Agenda of the 10th Genomic Standards Consortium Workshop

Venue:

Argonne National Laboratory
9700 S. Cass Avenue
Argonne, IL 60439, USA
Phone: 630/252-2000

This workshop is sponsored by:

DAY 1, Mon Oct 4th - for working group and Board meetings

10.00 - 12.00 ISA Infrastructure Workshop - All Welcome

The Investigation/Study/Assay (ISA) tools (www.isa-tools.org; Rocca-Serra et al, 2010; Bioinformatics paper) offer a way to capture MIGS/MIMS/MIENS compliant metadata as well as metadata describing a range of other types of investigations.

The ISA tools are freely available desktop software suite targeted to curators and experimentalists that:

* assists in the reporting and local management of experimental metadata (i.e. sample characteristics, technology and measurement types, sample-to-data relationships) from studies employing one or a combination of technologies;
* empowers users to uptake one or more community-defined minimum information checklists and ontologies, as required;
* formats studies for submission to a growing number of international public repositories endorsing the tools, currently ENA (genomics), PRIDE (proteomics) and ArrayExpress (transcriptomics).

The scope of this workshop is to provide an overview of the ISA Infrastructure tool kit and discuss how it can be used within the GSC to promote compliance with MIGS/MIMS/MIENS.

This meeting will be organized by Susanna-Assunta Sansone and Dawn Field. Please contact us if you are interested in attending.

12.00-13.00 Lunch

13.00 - 16.00 GSC Board Meeting - Closed

The GSC established a board in April 2009 whose members have been selected from representatives within the wider community most active in driving GSC activities. Membership in GSC standing committees is currently defined by participation. Please contact the relevant member(s) of the GSC Board if you are in a position to contribute to one more GSC governance activities. The full list of members can be seen here along with GSC standing committees.

GSC Governance: http://gensc.org/gc_wiki/index.php/GSC_Board

16.00 - 18.00 GSC Developer Meeting - All welcome

The GSC has recently formed a Developer's working group and GSC 10 will provide the forum for the first face-to-face meeting of this group. The role of this group is to push forward GSC projects, and GSC contributions to larger standardization projects (like MIBBI, ISA and the Environment Ontology), through technical discussions of the best solutions for implementation and work with Adopters towards implementation of GSC standards. This meeting will allow everyone to introduce themselves in person, give updates on their particular activities and will therefore set the stage for the rest of GSC 10. The focus of the meeting will be on the technical development of GSC projects, continuing links with the wider standardization community, but topics will also include:

- the formal collaboration with Digibio and the migration of GSC infrastructure to OSL
- launch of MTF - open discussion to wider group
- recruiting new Adopters
- developing ties between the Developer WG and the Board
- how to increase visibility of the GSC in the community?

This group has weekly telecons and membership is defined by participation and includes developers from a range of GSC projects and activities. Please consider joining.

Developer's Group: http://gensc.org/gc_wiki/index.php/Developer_Group

18.00 - 19.30 Registration and mixer with fingerfood

19.30 - 20.30 Keynote Talk of GSC 10

Microbial Earth - understanding diversity through genomics
Nikos Kyrpides (DOE Joint Genome Institute)

DAY 2, Tue Oct 5th

8.30-9.00am - Coffee and Registration

9.00 - 10.30 Session I: GSC welcome and project updates

Chair: Dawn Field (NERC CEH)

9.00 - 9.10 - Introduction to GSC
Dawn Field (NERC Centre for Ecology and Hydrology)

9.10 - 9.20 - Introduction to the GSC 10 theme - M5
Folker Meyer (Argonne National Laboratory)

9:20 - 9:30 - The MIGS/MIMS/MIENS family of standards
Peter Sterk (Genomic Standards Consortium)

9:30 - 9:40 The Genomic Contextual Data Markup Language (GCDML)
Renzo Kottmann (MPI-Bremen)

9:40 - 9:50 Genomic Rosetta Stone (GRS): How to register your database identifiers
Peter Dawyndt/Wim de Smet (University of Ghent)

9:50 - 10:00 Ontologies within the GSC (including Environment Ontology and the description of Habitat
Norman Morrison (University of Manchester)

10:00 - 10:10 Launching Digibio - Working with the GSC
Brian Bramlett (Lux Bio)

10:10 - 10:20 Update on M3/BioSharing SIG: launch of the BioSharing Community
Dawn Field (NERC CEH) and Susanna Sansone (University of Oxford)

10:20 - 10:30 Biodiversity working group update
Norman Morrison (University of Manchester)

10.30 - 11.00 Coffee break and Poster Session

11.00 - 13.00 Session II: Developers and Adopters

Chair: Nikos Kyrpides (DOE JGI)

11.00 - 11.15 Introducing the GSC Developers Group
Renzo Kottmann (MPI-Bremen) and Dawn Field (NERC CEH)

Flash Talks
Short introductions from the people involved, see:

http://gensc.org/gc_wiki/index.php/Developer_Group

11:15 - 11:30 Collecting MIGS/MIMS/MIENS on SMART Phones: the Epicollect system
Norman Morrison (University of Manchester) and David Aanensen (Imperial College, London)

11: 30 - 11:45 RDP - submission of MIENS compliant data
James Cole (Michigan State University)

11:45 - 12:00 ISA tools: MIGS/MIMS/MIENS compliance and use of ontologies
Philippe Rocca-Serra (University of Oxford)

12:00- 12:15 The RCN4GSC: an overview and call for new participants
Lynn Schriml (University of Maryland)

12:15 - 12:30 CAFAE - defining the role of the GSC
Owen White (University of Maryland)

12: 30 - 13:00 Open Discussion

13.00 - 14.00 Lunch

14.00 - 15.30 Session III: M5

Chair: Owen White (University of Maryland)

The Vision of M5

14.00 - 14.15 The need for computational standards in metagenomics - the vision of M5
Folker Meyer (Argonne National Laboratory)

14.15 - 14.30 An M5 Pilot Project: A proposed exchange format for metagenomics (MTF format)
Andreas Wilke (Argonne National Laboratory)

Building the M5 consensus technical platform

14.30 - 14.40 ELIXIR and the EBI vision for metagenomics
Sarah Hunter (EBI)

The promise of the cloud

14.40 - 14.50 CLOVR - putting pipelines in the cloud
Sam Anguioli (University of Maryland)

14:50 - 15.00 Parallelizing CLOVR in clouds and clusters with A.W.E.
Jared Wilkening (Argonne National Laboratory)

The promise of workflows

15.00 - 15.10 Workflows in a CAMERA context
Jeff Grethe (UCSD)

15.10 - 15.20 Kepler Workflows in a CAMERA context and beyond
Ilkay Altintas (UCSD)

15.20 - 15.30 Taverna, MyExperiment and workflows
Katy Wolstencraft (University of Manchester)

15.30-16.00 Coffee Break and Poster Session

16.00 - 17.30 Session IV: M5 and co-ordinated megasequencing projects

Chair Hans-Peter Klenk (DSMZ)

16.00 - 16.10 The Microbial Earth Project - towards sequencing 9000+ type strains
Nikos Kyrpides (DOE JGI)

16:10 - 16.20 Sloan indoor metagenomics meeting - towards understanding the microbes of man-made spaces
Folker Meyer (Argonne National Laboratory) and Owen White (University of Maryland)

16.20 - 16.30 A proposal for an Earth Microbiome Project - the outcome of the Terabase Metagenomics meeting
Jack Gilbert (Argonne National Laboratory)

16:30 - 16:40 BGI - the 1000 genomes project and beyond
TBA (BGI)

16.30 - 17.30 Discussion: Bringing it all together to build M5

- How do data and tools relate?
- What other standards are needed?
- Do we need a centralized GSC repository for metadata?
- Putting the M5 vision into a whitepaper

17.30 - 18.00 Technical platforms and Demonstrations (Parallel sessions)

17.30 - 18.00 A platform for bio-computing: Magellan
Narayan Desai (Argonne National Laboratory)

17.30 - 18.00 Standards compliance: ISA tools
Philippe Rocca-Serra and Eamonn Maguire (University of Oxford)

Dinner and Plenary talk by Oliver Ryder on Genomes 10k Project

8:30-9:00 Plenary Talk after dinner

The Genomes 10k Project - completing 10,000 Vertebrate Genomes
Oliver Ryder (San Diego Zoo, UCSD)

DAY 3, Wed Oct 6th

9.00 - 10.30 Session V: M5 - Building the roadmap

9:00-9:20 The promise of metadata for science
Rob Knight (University of Colorado)

9:30-10:30 M5 Open Discussion Continued

10.30 - 11.00 Coffee Break and Poster Session

11.00 - 13.00 Session VI: Break out groups

M5 - main auditorium
Folker Meyer (Argonne National Laboratory) and Sarah Hunter (EBI)

GSC Developers group
Renzo Kottmann (MPI-Bremen)

Biodiversity working group
Norman Morrison (University of Manchester)

CAFAE/GSC
Owen White (University of Maryland)

13.00 - 14.00 Lunch

14.00 - 15.30 - Wrap up: Final Review of Actions

Reviews of meeting, setting actions and thank yous

Discussion

* Summary of sessions and actions
* Contributions to GSC Special issue of SIGS: specific submissions
* GSC 10 meeting report: authors and content (top outcomes of meeting)
* Future meetings

Workshop Co-organizers and GSC Board

15.30 Formal close of workshop (Organizers) and Coffee

Poster abstracts

Gene-Centric Association Analysis for the Correlation between the Guanine-Cytosine Content Levels and Temperature Range Conditions of Prokaryotic Species

Hao Zheng and Hongwei Wu

School of Electrical and Computer Engineering, Georgia Institute of Technology

The environment has been playing an instrumental role in shaping and maintaining phenotypic and genotypic diversities of prokaryotes. It has been debatable whether the whole-genome Guanine-Cytosine (GC) content levels of prokaryotic organisms are correlated with their optimal growth temperatures. Since the GC content is variable within a genome, we here focus on the correlation between the genic GC content levels and the temperature range conditions of prokaryotic organisms. The GC content levels in the coding regions of four genes were consistently identified as correlated with the temperature range condition when the association analysis was applied to (i) the 722 mesophilic and 93 thermophilic/hyperthermophilic organisms regardless of their phylogeny, oxygen requirement, salinity, or habitat conditions, and (ii) partial lists of organisms when organisms with certain phylogeny, oxygen requirement, salinity or habitat conditions were excluded. These four genes are K01251 (adenosylhomocysteinase), K03724 (DNA repair and recombination proteins), K07588 (LAO/AO transport system kinase), and K09122 (hypothetical protein). To further validate the identified correlation relationships, we examined to what extent the temperature range condition of an organism can be predicted based on the GC content levels in the coding regions of the selected genes. The 84.52% accuracy for the complete genomes and the 84.09% accuracy for the in-progress genomes, especially when being compared to the 50% accuracy rendered by random guessing, suggested that the temperature range condition of a prokaryotic organism can generally be predicted based on the GC content levels of the selected genomic regions. The results rendered by various statistical tests and prediction tests indicated that the GC content levels of the coding/non-coding regions of certain genes are highly likely to be correlated with the temperature range conditions of prokaryotic organisms. Therefore, it is promising to carry out \reverse ecology" and to complete the ecological characterizations of prokaryotic organisms, i.e., to infer their temperature range conditions based on the GC content levels of certain genomic regions.

The global catalog of microbial material Straininfo

Bert Verslyppe and Peter Dawyndt

University of Ghent

StrainInfo (http://www.straininfo.net) is a global catalog of microbial material, building upon the catalogs of Biological Resource Centers (BRCs) by integrating catalog entries of equivalent microbial material. The adoption of Microbiological Common Language (MCL) XML synchronization quickly increased the volume of semantic information in StrainInfo. Semantic information is information corresponding with fine-grained, precisely defined fields such as for example isolation sample location and habitat, oxygen relationship and optimal growth temperature. As the effective data values of the different semantic fields entering StrainInfo are raw textual entries, are of varying detail, can have different forms or languages, and sometimes contain inconsistencies, they need to be converted to a semantic representation based on ontologies. Using a specialized semantic integration algorithm, these values then can be converted to a strain level consensus value for each field. As a case study, the focus was put on the isolation habitat and location fields. These strain level consensus values allow to use ontological knowledge when searching and therefore increase precision and recall compared to full text search. For example, it allows searching for all strains isolated in a given continent or in the neighborhood of a particular place, even if this additional information is not mentioned in the original catalog entries. The Environment Ontology can be used to immediately retrieve all different types of diary products when searching using the general term. In addition, the ontologies enrich the data by providing or linking additional information (e.g. GPS coordinates for geographical locations). The consensus values will be made available to end-users by displaying them on the corresponding strain passports. Geographical locations can be visualized on a map. Advanced search functionality will be made available to allow users to perform true semantic search based on ontologies. The integration results will also be available for electronic processing from the MCL XML exports. As the coverage and the quality of the system improves with the addition of more semantic information, we invite users administering datasets containing semantic information to consider making this information universally available to the complete microbiological community through StrainInfo!

Comparative metagenomic analysis of viral metagenomes from two methanotrophic communities

Blair Paul and David Valentine

Marine Science Institute, UC Santa Barbara

Methanotrophic microbial communities are estimated to consume 75–310 Tg of methane each year in the marine environment. Methanotrophic organisms span a vertical horizon from the anaerobic marine subsurface, dominated by archaea, to the aerobic water column, dominated by bacteria. While some methanotrophic bacteria have been isolated in pure culture to yield full genomes, no methanotrophic archaea have been, and we still know little about the diversity of marine methanotrophs and their viruses. We conducted a comparative metagenomic analysis of viral metagenomes from two methanotrophic communities, an aerobic benthic microbial mat and anaerobic sediment. Purification of viral DNA and pyrosequencing yielded over 400,000 sequences from the two metagenomes. Using MG-RAST and CAMERA, we show that both methanotrophic viromes encode host metabolic functions, but that the functions encoded differ between communities. While few viruses of Archaea have been studied to date, the anaerobic methanotrophic virome includes numerous sequence matches to Archaeal genes, indicating the ability for phage genomes to acquire auxiliary sequences from these hosts. The presence of host functions in both viromes suggests that phage-mediated horizontal gene transfer, a process common to other marine settings, also occurs in methanotrophic communities.

Accuracy of reading-frame calls on metagenomic fragments

William Trimble, Andreas Wilke, Kevin Keegan and Folker Meyer

Argonne National Laboratory

Models which can accurately predict the reading frame of metagenomic fragments have the potential to reduce the computational burden of similarity searches and metagenomic assembly efforts. We here test the accuracy of Markov models (like GeneMark) and FragGeneScan at guessing the reading frame of simulated and real metagenomic fragments whose reading frames are accurately known.

Integrating MIGS/MIMS into MG-RAST

Andreas Wilke [1,2] , Travis Harrison [1,2], Renzo Kottmann [4], Tobias Paczian [1,2], Elizabeth Glass [1,2], Folker Meyer [1,2,3]

[1] Argonne National Laboratory, Mathematics and Computer Science Division, Argonne, IL.
[2] Computation Institute, University of Chicago, Chicago, IL.
[3] Institute for Genomic and Systems Biology, Argonne National Laboratory, Argonne, IL.
[4] Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany.

Analysis of metagenomic data sets requires computational tools and infrastucture which enable sequence comparison, but also critical is related metadata. It is increasingly apparent that the full potential of comparative metagenomic analysis can be achieved only in the context of the metadata. Therefore, data describing such information as a sample's environment, sample origin, isolation, and treatment, as defined in MIGS/MIMS, are an important resource to link to sequence data in order to enable meaningful analysis and bioprospecting. MG-RAST v3.0 provides an environment for incorporating metadata into analyses. Here we describe the GSC-compliant MG-RAST Meta-Data Editor, integration and import/export functions.

New-Generation Metagenomic Analysis Environment: CAMERA 2.0 Workflow System

Shulei Sun, Jing Chen, Weizhong Li, Jeff Grethe, Ilkay Altinatas, Abel Lin, Steve Peltier, Karen Stocks, Eric Allen, Mark Ellisman and John Wooley

University of California San Diego

CAMERA 2.0, the latest release and a major upgrade for the environmental metagenomics data resource, covers primary sequence data and its associated metadata (1). Data sharing within the metagenomics community has always been a core focus of CAMERA. However, as the field of metagenomics expands it is becoming more important to not only share the underlying data itself but also the new and improved analysis methods being developed in the community. To that end, CAMERA 2.0 supports an enhanced and improved data sharing environment (for further details about managing the data challenge, see the poster: Managing the Primary Data and Metadata Challenges in Metagenomics) along with a new collaborative analysis environment that provides an extensible collection of bioinformatics tools and workflows to address the unique challenges of metagenomics and enabling researchers to collaborate in new ways through CAMERA.

CAMERA 2.0 organizes data analysis tools through a collaborative scientific workflow system. At the core of this environment is the Kepler scientific workflow system(2, 3, 4) that supports the integration of data and automated computational tools providing capabilities to record all steps in data the analysis process, i.e., provenance. This data-oriented view of an analysis enables communication (to collaborators) of what has occurred within a given workflow, and allows for the exchange and reproducibility of the computation itself. This capability makes sharing of data analyses between centers and researchers possible which is one of the main focuses of the "Metagenomics, Metadata, MetaAnalysis, Models and MetaInfrastructure" (M5) community. CAMERA is working with MGRAST (5), img/m (6), and other centers as a part of M5 to take data exchange and analysis sharing forward. Through the CAMERA portal, users can create, share, retrieve and run the processing workflows specific to their own experiment without having to install special software. In addition, this workflow-based environment reduces the cost of moving from a stand-alone scientific application to a workflow-based community resource by (i) designing and publishing workflows based on application services; (ii) executing workflows based on local or online data; (iii) saving and querying workflow results; (iv) saving and viewing data and process provenance; (v) creating ad-hoc collaborations and project spaces; and (vi) publishing uploaded workflows or sharing workflows with workspace members.

The core workflow system currently makes the following metagenomic analyses available to researchers: data quality control (specifically, QC Filter and 454 Duplicate Clustering), read assembly (454 Read Assembly), functional annotation and clustering (Metagenomic Data Annotation and Clustering), taxonomy binning (Taxonomy Binning), BLAST, and additional downstream analysis methods. The QC Filter takes FASTA and qual files or a FASTQ file as input, calculates the average score for each read, then fetches high quality reads, filters out reads shorter than the minimum read length, and generates a statistical analysis. The 454 Duplicate Clustering identifies exact and near identical duplicates to remove sequencing artifacts (7). The 454 Read Assembly first runs seven independent assembly programs to get a pool of contigs, then re-assembles these contigs to have consensus; this re-assembling concept significantly captures the benefit of all individual assemblers (8). The Metagenomic Data Annotation and Clustering identifies tRNAs, rRNAs, and ORFs from the input reads, performs clustering on the reads and ORFs, then annotates against Pfam, TIGRFAM, KOG, and COG (9). Taxonomy Binning assigns a taxonomy path to a read using third party tools like MEGAN (10), RDP Classifier (11), Greengene (12), and the Silva database (13). CAMERA's BLAST analysis offers all six BLAST programs, can takes as many as several hundred thousand sequences against metagenomic datasets or reference genomes including KEGG database, and has a graphic output interface as well as export functions.

Another important aspect of the workflow environment is that these workflows are organized into a systematic network, in which the output for one functional unit can be used as an input for the next workflow. This allows researchers to build a complete end-to-end analysis stream by choosing to use different combinations of workflows based on their specific needs of that data analysis. For example, one possible end-to-end analysis stream (Figure 1) for researchers with raw sequencing data may entail: i) use of the QC filter to do data quality control; ii) assemble the resultant reads to longer contigs; iii) assign taxonomy to each of these contig; iv) annotate genes against COG, Pfam, TIGRFAM, and other reference databases, then cluster the genes to the desired level; v) run a statistical comparison, obtain a graphic view, and so on. Researchers can take a "one stop shopping" advantage using this workflow linkage capability. CAMERA workflow and tools are available at http://camera.calit2.net/

REFERENCE:

1. Shulei Sun, Jing Chen, Weizhong Li, Jeff Grethe, Ilkay Altinatas, Abel Lin, Steve Peltier, Karen Stocks, Eric Allen, Mark Ellisman and John Wooley, Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis - The CAMERA Resource, Nucleic Acid Res., in press, 2010
2. Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers. Examining the Challenges of Scientific Workflows. Computer, 40(12):24–32, 2007.
3. E. Deelman, D. Gannon, M. Shields and I. Taylor. Workflows and e-Science: An Overview of Workflow System Features and Capabilities, FGCS, 25(5):528-540, 2009.
4. Bertram Ludaescher and Ilkay Altintas and Chad Berkley and Dan Higgins and Efrat Jaeger and Matthew Jones and Edward A. Lee and Jing Tao and Yang Zhao , Scientific workflow management and the Kepler system. (2006) Concurrency and Computation: Practice and Experience. Volume 18. Pages: 1039-1065.
5. Meyer,F., Paarmann,D., D'Souza,M., Olson,R., Glass,E.M.,Kubal,M., Paczian,T., Rodriguez,A., Stevens,R., Wilke,A. et al.(2008) The Metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes.BMC Bioinformatics, 9, 386.
6.
7. Markowitz,V.M., Ivanova,N.N., Szeto,E., Palaniappan,K., Chu,K.,Dalevi,D., Chen,I.M.A., Grechkin,Y., Dubchak,I., Anderson,I.et al. (2008) IMG/M: a data management and analysis system for metagenomes. Nucleic Acid Res., 36, D534–D538.
8. Niu B., Fu L. Sun S., and Li W. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinformatics 2010, 11:187
9. Sitao Wu, Limin Fu, Beifang Niu and Weizhong Li, MetaRAC: meta-assembly of pyrosequencing reads from metagenomic samples. Unpublished
10. Li W. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. BMC Bioinformatics 2009, 10:359
11. D.H. Huson, A.F. Auch, Ji Qi and S.C. Schuster, MEGAN Analysis of Metagenomic Data. Genome Research. 17:377-386, 2007.
12. Wang, Q, G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol.73(16):5261-7.
13. http://greengenes.lbl.gov/cgi-bin/nph-index.cgi
14. Silva: http://www.arb-silva.de/

Managing the Primary Data and Metadata Challenges in Metagenomics

Jing Chen, Shulei Sun, Weizhong Li, Jeff Grethe, Ilkay Altinatas, Abel Lin, Steve Peltier, Karen Stocks, Eric Allen, Mark Ellisman and John Wooley,

University of California San Diego

The Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA)[1] provides a single system for depositing, locating, analyzing, visualizing and sharing data about microbial biology through an advanced web-based analysis portal. Metagenomic researches need to access large amount of primary sequence data together with complex metadata. Metadata provide the context of sequence data and is essential for Metagenomics studies.

Meeting the requirement of metadata management is one important aspect of CAMERA. The data model in the CAMERA database is fully compliant with the MIMS/MIGS/MIENS (MIxS)[2,3] standards promulgated by the Genomic Standards Consortium (GSC). In the CAMERA database, we have created a data dictionary that not only manages the genome/metagenome metadata as defined in GSC's MIxS check list, but also manages metadata that have been standardized by CAMERA staff. In particular, the CAMERA database connects metadata terms and concepts with external ontologies such as the Environmental Ontology (ENVO)[4]. It links metadata relevant to environmental metagenome datasets with annotations in a semantically-aware environment. Two query interfaces have been created in the CAMERA portal for metadata queries; namely, a GIS based graphical user interface and a form based query builder interface. These interfaces allow users to write expressive queries on a broad range of metadata categories such as habitat, sample type, time, location, and other environmental physicochemical parameters.

CAMERA participates in the GSC developer group whose goal is the adoption of GSC standards in software development that helps users utilize GSC standards in their research. CAMERA currently collaborates with the GSC developer group on creating a GSC logical data model which is a design template for metagenomic data/metadata storage and exchange. Within the CAMERA environment, the focus is the application of these standards for the content and the format of the metagenomic data and metadata and its submission to the CAMERA repository. To ensure wide dissemination and ready access to the primary data and its annotation, CAMERA provides data submission tools to allow researchers to share their data. In the future, based on these GSC interactions, CAMERA will have the ability to easily share data with other metagenomics sites. CAMERA currently has multiple interfaces for easy submission of large or complex datasets, and supports pre-registration of samples for sequencing.

In addition to metagenomic sequence data, the CAMERA database integrates a broad collection of reference data, which includes reference genomes and proteins from NCBI Genbank and Refseq, taxonomy, PFAM[5], Gene Ontology[6] and COG[7]. These reference data are updated on a regular basis. CAMERA is not only a database, it is also a computational environment which integrates a collection of tools and viewers for analyzing, annotating, and comparing metagenome and genome data (for further details about managing the data analysis challenge, see the other poster: New-Generation Metagenomic Analysis Tools: CAMERA 2.0 Workflow System). CAMERA is available at http://camera.calit2.net/

REFERENCE:

1. Shulei Sun, Jing Chen, Weizhong Li, Jeff Grethe, Ilkay Altinatas, Abel Lin, Steve Peltier, Karen Stocks, Eric Allen, Mark Ellisman and John Wooley, Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis - The CAMERA Resource, Nucleic Acid Res., In press, 2010

2. Dawn Field, George Garrity, Tanya Gray, Norman Morrison et al. The minimum information about a genome sequence (MIGS) specification Nature Biotechnology *26,* 541 - 547 (2008) | doi:10.1038/nbt1360

3. http://gensc.org/gc_wiki/index.php/MIENS

4. http://www.environmentontology.org/

5. R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunesekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy, A. Bateman The Pfam protein families database. Nucleic Acids Research (2010) Database Issue 38:D211-222

6. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat. Genet.. May 2000;25(1):25-9.

7. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 2001 Jan 1;29(1):22-8.

GSC-Standards-Compliant RDP Tools

Jordan Fish, Qiong Wang, Benli Chai, James Tiedje and James Cole

Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824

RDP's Genome Browser offers a convenient method for researchers to locate rRNA sequences from completed genomes. It provides a detailed display for each organism with links to additional genome-specific information. The initial version of the Genome Browser retrieved this information from the GSC Rosetta Stone hosted by NERC. The external information was cached on the server side to reduce the load on the NERC server, and was accessible even when the NERC server was unavailable. When GSC switched to NCBI linkouts for a new GSC Rosetta Stone, RDP was among the first GSC members to submit linkouts to NCBI. The current RDP Genome Browser uses this linkout-based Rosetta Stone. Links are retrieved by the client's browser using AJAX requests directly to NCBI's service. We found it straightforward to integrate NCBI's linkout REST service with the RDP Genome Browser using AJAX. At this time fewer GSC members provide NCBI linkouts than provided information to the NERC Rosetta Stone system.

Submission of gene-targeted metagenomics data to the NCBI, ENA and DDBJ Short Read Archive (SRA) requires a complex set of XML metadata files along with trace data files. Preparing these XML files without specialized tools is cumbersome and beyond the technical ability of many bench scientists. myRDP SRA Prepkit offers researchers tools to prepare the complicated XML documents required. The SRA Prepkit web interface was developed using xforms and is displayed using Orbeon Forms serverside preprocessing. It gives uses a set of interrelated forms for entering the required metadata along with tools for preparing the final submission. Soil sample attributes are included in the metadata forms to assist researchers in conforming to the Minimal Information about an ENvironmental Sequence (MIENS) specification. In the future we plan to add MIENS attributes for all environment specializations.

The vision of the Genomic Standards Consortium in today's ultra high-throughput sequencing era

Peter Sterk and the Genomic Standards Consortium

Wellcome Trust Sanger Institute, Hinxton, UK

The rapid pace of genomic and metagenomic sequencing projects, which include studies of microbiomes, will only increase as the use of ultra-high-throughput sequencing methods becomes commonplace. It is clear that we need new standards to capture additional contextual data as well as tools to support its use in downstream computational analyses. It is also clear that these standards will be vital to exploring the complex interactions that take place in communities – both microbial communities, such as those sampled in marine environments, and host-microbial communities, such as those now being sampled in the Human Microbiome Project.
The Genomic Standards Consortium (GSC) is an open-membership community of researchers working in a range of research fields that include biologists, computer scientists, those building genomic databases and conducting large-scale comparative genomic analyses, and those building community-based standards. The vision of the GSC is to broaden the possible research questions that can be asked in comparative analysis studies by adding more contextual metadata to the ever growing number of sequence data sets. At the same time, the aim of the GSC is to greatly simplify data management and data integration by leveraging metadata and other standards and thus drive forward computational studies of genomes, metagenomes and marker sequences and related data.

The mission of the GSC is to work with the wider community towards (1) the implementation of a new genomic standards; (2) the development of methods for capturing and exchanging metadata and (3) harmonization of metadata collection and analysis efforts across the wider genomics community. In addition, The GSC encourages the use of relevant terminologies (ontologies) and contributes to projects that develop these. It sees the need for new methods of publishing that aid computational studies, and has as one of its activities lauched a new electronic journal SIGS (Standards in Genomic Sciences; http://standardsingenomics.org) in order to provide an open-access publication for the rapid dissemination of both genome and metagenome reports compliant with the minimal information standards developed by the GSC: MIGS/MIMS/MIENS for respectively genome, metagenome and environmental sequences. The GSC actively supports standards-related software development.

The GSC has been organizing workshops on a regular basis during which participants have the opportunity to advance its core projects, propose new ones and establish linkages between the GSC and relevant scientific projects. The GSC maintains its own wiki at http://gensc.org/, which has detailed information about its activities and related projects and links to its own and related scientific publications.

GSC 10 Participants

Name Email	
David Aanensen	d.aanensen@imperial.ac.uk
Karen Abraham	karen.abraham@unilever.com
Ilkay Altintas	altintas@sdsc.edu
Sam Angiuoli	angiuoli@gmail.com
Matthew Bietz	mbietz@u.washington.edu
Brian Bramlett	brian.bramlett@luxbiogroup.com
J. Gregory Caporaso	gregcaporaso@gmail.com
Jing Chen	jic002@ucsd.edu
James Cole	colej@msu.edu
Mark D'Souza	dsouza@mcs.anl.gov
Wim De Smet	Wim.DeSmet@UGent.be
Narayan Desai	desai@mcs.anl.gov
Dawn Field	dfield@ceh.ac.uk
Jordan Fish	fishjord@msu.edu
Jack Gilbert	gibertjack@anl.gov
Elizabeth Glass	marland@mcs.anl.gov
Frank Oliver Glöckner	fog@mpi-bremen.de
Markus Goeker	markus.goeker@dsmz.de
Antonio Gonzalez Pena	Antonio.gonzalezpena@colorado.edu
Jeffrey Grethe	jclawren@@ucsd.edu
Travis Harrison	teharrison@anl.gov
Lynette Hirschman	lynette@mitre.org
Sarah Hunter	hunter@ebi.ac.uk
Kevin Keegan	kkeegan@anl.gov
Hans-Peter Klenk	hpk@dsmz.de
Rob Knight	rob.knight@colorado.edu
Renzo Kottmann	rkottman@mpi-bremen.de
Nikos Kyrpides	nckyrpides@lbl.gov
Konstantinos Liolios	kliolios@lbl.gov
Eamonn Maguire	eamonnmag@googlemail.com
Anup Mahurkar	amahurkar@som.umaryland.edu
Konstantinos Mavrommatis	KMavrommatis@lbl.gov
Daniel McDonald	mcdonadt@colorado.edu
Folker Meyer	folker@mcs.anl.gov
Ilene Mizrachi	mizrachi@ncbi.nlm.nih.gov
Norman Morrison	norman.morrison@manchester.ac.uk
Sarah O'Brien	sobrien@anl.gov
Sarah Owens	Sarah.Owens@anl.gov
Ioanna Pagani	ipagani@lbl.gov

Blair Paul	bgpaul@umail.ucsb.edu
Philippe Rocca-Serra	proccaserra@googlemail.com
Oliver Ryder	oryder@ucsd.edu
Susanna Sansone	sa.sansone@gmail.com
Lynn Schriml	lschriml@som.umaryland.edu
Maulik Shukla	mshulka@vbi.vt.edu
Bruno Sobral	sobral@vbi.vt.edu
Peter Sterk	psterk1@googlemail.com
Jesse Stombaugh	jesse.stombaugh@colorado.edu
Daniel Sullivan	dsullivan@vbi.vt.edu
Shulei Sun	s2sun@ucsd.edu
Mathangi Thiagarajan	mathangi@jcvi.org
William Trimble	trimble@anl.gov
Bert Verslyppe	Bert.Verslyppe@UGent.be
Qiong Wang	wangqion@msu.edu
Owen White	owhite@som.umaryland.edu
Andreas Wilke	wilke@mcs.anl.gov
Jared Wilkening	jared@mcs.anl.gov
Katherine Wolstencroft	Manchester@som.umaryland.edu
K. Eric Wommack	keric.wommack@gmail.com
Hongwei Wu	hongwei.wu@gatech.edu