Pacific Symposium on Biocomputing 15:481-484(2010)

GENOMIC STANDARDS CONSORTIUM WORKSHOP: METAGENOMICS, METADATA AND METAANALYSIS (M3)

PETER STERK

NERC Centre for Ecology and Hydrology,
Oxford, OX1 3SR, United Kingdom

LYNETTE HIRSCHMAN

Information Technology Center, The MITRE Corporation, 202 Burlington Road
Bedford, MA 01730, USA

DAWN FIELD

NERC Centre for Ecology and Hydrology,
Oxford, OX1 3SR, United Kingdom

JOHN WOOLEY

University of California San Diego, 9500 Gilman Drive
La Jolla, CA 92093, USA

The M3 workshop has, as its primary focus, the rapidly growing area of metagenomics, including the metadata standards and the meta-analysis approaches needed to organize, process and interpret metagenomics data. The PSB Workshop builds on the first M3 meeting, a Special Interest Group (SIG) meeting at ISMB 2009, organized by the Genomics Standards Consortium.

1. M3: Metagenomics, Metadata, MetaAnalysis

1.1. *Background*

There are now thousands of genomes and metagenomes available for study (see the Genomes Online Database (http://www.genomesonline.org/) [1]. Interest in improved sampling of diverse environments (e.g. ocean, soil, sediment, and a range of hosts) combined with advances in the development and application of ultra-high throughput sequencing methods will vastly accelerate the pace at which new metagenomes are generated. For example, in 2007, the Global Ocean Survey published scientific analyses of 41 metagenomes [2], and as of October 2008, the submission of user-generated metagenomes to the public MG-RAST annotation server surpassed 1300 [3]. We have entered an era of ‘mega-sequencing projects’ that now include funded projects like the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project and the Human Microbiome Project, with many more equally visionary projects on the horizon.

While a genome represents the full genetic (DNA) complement of a single organism, metagenomes represent the DNA of an entire community of organisms. Metagenomes are partial samples of complex and largely unknown communities that can often only be poorly assembled. Genome and metagenomes are now also being complemented with studies of metatranscriptomes (community transcript profiles) and metaproteomes (community protein profiles). The integrative study of these datasets including those from multi-omic experiments of the same biological samples, bring with them the demand for new computational approaches. These data hold the promise of unparalleled insights into fundamental questions across a range of fields including evolution, ecology, environmental biology, health and medicine. Advances stem from improved understandings of the combinations, abundances and functions of the organisms in these communities and their genes and pathways. We are just starting to exploit these new technologies to understand the microbial world, their role in climate and biogeochemical processes and potential for bioenergy sources, and their implications for human health. The data sets will transform our knowledge of microbial evolution and ecology, since we have only scratched the surface in terms of sampling natural microbial diversity in terms of space and time. At the same time, the emerging science of metagenomics offers insight into eukaryotes by way of the roles of their microbiota, both mutualistic and pathogenic. The rapid pace of genomic and metagenomic sequencing projects [4], which now include studies of microbiomes, will only increase as the use of ultra-high-throughput sequencing methods becomes more common place. Therefore, the role of standards becomes even more vital to scientific progress and data sharing. 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.

1.2. *The PSB M3 Workshop*

The M3 Workshop at PSB 2010 builds directly on the past GSC workshops and the ISMB M3 SIG. Its focus is on comparative studies of (meta)genomes that bring these sequences into "context" (i.e., by geolocation, habitat, organism phenotype, etc). For example, a seminal paper, illustrating the power of this approach, has recently been published in PNAS [5]. It reports a study aimed at elucidating the relationships between metabolic pathways and environmental parameters in microbial communities using the data and metadata from the Global Ocean Survey (GOS), an earlier landmark paper in the history of the field of metagenomics [2]. The kick-off of the Human Microbiome Project and the resulting data sets will open enormous new possibilities for integration and analysis of metagenomic data sets in context.

The agenda of this M3 workshop has been designed to cover the growing intersection of science and standards. The workshop combines talks selected from abstract submissions, and a panel discussion to give a "voice" to the community. Building such community-driven consensus, in the form of standards that support and accelerate scientific discovery in biology, is of growing importance. This is especially true given the rapid growth of experimental data, most notably including both genomic and metagenomic sequences.

1.3. *The Genomic Standards Consortium*

The establishment of the Genomic Standards Consortium in late 2005 and its growing membership and activities attests to the growing interest in this area and the willingness of a wider range of researchers to become involved in this area of work. The GSC has largely been an activity centered in Europe to date (the UK) with strong involvement from the US. We feel it is essential to encourage the involvement of researchers across Asia, especially given the growing investments in genomic technologies in this area of the world. The PSB offers an ideal opportunity to engage bioinformaticians from around the world, and notably, to begin discussions with leading scientists from Asia.

The Genomic Standards Consortium (GSC) is organizing the M3 workshops as part of its goal to create richer descriptions of our collection of genomes and metagenomes through the development of standards and tools for supporting compliance and exchange of contextual information [6]. Established in September 2005, this international community includes representatives from the International Nucleotide Sequence Database Collaboration (INSDC), major genome sequencing centers, bioinformatics centers and a range of research institutions.

The GSC has been responsible for promulgating the MIGS/MIMS standard (Minimal Information about Genomic/Metagenomics Sequences), and, at the latest GSC meeting in September 2009, a new standard MIENS (Minimal Information about Environment Sequences). These standards are being incorporated into the INSDC (International Nucleotide Sequence Database Collaboration) as part of a new “structured comment field”. This development will be explored in a panel session that will be part of the workshop, involving representatives from EBI, Genbank and DDBJ. The GSC has also launched 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 MIGS/MIMS standards; the first issue (July 2009) contains reports on seven sequenced bacterial genomes.

2. M3 Workshop Structure

The workshop goal is to attract experimentalists and computational researchers making "next-generation" use of contextual metadata. The workshop is divided into two parts – a set of contributed talks that highlight specific research activities, and a panel of leaders in the metagenomics community who will discuss the broad issues related to generation of metagenomics data, metadata standards and tools to support the metaanalysis. In addition, the workshop includes a poster session to highlight recent advances related to the M3 goals and GSC activities.

The contributed talks describe comparative metagenomic studies that demonstrate the power provided by data curated (e.g., habitat or host) and measured (e.g. geographic location, salinity, temperature, or pH) using appropriate metadata standards. Likewise, we have welcomed studies describing new approaches, tools, databases, standards, ontologies or substantial new sets of curated metadata that aid in the integration and inter-operability of disparate datasets. We have also included discussion of research focused on capture and organization of metadata, for example through text-mining and ontology development, that enables new understanding of the interaction of organisms in their ecological context.

The talks will cover the three “M”s:

Metagenomics

- *Using 100 years of data to contextualize metagenomics in the Western English Channel*. Jack Gilbert, Plymouth Marine Laboratory, UK
- *Metagenomics reveals functional shifts in the bovine rumen microbiota composition with propionate intake.* Michael E. Sparks, Animal and Natural Resources Institute, USDA, Agricultural Research Service, Beltsville, USA

Metadata

- *Gemina: Ontology and Metadata Standards Development provide core of Infectious Pathogen Surveillance and Geospatial Tool.* Lynn Schriml, University of Maryland School of Medicine, Baltimore, USA

MetaAnalysis

- *Comparative Microbial Genomics of resistance genes in Staphylococcus aureus*. Anja Stausgaard, The Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
- *More accurate taxonomic assignment of short reads*. Gabriel Valiente, Algorithms, Bioinformatics, Complexity and Formal Methods Research Group, Technical University of Catalonia, E-08034 Barcelona, Spain

The panel will include GSC board members and metagenomic data producers, organizers from the main (meta)genomic databases, and tool producers. The central theme for discussion is “unifying access to our current collection of genomes and metagenomes.”

Acknowledgments

The organizers gratefully acknowledge the support from the NSF grant RCN4GSC, DBI-0840989. LH has also been supported in part by NSF IIS 0844419: SGER for Utility and Usability of Text Mining for Biological Curation. PS is supported by NERC grant (NE/E007325/1) to DF.

References

1. K. Liolios, K. Mavromatis, N. Tavernarakis and N. C. Kyrpides, Nucleic Acids Res **36**, D475-479 (2008).
2. D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcon,

V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier and J. C. Venter, PLoS Biol **5** (3), e77 (2007).

3. F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening and R. A. Edwards, BMC Bioinformatics **9**, 386 (2008).
4. D. Field, G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M. J. Allen, S. V. Angiuoli, M. Ashburner, N. Axelrod, S. Baldauf, S. Ballard, J. Boore, G. Cochrane, J. Cole, P. Dawyndt, P. De Vos, C. DePamphilis, R. Edwards, N. Faruque, R. Feldman, J. Gilbert, P. Gilna, F. O. Glockner, P. Goldstein, R. Guralnick, D. Haft, D. Hancock, H. Hermjakob, C. Hertz-Fowler, P. Hugenholtz, I. Joint, L. Kagan, M. Kane, J. Kennedy, G. Kowalchuk, R. Kottmann, E. Kolker, S. Kravitz, N. Kyrpides, J. Leebens-Mack, S. E. Lewis, K. Li, A. L. Lister, P. Lord, N. Maltsev, V. Markowitz, J. Martiny, B. Methe, I. Mizrachi, R. Moxon, K. Nelson, J. Parkhill, L. Proctor, O. White, S. A. Sansone, A. Spiers, R. Stevens, P. Swift, C. Taylor, Y. Tateno, A. Tett, S. Turner, D. Ussery, B. Vaughan, N. Ward, T. Whetzel, I. San Gil, G. Wilson and A. Wipat, Nat Biotechnol **26** (5), 541-547 (2008). J. Raes, K. U. Foerstner and P. Bork, Curr Opin Microbiol **10** (5), 490-498 (2007).
5. J. Raes, K. U. Foerstner and P. Bork, Curr Opin Microbiol **10** (5), 490-498 (2007).
6. D. Field, G. M. Garrity, S. A. Sansone, P. Sterk, T. Gray, N. Kyrpides, L. Hirschman, F. O. Glockner, R. Kottmann, S. Angiuoli, O. White, P. Dawyndt, N. Thomson, I. S. Gil, N. Morrison, T. Tatusova, I. Mizrachi, R. Vaughan, G. Cochrane, L. Kagan, S. Murphy and L. Schriml, OMICS **12** (2), 109-113 (2008).