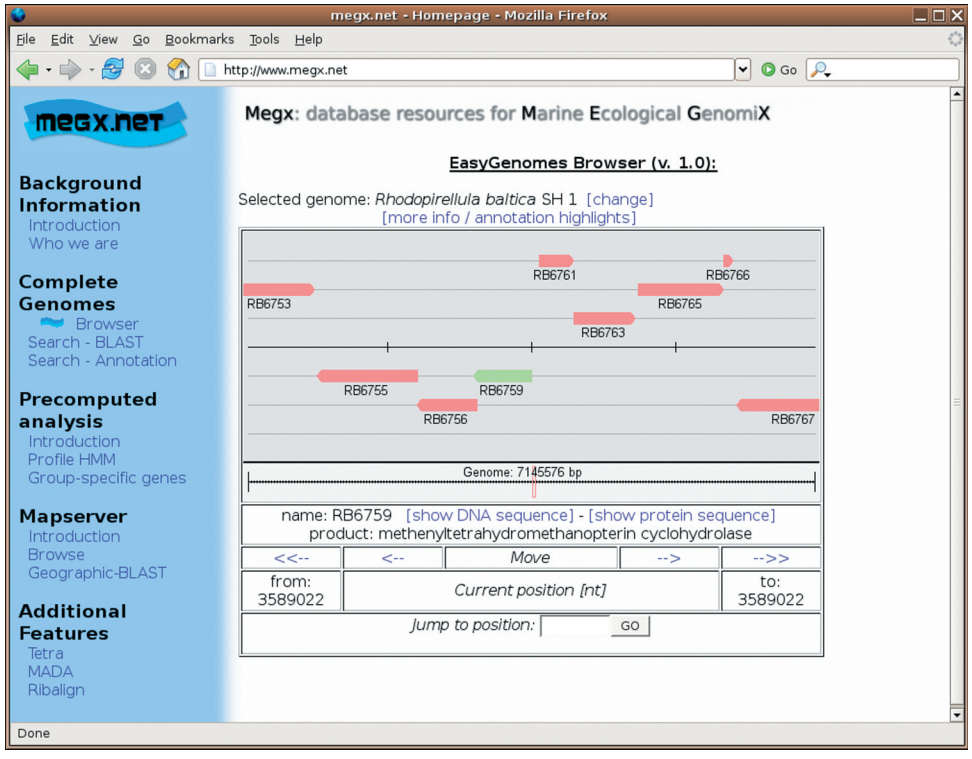


Figure 1. Fast access to the annotation highlights of marine microorganisms. Here, the archaea-like C1 metabolism key gene is *R.baltica*.

PRECOMPUTED INFORMATION

Environmentally relevant protein families

Some gene families are of particular interest for ecological genomics, as they play key roles in the environment or give insights into the adaptation of microorganisms to their respective niche. Glycosylhydrolases, sulphatases, peptidases and transcriptional regulators are some examples of gene groups that have been automatically extracted based on selected profile hidden Markov models originating from the Pfam database (10). The results can be browsed graphically on our web page. This search strategy allows consistent quantitative comparisons, as the publicly available original annotation can not easily be compared. For example, the outstanding number of genes encoding sulphatases in the genome of *R.baltica* (11) or the reduced dataset of transcriptional regulators in *Prochlorococcus marinus* strains (12,13) can be compared with the corresponding gene content of other marine microorganisms.

Group-specific genes

Group-specific genes are defined as those found exclusively in a defined subset of genomes. The definition of groups is variable and can be based on a phylogenetic affiliation, a common metabolism or related habitats. An example for group specific genes for phylogenetically closely related organisms are the three available *P.marinus* strains. The results show that some light-inducible proteins are exclusively found in those organisms (13). Moreover, we present a set of proteins of yet unknown function which are *P.marinus* specific. The corresponding genes represent interesting targets for functional genomics and further wet-lab experiments.

TETRA SERVER

TETRA is a software tool for genomic and metagenomic analysis. It can assess the relatedness of genomic fragments by computing correlations between their tetranucleotide usage patterns (i.e. statistical over- and under-representation of tetranucleotides) (14,15). The new version includes chaos game plot representations for DNA sequences, which can be used to get additional information on the relatedness of genomic fragments. Moreover, TETRA can plot fluctuations of tetranucleotide usage patterns within DNA sequences. This is particularly useful to identify irregular regions in entire genomes or larger genomic fragments like laterally transferred genes or transposase and phage insertions.

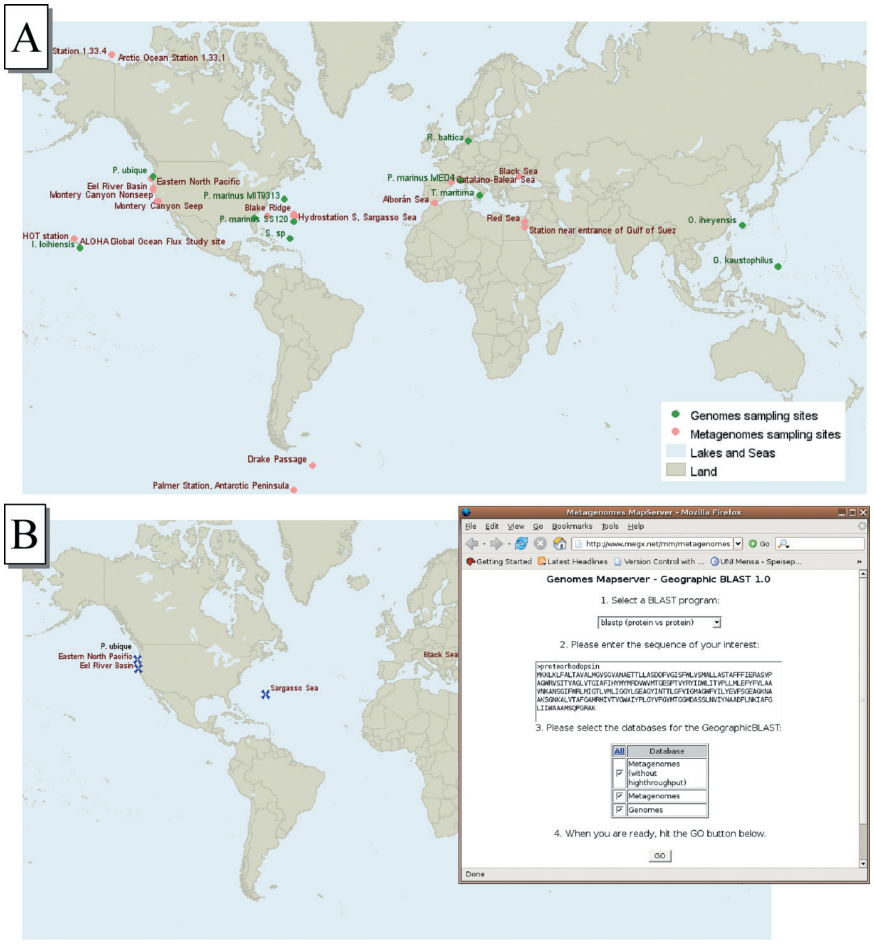


Figure 2. The Genomes Mapserver. (**A**) Marine genomes and metagenomic fragments can be browsed and searched on a world map on our web-based system. (**B**) An example showing a Geographic-BLAST search for genes encoding proteorhodopsins in the currently available dataset.

GENOMES MAPSERVER

Geographic information systems (GIS) are commonly used in the field of geology for data integration. A GIS is a combination of elements designed to store, retrieve, analyse and display geographic data. We introduce here the Genomes Mapserver, a GIS that allows access to genomic and metagenomic sequence data in their geographic and ecological contexts. The sampling sites of marine (meta)-genomic studies are displayed within a browsable world map (Figure 2). Each sampling site can be selected to display the corresponding sequences and additional contextual information. The underlying database is designed to enable future data mining tasks to reveal possible gene patterns associated with a particular environmental context. For targeted searches, a geographic-BLAST tool has been developed, allowing to perform 'spatial' queries for sequences based on the popular BLAST algorithm (16). The Geographic-BLAST/Genomes Mapserver combination allows to systematically study the biogeography of particular genes in the environment (Figure 2).

ADDITIONAL FEATURES

A software tool for microarray data evaluation and a database of aligned ribosomal proteins for phylogenetic analysis (Ribalign) will soon be available on the webpage.

DATABASES ACCESS

The precomputed genome searches and group-specific genes, the TETRA server and the Metagenomes Mapserver are freely available through http://www.megx.net.

ACKNOWLEDGEMENTS

We thank the Max Planck Society for initial funding and the EU Sixth Framework Programme (FP6-NEST) for providing financial support for further development of the Genomes Mapserver (contract no. 511784). Funding to pay the Open Access publication charges for this article was provided by the Max Planck Society.

Conflict of interest statement. None declared.

REFERENCES

1. Woese,C.R. (1987) Bacterial evolution. *Microbiol. Rev.*, **51**, 221–271.
2. Olsen,G.J., Woese,C.R. and Overbeek,R. (1994) The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.*, **176**, 1–6.
3. Torsvik,V., Goksoyr,J. and Daae,F.L. (1990) High diversity in DNA of soil bacteria. *Appl. Environ. Microbiol.*, **56**, 782–787.
4. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae*. *Science*, **269**, 496–512.
5. Cary,C. and Chisholm,P. (2000) *Report of a Workshop on Marine Microbial Genomics to Develop Recommendations for the National Science Foundation*. Arlington, VA.
6. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D.Y., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
7. DeLong,E.F. and Karl,D.M. (2005) Genomic perspectives in microbial oceaonography. *Nature Insight*, **437**, nature04157.
8. Kanz,C., Aldebert,P., Althorpe,N., Baker,W., Baldwin,A., Bates,K., Browne,P., van den Broek,A., Castro,M., Cochrane,G. *et al.* (2005) The EMBL nucleotide sequence database. *Nucleic Acid Res.*, **33**, D29–D33.
9. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2005) GenBank. *Nucleic Acid Res.*, **33**, D34–D38.
10. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acid Res.*, **32**, D138–D141.
11. Glöckner,F.O., Kube,M., Bauer,M., Teeling,H., Lombardot,T., Ludwig,W., Gade,D., Beck,A., Borzym,K., Heitmann,K. *et al.* (2003) Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc. Natl Acad. Sci. USA*, **100**, 8298–8303.
12. Dufresne,A., Salanoubat,M., Partensky,F., Artiguenave,F., Axmann,I.M., Barbe,V., Duprat,S., Galperin,M.Y., Koonin,E.V., Le Gall,F. *et al.* (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc. Natl Acad. Sci. USA*, **100**, 10020–10025.
13. Rocap,G., Larimer,F.W., Lamerdin,J., Malfatti,S., Chain,P., Ahlgren,N.A., Arellano,A., Coleman,M., Hauser,L., Hess,W.R. *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*, **424**, 1042–1047.
14. Teeling,H., Meyerdierks,A., Bauer,M., Amann,R. and Glöckner,F.O. (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.*, **6**, 938–947.
15. Teeling,H., Waldmann,J., Lombardot,T., Bauer,M. and Glöckner,F.O. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, **5**, 163.
16. 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.