NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION
NATIONAL LIBRARY OF MEDICINE, NIH

BOARD OF SCIENTIFIC COUNSELORS
MEETING MINUTES
November 13, 2018
9:00 a.m. – 2:00 p.m.

The Board of Scientific Counselors of the National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), convened on November 13, 2018, in the NLM Board Room, Bethesda, Maryland. The meeting was open to the public.

BSC Members Present
David Relman, M.D., Stanford University *(BSC Chair)* – participated by teleconference/webex
Michael Boehnke, Ph.D., University of Michigan – participated by teleconference/webex
Kateryna Makova, Ph.D., Penn State University
Katherine Pollard, Ph.D., University of California
Steven Salzberg, Ph.D., Johns Hopkins University
Donna Slonim, Ph.D., Tufts University
Pamela Soltis, Ph.D., University of Florida
Jianzhi Zhang, Ph.D., University of Michigan
James Ostell, Ph.D., NCBI, NLM *(BSC Executive Secretary)*

NLM Staff Present
Dennis Benson, Ph.D., NCBI, NLM
Olivier Bodenreider, M.D., Ph.D., LHC, NLM
Max Burroughs, Ph.D., NCBI, NLM
Janet Coleman, NCBI, NLM
L. Aravind (aka Aravind Iyer) Ph.D., NCBI, NLM
Laks Iyer, Ph.D., NCBI, NLM
David Landsman, Ph.D., NCBI, NLM
Leonardo Marino-Ramirez, Ph.D., NCBI, NLM
Kim Pruitt, Ph.D., NCBI, NLM
Jerry Sheehan, NLM
Bart Trawick, Ph.D., NCBI, NLM
Roberto Vera Alvarez, Ph.D., NCBI, NLM

Others Present
Charles Dearolf, M.D., OD, NIH

I. Welcome and Introductions

Dr. Relman called the meeting to order at 9:00 a.m. Members introduced themselves. Dr. Ostell thanked the Board for their work, and, because there were several new members, provided background on the general functions of NCBI's Information Engineering Branch (IEB) and Computational Biology Branch (CBB). He also explained how Board meetings typically involve reviews of senior staff members and/or focus on NCBI production resources.

Dr. Landsman reported that the next BSC meeting would be April 9, 2019.

The BSC voted to approve the minutes of the last meeting, held April 24, 2018.

II. Presentation and Review of L. Aravind, Ph.D., Senior Investigator, CBB

Dr. Aravind began his presentation with an introduction to the broad themes his research covered over the last four years. The research included a combination of long-standing multi-cycle projects, new investigations, and work with collaborators. He then focused on three major projects to give the committee an in-depth view of his research.

First, he spoke about the work his group has done on nucleotide signaling. The primary highlight of this research was the discovery of novel signaling-nucleotide-generating enzymes and sensory domains that recognize these nucleotides to further transmit the signal. In particular, he emphasized the discovery of a common functional principle based on nucleotide signaling, unifying diverse CRISPR/Cas systems and interferon-induced oligo 2'-5' A signaling in animals.

Dr. Aravind then described his research on biological conflicts and how enzymatic effectors are used to attack RNA in such conflicts. He elaborated on the discoveries he and his group made on novel enzymes that repair RNA damaged in such conflicts.

The third project he described was the discovery of a novel translation release factor that had been sought after for more than a decade. He presented evidence how this release factor uses a conserved glutamine to catalyze the release of stalled peptides from the tRNA as part of a quality control system in the cell that helps clear defective translation products at stalled ribosomes.

He ended his presentation with a quick summary of his publications during the review period and a summary of the research done with collaborators. Following Dr. Aravind's presentation, the BSC met in closed session with him.

III. Presentation and Review of David Landsman, Ph.D., Senior Investigator, CBB

Dr. Landsman focused his presentation on two of the seven key projects his group has worked on over the past four years.

The first project involved HMGN1 and HMGN2 proteins. These proteins, Dr. Landsman said, bind to chromatin in a tissue-specific manner and stabilize cell identity. The research, which will be published in *Nature Communications*, found that HMGN proteins mark tissue-specific promoters, active enhancers, and super enhancers. Also, the research found that MEF cells lacking both HMGN1 and HMGN2 are induced into pluripotent cells significantly faster than wild type MEF. The researchers speculate that the HMGN proteins help to maintain cell identity by binding to tissue-specific promoters and enhancers.

The second project Dr. Landsman described was development of a database called "HistoneDB 2.0 – with Variants" (https://www.ncbi.nlm.nih.gov/research/HistoneDB2.0/index.fcgi/). The

database was established to better understand how sequence variation may affect functional and structural features of nucleosomes; it can be used to explore the diversity of histone proteins and their sequence variants in many organisms.

HistoneDB 2.0 replaced an earlier database, "Histone Database," which included histones and non-histone proteins containing histone folds. HistoneDB2 consists of manually curated histone variants and their multiple sequence alignments with annotated characteristic features and descriptions of their functions. Hidden Markov Models constructed based on these alignments can be used to annotate any histone-like sequence of interest, allowing for automatic annotations of histone variants. The database also allows users to compare variants and their features and provides phylogenetic trees of histone variants.

Following Dr. Landsman's presentation, the BSC met in closed session with him.

IV. Report on the 2018 NLM Blue Ribbon Panel – Jerry Sheehan, Deputy Director, NLM

Mr. Sheehan opened his presentation by thanking the BSC for their service. He noted how issues of data science are becoming prominent across NIH and that NLM is trying to help NIH achieve its data science objectives, in part by being a platform for data discovery and data-powered health.

Mr. Sheehan provided general background on the Blue Ribbon Panel review of NLM. He explained that the reviews are an NIH-wide activity that are conducted for all ICs that have an intramural research program (24 of the 27 ICs), and they are intended as a way to get external advice and input on the operation of the program. The reviews typically are done about every 10 years. The Blue Ribbon Panel review of NLM's research program is expected to be posted online shortly (possibly before the end of the day).

The charge to the panel was:

- Review the strengths and weaknesses of the intramural and training programs
- Consider the optimal balance among research, system development, and information services such as dbGaP and PubMed
- Identify priority areas in biomedical informatics & data science research and recommend ways to support training in these two areas
- Recommend any warranted changes to NLM's organizational structure, budget, staffing, internal and external partnerships
- Suggest ways to assess outcomes and impact of research and training
- Align NLM's intramural research with NLM and NIH Strategic Plans

The panel was impressed with the quality of NLM's scientists and their work and said it was critical that NLM have a vibrant and aggressive intramural research program, Mr. Sheehan said. Key recommendations from the review include:

1. Work with NIH to significantly boost NLM's investment in intramural research to support new independent investigators and embrace new opportunities for data science, informatics, and computational biology
2. Manage the intramural research program as one seamlessly connected, unified intramural research program with a single Scientific Director

3. Adopt one or more audacious, high-risk, high-reward projects to galvanize research across the organization and inspire the larger scientific community
4. Engage in a research portfolio evaluation and strategic plan to align NLM research priorities with the NLM Strategic Plan and research priorities of NIH and the broader biomedical research community
5. Engage in a vigorous program of joint investigator appointments with other ICs to create a cadre of biomedical data science and informatics investigators with specific domain expertise who consider NLM to be their technical home
6. Create mechanisms for identifying opportunities for moving research tools into services, based on current need and anticipated impact
7. Work with NIH to develop policies and procedures for optimizing successful recruitments
8. Restructure intramural research training into a single, unified training program with a designated Training Director
9. Include broad metrics of scientific outcome, leadership, and impact, as well as publications and citations, in performance assessments of NLM's intramural research program and its researchers
10. Convene a single Board of Scientific Counselors with sufficient scientific breadth and expertise to evaluate the full set of activity within a unified intramural research program
11. Work with NIH to renovate and redesign intramural research labs and shared spaces to promote greater collaboration among research groups, optimize collective use of research equipment, and enable 21st-century team science

Mr. Sheehan noted that NIH Director Dr. Collins, the Lister Hill Center BSC, and the NLM Board of Regents already have been briefed on the findings and that NLM has begun recruiting for three new investigators.

Following Mr. Sheehan's presentation BSC members commented that NLM might want to carefully consider the structure of co-research arrangements with other ICs.

V. Poster Session

Postdoctoral/research fellows for Drs. Aravind and Landsman presented posters. The 3 posters were:

Nucleotide and codon background mutability shapes cancer mutational spectrum and advances driver mutation identification – Anna-Leigh Brown, Minghui Li, Alexander Goncearenco, Anna R. Panchenko (from Dr. Landsman's group, except Dr. Li, who is from Soochow University, China)

HU/IHF and Chromo-domains: a tangled tale of evolutionary convergence and divergence in chromatin proteins – Gurmeet Kaur, A. Maxwell Burroughs, Lakshminarayan M. Iyer, Srikrishna Subramanian, L. Aravind (from Dr. Aravind's group)

Extensive diversity of AID/APOBEC-like deaminases: Evidence for widespread mutagenesis-based immunity mechanisms – Arunkumar Krishnan, Lakshminarayan M. Iyer, L. Aravind (from Dr. Aravind's group)

VI. Tier 1 Project Updates – Dr. Kim Pruitt

Dr. Pruitt provided an update on NCBI's experiment with establishing a set of Tier 1 projects in Fiscal 2018. The effort involved selecting high-priority projects on which to focus, with specific deliverables and metrics of success. Projects were selected based on customer value. The Tier 1 projects grew out of a larger NCBI reorganization over the last two years that optimized how the organization works, with an emphasis on customer value. She noted that the NCBI is doing much more formal project management and has pivoted strongly to doing customer-centered design. Also, there is increased focus on prioritizing projects and resources.

Results of the experiment were successful, with much learned and new skills developed. More specifically:

- Teams were more efficient and effective in their work
- Use of agile development methods let teams iterate and produce results in shorter time
- Project management skills were improved
- Teams met most of their deliverable goals within the specified time periods
- Throughout the projects customer value was continuously assessed (e.g. through user interviews, usability testing, A/B lab & analytics)

Overall goals of the five projects were to improve public health outcomes, improve search and retrieval, and to modernize NCBI technology. The five initial Tier 1 projects were as follows:

Pathogen surveillance

The Pathogen project aims to integrate bacterial pathogen genomes originating in food, environmental sources and patients, and then cluster and identify related sequences to uncover potential food contamination sources. Partners in the project – CDC, FDA, USDA, state health labs – submit the sequences of bacterial samples to NCBI, which integrates the data to provide high-resolution rapid clustering of related isolates to aid in traceback and outbreak investigations.

NCBI's goals for the project included improving the pipeline processing time to provide Rapid Reports within an hour of sequence submission and to provide SNP trees within 24 hours. A dashboard monitors when NCBI fails to meet the goal, which then is analyzed to understand the cause of failure and to make changes to address the problem. Another goal, also met, was to improve usability of the pathogen isolates browser so that people could more easily identify information about isolates and ones in need of follow-up. NCBI also measures how fast it can process submissions; this year the goal, which was met, was to accommodate submission growth and process at least 90,000 isolates/year.

Virus annotation & access

One goal of this project was to improve ease and quality of viral sequence submissions by providing viral annotation and submission tools, Dr. Pruitt said. The success metric – sequences processed by the Viral annotation and submission tools (vAST) – was achieved, with more than 52,000 influenza sequences processed in FY18. Another goal was improved search and retrieval of viral sequence data. NCBI improved usability of the user interface, with a new web design that is very interactive and can be quickly used to filter for host, country, and other parameters.

Known Item Search (KIS)
The goal of KIS is to deliver expected, high-value results for text searches where users have known sequence items in mind (e.g. the human genome) but don't have an ID number. One measure of success was the number of clicks on results suggested by sensors. Dr. Pruitt presented a slide that showed click through rates on the KIS sensor ranged from 19% (in the Protein database) to 47% (in the Genome database). In addition to Protein and Genome, KIS was implemented in Gquery (the global search page), Assembly, Nuccore, and Gene.

PubMed 2.0
This project started with a goal of improving PubMed search and then pivoted towards improving the usability of the website. Changes were deployed in an experimental format in PubMed Labs. The goals included: simplifying the backend content so that search is integrated across PubMed and PubMed Central; improving search quality and user experience, with Best Match and Auto-suggest, which have been implemented; improving UI/UX (user interface and user design), which included reusable components, text snippets, the abstract page and mobile access; and updating the technology, with SOLR indexing and AWS cloud access. Success metrics included fewer outages (downtime) and increased use of Best Match. Dr. Pruitt presented a slide showing improvements in both metrics.

Key services in the cloud
The goal of this project was to establish the Google and Amazon architecture for cloud services, which was successful. The initial projects were PubMed Labs and PubMed Data Management, KIS, BLAST, and the Sequence Data Delivery Pilot (SDDP, which will become the Sequence Data Delivery Project). SDDP is a way for NCBI to broker access to controlled-access data in a cloud environment; it builds upon the authorization process already in place in the Sequence Read Archive (SRA) and the database of Genotypes and Phenotypes (dbGaP). The system also allows NCBI to provide access to data that is hosted outside NCBI.

Dr. Pruitt concluded by saying that the 2018 approach was a major experiment in how the NCBI does its work and how it manages its projects, with deliverables and tracking. Results were very successful, and NCBI will be continuing this approach as its operating model going forward.

VII. Q&A/Discussion

Selection of Tier 1 projects
In response to a question about how the five Tier 1 projects were selected, Dr. Pruitt said that there was a proposal period and a review period. A production services operating board, which was started as part of reorganization and includes staff from IEB and CBB, selected the projects with input from the NCBI Director. The same process was used to select projects for FY2019.

Reorganization impetus and the cloud
Dr. Ostell commented that part of the impetus for the reorganization and the shift in operations was wanting to empower people at different levels of NCBI to make informed decisions. This need stemmed in part from succession planning for senior leadership. The reorganization also resulted from an interest in increasing engagement with NIH and the outside community as a way to scale services without a larger staff. "Pivoting outwards" requires strong project management in order to deliver work in a timely and trackable way. Lastly, with the huge growth in data, NCBI is no longer able to pay for storing it all. The SDDP project will allow the

institutes to pay for putting their own data in the cloud while using NCBI's dbGaP authorization process to control access. Similarly, having data in the cloud also allows it to be available for computation without NCBI having to bear the compute costs. Another advantage to use of the cloud is that tools from outside sources can be used.

Cloud security

Asked if there are security concerns with having data in the cloud, Dr. Ostell pointed out that the cloud likely is more secure than university systems and federal systems. He added that dbGaP is connected to grant management IDs, and those IDs are verified. Also, NIH has enforcement capabilities if an institution violates terms of its grants. Independent of the system used, there is no technical protection against a grad student violating the terms of data access, he noted.

VIII. Closed debriefing

The board met in closed session with Dr. Charles Dearolf, from the NIH Office of Intramural Research, to discuss the Senior Investigator reviews.

IX. Adjournment

The BSC adjourned at approximately 2:00 p.m.

David Relman 2/13/19

Dr. David Relman, Chair (Date)
Board of Scientific Counselors

James Ostell 2/25/2019

Dr. Jim Ostell, Director (Date)
National Center for Biotechnology Information