CONTRACT CONCEPT REVIEW

NTP Board of Scientific Counselors Meeting
December 2, 2015

Concept Title: Bioinformatics Support for DNTP and DIR

NTP Scientist: B. Alex Merrick, Biomolecular Screening Branch, DNTP

Purpose

We propose a new contract to provide bioinformatics support for research within the NIEHS Division of the National Toxicology Program (DNTP) and Division of Intramural Research (DIR):

1. **Bioinformatics support for DNTP studies (67%).** Bioinformatics supports data analysis from a diversity of genomic and high throughput assay studies requiring innovative genomics, toxicogenomics, cheminformatics, biostatistical, textual and visual analysis needs. Some examples of bioinformatic needs include pipelines for Tox21 screening data analysis and chemical prioritization, NextGen and high throughput expression analysis, DNA methylation and epigenetics (e.g., Mouse Methylome Project), and chemical effects upon biochemical pathways and networks. Datastreams may come from animal studies (from short-term and long-term rodent studies from NTP contracts or NTP Laboratory experiments), *in vitro* high throughput studies, or human studies.
2. **Bioinformatics support for DIR studies (33%).** Bioinformatics support would include analysis of molecular, genomic and epigenomics data from laboratory experiments and human studies from DIR investigators.

Background and Significance

The National Toxicology Program (NTP) is a federal, interagency program, headquartered at the NIEHS, whose goal is to provide information about potentially toxic chemicals to health, regulatory, and research agencies, scientific and medical communities, and the public. This information includes experimental and large-scale biological data about potentially hazardous agents obtained by taking advantage of technological advances in molecular biology and computer science. Current NTP initiatives include examining the effects of cell phone radiation, endocrine disruptors, herbal supplements, chemical mixtures and nanomaterials, as well as developing new approaches to advance high throughput (high speed and high quantity) screening of chemicals, and to reduce the number of animals used in research. As part of its mission, NTP conducts short-term and long-term rodent studies primarily through contract laboratories, which are managed by NTP staff. NTP also conducts high and medium throughput *in vitro* assays through the Tox21 program in partnership with the NIH Chemical Genomics

Center (NCGC) and the Environmental Protection Agency (EPA). Several NTP Offices are engaged in evaluating scientific literature and completing systematic reviews on selected topics that require bioinformatic support. Additionally, the NTP Laboratory, DNTP collects a variety of mission-related data, including *in vitro* screening assays and bioassays using *Caenorhabditis. elegans,* zebrafish, and other alternative model species.

The DIR at NIEHS conducts *in vitro* and *in vivo* research on biological and chemical processes, the role of environmental agents in human disease and dysfunction, underlying mechanisms of environmentally associated diseases, and intervention/prevention strategies to reduce the effects of exposures to environmental agents. An example of a collaborative DIR-DNTP study involving the bioinformatic support is the Mouse Methylome Project. The goal of this project to map, visualize, and interpret DNA methylation patterns using whole genome bisulfite sequencing of parental C57 and C3H mouse strains and F1 progeny to gain insight into strain specific differences for hepatocellular carcinoma. Innovative approaches were developed to group DNA methylations into differential methylated regions (DMRs) that would alter gene expression appear related to injury response and regeneration.

NIEHS invests substantial time and funding into the collection of mission-related data; appropriate bioinformatic data analyses are required to support accurate and precise interpretation of study results.

Bioinformatic support for NTP studies

NIEHS researchers conduct studies that produce large amounts of data, varying in size and complexity. Fields of scientific study are diverse and include toxicology, transcriptomics, cheminformatics, genomics and high throughput screening data that require innovative bioinformatics analysis and development of new software tools. NTP conducts work in gene expression and DNA sequencing, as well as other genomics, proteomics, metabolomics, and epigenetics work that require whole genome bioinformatics data analysis. These studies involving next-generation (NextGen) sequencing of genomic material and bioinformatic analysis are highly specialized and rapidly evolving fields. The NIEHS Biostatistics and Computational Biology Branch provides statistical support to NTP; however, the quantity, complexity, and variety of NTP data require additional contract support to span the bioinformatics needs of NIEHS researchers. Thus, a bioinformatics contract separate from the statistics contract is needed to provide support for the magnitude and range of data and bioinformatics analysis for the studies conducted at NIEHS.

There are several bioinformatic strategies and processes needed by NTP. For example, public genomic databases like GEO (gene expression omnibus) and SRA (sequence read archive) are large repositories of microarray and NextGen data that need to be searched, selectively

extracted for specific chemical exposure and expression studies, and normalized for comparison with NTP expression data. A recently published paper from the Sequencing Quality Control (SEQC) Consortium involving NIEHS researchers relied upon these bioinformatic processes to show concordance between RNA-seq and microarray data (Wang et al., 2014, Nature Biotech *32*:926-932). A second example, is the need for creating consistent bioinformatic workflows or 'pipelines' for differential expression analysis and pathway analysis using NextGen sequencing methods such as RNASeq for both fresh and FFPE - formalin fixed paraffin embedded tissues as recently described by NTP researchers (Merrick et al., 2013, PlosOne *8*:e61768; Auerbach et al., 2015, J Appl Toxicol *35*:766-80). Pipelines are also being developed for selected projects involving exome sequencing of extracted DNA from mouse tumors with the objective of distinguishing the gene mutational spectra between chemical-mediated and spontaneous tumors.

Additionally, NTP is a primary participant in the Tox21 program that conducts high throughput and medium throughput screening assays on thousands chemicals. This large volume of data requires computational support to assess data analysis methodologies, structure-activity-relationships among large chemical features and classes, models of dose-response associations, and evaluation of responses across assays chemicals. NTP will use computational approaches to relate chemical features in the Tox21 dataset with high throughput screening (HTS) data for chemical prioritization. Further, bioinformatics approaches will be needed to identify and extract relevant information from the vast amount of scientific published literature relevant to Tox21 chemicals under various exposure conditions.

The next phase of Tox21 will include high throughput transcriptomics (HTT). This will involve analyzing multiplexed transcriptomic data (e.g. 2500 transcripts per sample well) for in vitro experiments across multiple concentrations with environmental chemicals. These transcriptomic data will require specialized bioinformatic procedures to contextualize expression data with activation of functional genomic groups (e.g., nuclear receptors), pathways (e.g., DNA damage response), and networks (e.g., cell differentiation). This will first require bioinformatic approaches for selection of sentinel genes and appropriate cells types to be used for HTT screens from multiple species. A plan is being developed for processing and analysis of the vast amounts of transcriptomic data arising from an HTT platform.

Bioinformatic support for DIR studies

DIR researchers conduct basic research at the molecular and cellular level and integrate data into systems and networks at the tissue and organism level. DIR researchers require bioinformatic support activities that include assistance with genomics and epigenomics studies.

For example, antibodies directed toward transcription factors or specific histones can capture regulatory portions of DNA that can be identified using NextGen sequencing methods like ChIP-Seq. Another common NextGen sequencing application is exome sequencing that involves the capture of coding regions of DNA to determine base pair variations and possible mutations at a genome wide level. These and many other molecular techniques make use of NextGen sequencing technologies in a way that is heavily reliant upon knowledge of species-specific genomics, alignments, comparative sequence and variant analysis, and transcriptional and translational biology. Epigenomics involving DNA methylation and histone modification is a relatively new and complex field of gene regulation that also requires biological knowledge and specialized algorithms. Bioinformatic methods in genomics and epigenomics are evolving rapidly and can vary depending on goals of each DIR study, so it shall be necessary for the bioinformatics contractor to stay current with developments in the field and to be able to customize data analytic methods.

In addition, DIR researchers are generating experimental data with human materials and need to compare their bioinformatics data with similar data on a genomic scale. Bioinformatic support for this research involves downloading, accessing and manipulating genomic and phenotypic data stored in public databases, such as GEO, SRA, dbSNP, COSMIC, dbGAP and many others.

Proposed Statement of Work

The Bioinformatics Support for DNTP and DIR contract is a new requirement that will provide the necessary resources to ensure appropriate bioinformatics support to DNTP and DIR. The approximate proportions of support for DNTP and DIR are, respectively, 67% and 33%, and are expected to remain stable over time.