Information and Data Sharing Policy

Genomics: GTL Program

Office of Biological and Environmental Research

Office of Science

Department of Energy

Final Date: April 4th , 2008

Introduction

Experimental biology has evolved in the past 20 years to include a rapid-access, global scientific community hyper connected through the Internet. The changing scope of scientific inquiry and the astonishing rate of data production drives the development of a new type of cyber infrastructure, which, in turn has promoted the formation of e-science (1). Journals, funding agencies and governments correspondingly have developed information standards and sharing policies, all of which in one way or another address research conducted in an open-access environment. A key hallmark of these policies is the requirement that scientific inquiry and publication must include the submission of publication relevant information and materials to public repositories. For the most part, the policies follow the uniform principle for sharing integral data and materials expeditiously (called UPSIDE)(2). Conversely, when research information is not made publicly available to a global scientific community, a corresponding price is paid in lost opportunities, barriers to innovation and collaboration, and the obvious problem of unknowing repetition of similar work (3).

This statement summarizes the information and data-sharing policy within the Genomics: GTL (GTL) program at the Department of Energy's Office of Biological and Environmental Research (OBER). OBER recognizes that successful implementation of this policy will require the development of new technologies such as software tools and database architectures, and will be funded, as necessary, from the GTL program subject to funding availability. We affirm our support for the concept of information and data standards and sharing and we believe that a comprehensive policy can be constructed that will encourage GTL researchers to exchange new ideas, data and technologies across the GTL program and the wider scientific community.

Research information obtained through public funding is a public trust. As such, this information must be publicly accessible. The GTL information-sharing policy requires that all publication related information and materials be made available in a timely manner. All Principal Investigators (PIs) within the GTL program will be required to construct and implement an Information and Data-Sharing Plan that ensures this accessibility as a component of their funded projects.

Policy Statement

The Office of Biological and Environmental Research (OBER) will require that all publishable information resulting from GTL funded research must conform to community recognized standard formats when they exist, be clearly attributable, and be deposited within a community recognized public database(s) appropriate for

the research conducted. Furthermore, all experimental data obtained as a result of GTL funded research must be kept in an archive maintained by the Principal Investigator (PI) for the duration of the funded project. Any publications resulting from the use of shared experimental data must accurately acknowledge the original source or provider of the attributable data. The publication of information resulting from GTL funded research must be consistent with the Intellectual Property provisions of the contract under which the publishable information was produced

I. Applicability

This policy shall apply to all projects receiving funding in the Genomics: GTL program as of October 1, 2008. For cases where information sharing standards or databases do not yet exist, the information sharing and data archiving plan provided by a project's PI must state these limitations. Data and information that are necessary elements of protected intellectual property and related to a pending or future patent application are explicitly exempt from public access until completion of the patenting process. Adherence to this policy will be monitored through the established procedure of yearly progress reports submitted to GTL program managers. All information regarding data shared by GTL-funded research projects will be made publicly available at genomicsgtl.energy.gov/datasharing.

II. Submission of Information and Data

All investigators are expected to submit their publication related information to a national or international public repository, when one exists, according to the repository's established standards for content and timeliness but no later than 3 months after publication. This includes:

Experimental protocols,
Raw and/or processed data, as required by the repository,
Other relevant supporting materials.

OBER will maintain a website listing all published peer reviewed papers and published patents resulting from GTL funding and PIs are expected to inform OBER on a regular basis when a publication appears in print. OBER is encouraged by the development of the National Institutes of Health open-access policy and, when possible, OBER will link to open-access GTL funded publications. PI's, however, are encouraged to publish in journals appropriate to their fields of research. OBER recognizes that sub-disciplines and experimental technologies have varying degrees of cyber-infrastructure and standard ontology to accommodate this policy. Specific guidelines and suggestions for GTL investigators are provided below.

II. A. Nationally and Internationally-Accepted Databases and Ontologies

II.A.1. Sequence Data

The field of genomic sequencing has a very well developed mechanism for public archiving of experimental data. Nucleotide sequence data will be deposited into GenBank, and protein sequence data will be deposited into the UniProtkb/Swiss-Prot Protein Knowledge database. Investigators should report to OBER the sequence identifier including the accession number and version. In addition, investigators are encouraged to use the gene ontology annotation database(4) when possible and OBER applauds the work of the Genomic Standards Consortium (GSC) in the development of minimum information about a genome sequence standards (MIGS).

Specifically for large-scale GTL sequencing projects, OBER will adopt the policy that whole genome sequencing data, where genome completion is the stated goal, must be made publicly available 3 months after first assembly of the sequencing reads for that genome. In the case of metagenomic sequencing, data must be deposited to the National Center for Biotechnology Information (NCBI) 3 months after completion of the last sequencing run, which must be specified in the JGI User Agreement. For other types of sequencing experiments, such as expressed sequence tags (ESTs), the data will fall under the guidelines for publication of relevant information and shall be deposited to NCBI 3 months after publication.

II.A.2 Three-Dimensional Structural Data

All coordinates and related information for structures of biological macromolecules and complexes are to be deposited in the Protein Data Bank (PDB) or Nucleic Acid Databank (NDB), as appropriate. Accession codes are to be reported back to OBER.

II.A.3 Microarray and Gene Expression Data

The Microarray and Gene Expression Data (MGED) Society recommends the use of a MGED ontology for the description of key experimental conditions as, for example, using a MIAME-compliant format (MIAME, Minimum Information About a Micro-array Experiment). OBER's policy will follow the MGED recommended ontology. We further strongly encourage GTL researchers to deposit raw and transformed data sets and experimental protocols to a public microarray database and report back to OBER the accession number and URL. Possible microarray databases for data deposition include the Gene Expression Omnibus (5), ArrayExpress (6) and the Stanford Microarray Database (7).

II.B. Information Sharing Systems and Databases Under Development

II.B.1 Proteomics

The Proteomics Standards Initiative (PSI), a working group of the Human Proteome Organization (HUPO), recently outlined two standard proteomics ontologies: minimum information about a proteomics experiment (MIAPE)(8) and minimum information required for reporting a molecular interaction experiment (MIMIx)(9). Because this is an evolving initiative and the field is still immature, we cautiously encourage GTL proteomics researchers to adopt the use of MIAPE and MIMIx in their research. We are further encouraged by the development of public proteomics repository databases such as the Open Proteomic Database(10) and PEDRo (Proteome Experimental Data Repository) (11) and encourage GTL researchers to engage with these databases. However, we recognize that standards and ontologies will evolve within the proteomics community and GTL's policy will follow guidelines set forth by HUPO as they develop.

II.B.2 Other Technologies

GTL research makes use of a large variety of technologies for which there are, as yet, no national or international information standards and archival formats. Scientists in the GTL program are encouraged to participate in the efforts of research communities to develop such standards for enabling information sharing.

GTL's long term objective is to encourage the development of infrastructure for technologies that do not as yet have nationally or internationally accepted information

sharing standards. In cases where there are no public repositories or community driven standard ontologies, OBER recommends that these types of data and information be made publicly available by the PI.

III. Protection of Human Subjects

Research using human subjects provides important scientific benefits but these benefits never outweigh the need to protect individual rights and interests. OBER will require that grantees and contractors follow the DOE principles and regulations for the protection of human subjects involved in DOE research. Minimally this will require an IRB review. These principles are stated clearly in the Policy and Order documents: DOE P 443.1A and DOE O 443.1A, which are available online at www.directives.doe.gov.

IV. Systems Biology and the GTL Knowledgebase

A long-term vision for the Genomics: GTL program, as outlined in the 2005 roadmap, is an integrated computational environment for GTL systems biology (12). OBER affirms our support for the development of an integrated framework to provide for data sharing, modeling, integration, and collaborations across the program. OBER also recognizes that continued support for development of community driven standard ontologies and data-sharing policies is inherent to the successful implementation of a systems biology network.

V. Computational Software

The International Society for Computational Biology (ISCB) recommends that funding agencies follow ISCB guidelines for open-source software at a "Level 0" availability. ISCB states that research software will be made available free of charge, in binary form, on an "as is" basis for non-commercial use and without providing software users the right to redistribute. OBER will follow ISCB recommendations at a Level 0 availability. OBER recommends that research software developed with GTL funding that result in a peer-reviewed software publication is to be made accessible through either an open source license (www.opensource.org) or deposited to an open source software community such as SourceForge.

VI. Laboratory Information Management Systems (LIMS) for Data Management and Archiving

GTL systems biology research projects involve high-throughput, data intensive research that necessitates use of a data management system to automatically handle this pipeline of data. OBER's goal is that researchers within the GTL program utilize a LIMS system for managing their research data and information. Because different research agendas require different information management systems, an overarching and restrictive policy could place an undue burden on PIs. Therefore, we expect that research projects that involve more than one senior investigator will be required to implement a LIMS or a similar type of electronic system for data and information archiving and retrieval. This plan should balance the clear value of data availability and sharing against the cost and effort of archive construction and maintenance.

VII. Summary

This document outlines the Genomics: GTL program policy and will require GTL funded principle investigators to construct an information and data-sharing plan as a component of their projects. The policy requires information to conform to existing community recognized standard formats wherever possible, to be clearly attributable, and to be deposited, in a timely manner, within a community recognized public database(s) appropriate for the research conducted. OBER is committed to encouraging development of public repositories and standard ontologies for the GTL research community. OBER recognizes that this policy necessarily will be revised to include new standards, data types, and other advances that are pertinent to maximizing availability of data and information across the GTL program. This information and data-sharing policy and related materials can be found at genomicsgtl.energy.gov/datasharing.

References

1. E-science refers to large scale science that is distributed through global collaborations and enabled by the Internet. (see www.research-councils.ac.uk/escience).

2. National Research Council. 2003. Sharing publication related data and materials: Responsibilities of authorship in the life sciences. The National Academy Press, Washington DC.

3. Uhlir, P. F. and P. Schröder. 2007. Open data for global science. *Data Science Journal,* **6**:OD36-OD53.

4. Camon, E., M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler. 2004. The gene ontology annotation (GOA) database: Sharing Knowledge in Uniprot with gene ontology. *Nucleic Acids Res.*, **32**:D262-D266.

5. Edgar, R., M. Domrachev, and A. E. Lash. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**:207-210.

6. Brazma, A., H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezcimen, P. Rocca-Serra, and S.-A. Sansone. 2003. ArrayExpress-A public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**:68-71.

7. Sherlock, G., T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J. C. Matese, S. S. Dwight, M. Kaloper, S. Weng, H. Jin, C. A. Ball, M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein and J. M. Cherry. 2001. The Stanford microarray database. *Nucleic Acids Res.*, **29**:152-155.

8. Taylor, C. F., N. W. Paton, K. S. Lilley, P.-A. Binz, R. K. Julian Jr., A. R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E. W. Deutsch, M. J. Dunn, A. J. R. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T. A. Neubert, S. D. Patterson, P. Ping, S. L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T. M. Vondriska, J. P. Whitelegge, M. R. Wilkins, I. Xenarios, J. R. Yattes III, and H. Hermjakob. 2007. The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology*, **25**:887-893.

9. Orchard, S., L. Salwinski, S. Kerrien, L. Montecchi-Palazzi, M. Oesterheld, V. Stumpflen, A. Ceol, A. Chatr-Aryamontri, J. Armstrong, P. Woollard, J. J. Salama, S. Moore, J. Wojcik, G. D. Bader, M. Vidal, M. F. Cusick, M. Gerstein, A.-C. Gavin, G. Superti-Furga, J. Greenblatt, J. Bader, P. Uetz, M. Tyers, P. Legrain, S. Fields, N. Mulder, M. Gilson, M. Niepmann, L. Burgoon, J. De Las Rivas, C. Prieto, V. M. Perreau, C. Hogue, H.-W. Mewes, R. Apweiler, I. Xenarios, D. Eisenberg, G. Cesareni, and H. Hermjakob. 2007. The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature Biotechnology*, **25**:894-898.

10. Prince, J. T., M. W. Carlson, R. Wang, P. Lu, and E. M. Marcotte. 2004. The need for a public proteomics repository. *Nature Biotechnology*, **22**:471-472.

11. Garwood, K., T. McLaughlin, C. Garwood, S. Joens, N. Morrison, C. F. Taylor, K. Carroll, C. Evans, A. D. Whetton, S. Hart, D. Stead, Z. Yin, A. J. P. Brown, A. Hesketh, K. Chater, L. Hannson, M. Mewissen, P. Ghazal, J. Howard, K. S. Lilley, S. J. Gaskell, A. Brass, S. J. Hubbard, S. G. Oliver, and N. W. Paton. 2004. PEDRo: A database for storing, searching and disseminating experimental proteomics data. *BMC Genomics*, **5**:68.

12. U.S. Department of Energy Office of Science Office of Biological and Environmental Research. 2005. *Genomics:GTL Roadmap: Systems Biology for Energy and Environment* (see http://genomicsgtl.energy.gov/roadmap/index.shtml).