BioCompute: A Standardized Method to Communicate Bioinformatic Workflow Information and Ease Organizational Burden

Jonathon Keeney, Ph.D.
keeneyjg@gwu.edu
Assistant Research Professor, George Washington University
Managing Director, BioCompute Executive Steering Committee
Computational Toxicology Communities of Practice
July 25th, 2019
NGS Data Flows

Ancestry  Cancer  Microbiome  Disease correlation  Agriculture  Synthetic biology  Livestock  Metagenomics  Personalized medicine
Wasted Time and Money

Problems/Questions/Issues/Clarifications

industry

$ millions of dollars + time

Submit/resubmit

FDA

Decision

BioCompute Objects
Submitting Next Generation Sequencing Data to the Division of Antiviral Products
Experimental Design and Data Submission

Acceptable Next Generation Sequencing Platforms
The division will accept Next Generation sequencing data generated from most standard Next Generation Sequencing (NGS) platforms provided the sponsor supplies the appropriate details for the sequencing platform, the protocols to be used for sample preparation, the raw NGS data, and the methods used to analyze the data. We recommend communicating with the division early in the process and providing these details prior to submitting the sequencing data. Please consider the following information when preparing your NGS submissions.

Data Transfer
1. Portable hard drive
   a. The raw NGS data in the fastq format should be sent to the division on a secured, portable hard drive following the guidelines outlined in this Guidance:
   b. Please note that only the raw NGS data, the frequency table, and a table of contents should be contained on the hard drive. Additional files, such as those with a .exe extension may result in rejection of the submission. In addition, if the hard drive is password protected (not required or recommended at this time), please consult with the division ahead of time to ensure that the password is provided to the appropriate personnel in the document room.
   c. All additional data should be submitted via the electronic document gateway.
Wasted Time and Money

Problems/Questions/Issues/Clarifications

$ millions of dollars + time

EPA

Submit/resubmit

Decision
Use of High Throughput Assays and Computational Tools; Endocrine Disruptor Screening Program; Notice of Availability and Opportunity for Comment

A Notice by the Environmental Protection Agency on 06/19/2015

AGENCY:
Environmental Protection Agency (EPA).

ACTION:
Notice.

SUMMARY:
This document describes how EPA is planning to incorporate an alternative scientific approach to screen chemicals for their ability to interact with the endocrine system. This will improve the Agency’s ability to fulfill its statutory...
Assessing and Managing Chemicals under TSCA

Alternative Test Methods and Strategies to Reduce Vertebrate Animal Testing

The Toxic Substances Control Act (TSCA), as amended by the Frank R. Launtenberg Chemical Safety for the 21st Century Act, directs EPA to:

- reduce and replace, to the extent practicable and scientifically justified, the use of vertebrate animals in the testing of chemical substances or mixtures; and
- promote the development and timely incorporation of alternative test methods or strategies that do not require new vertebrate animal testing.

TSCA also requires EPA to develop a strategic plan on this topic and provide a progress report on the implementation of the plan to Congress every five years since the date of the enactment of the Lautenberg Chemical Safety Act, i.e. beginning in 2021.

In 2018, EPA published its Strategic Plan to Promote the Development and Implementation of Alternative Test Methods within the TSCA Program. The Strategic Plan incorporated input from two public meetings and written comments submitted on the draft strategic plan.
A solution should...

• Be human readable: like a GenBank sequence record
• Be computer readable: structured information with predefined fields and associated meanings of values
• Contain enough information to understand the computational pipelines, interpret information, maintain records, and reproduce experiments
• Have a way to be sure the information has not been altered: immutable
Solution: BioCompute

- Standard for communicating computational analysis workflows
- Acts like an envelope for entire pipeline
  - Can incorporate other standards (e.g. CWL, FHIR Genomics)
- Built with FDA
- Human and machine readable
  - Written in JSON
- Categorized by domains
- Adheres to F.A.I.R. principles
- Adaptable
- Preserves data provenance
- Unique IDs for versioning
802.11 Analogy
workflow – defined in HIVE

parametrized pipeline with all internal arguments well defined

output domain

Verification Kit

Metadata
(not needed for computation)

workflow – defined in HIVE

parametrized pipeline with all internal arguments well defined

output domain

Verification Kit

Metadata
(not needed for computation)
BCO Example: HCV-1 drug resistance

- **Goal:** Identify SNPs, insertions, and deletions that correlate with reduced ledipasvir antiviral drug efficacy in Hepatitis C virus subtype 1
  - Genome sequencing data from a drug resistant cohort
BCO Example: HCV-1 drug resistance
BCO Example: HCV-1 drug resistance
BCO Example: HCV-1 drug resistance

```bash
fastq-dump -X 2 SRR001666 --split-3
Read 2 spots for SRR001666
Written 2 spots for SRR001666

go head SRR001666_1.fastq SRR001666_2.fastq

@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC

@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
GTTCAGGGATACGACGTTTGTATTTTAAGAATCTGA
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII6IBI

@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
AAGTTACCCTTAACAACTTAAGGGTTTTCAAATAGA
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIDIIIIIII>IIIIII/

@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
AGCAGAAGTCGATGATAATACGCGTCGTTTTATCAT
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
IIIIIIIIIIIIIIIIIIIIIIGII>IIIII-I)8I
```
BCO Example: HCV-1 drug resistance

Parameters

Scripts/drivers/prerequisites

Parameters

Scripts/drivers/prerequisites

RESULT
BCO Example: HCV-1 drug resistance

---

“usability_domain”: [
  "Identify baseline single nucleotide polymorphisms SNPs [SO:0000694], insertions [SO:0000667], and deletions [SO:0000045] that correlate with reduced ledipasvir [PubChem:67505836] antiviral drug efficacy in Hepatitis C virus subtype 1 [taxID:31646]",
  "Identify treatment emergent amino acid substitutions [SO:0000048] that correlate with antiviral drug treatment failure",
  "Determine whether the treatment emergent amino acid substitutions [SO:0000048] identified correlate with treatment failure involving other drugs against the same virus",
  "GitHub CWL example: https://github.com/mr-c/hive-cwl-examples/blob/master/workflow/hive-viral-mutation-detection.cwl#L20"
]

"extension_domain": {"github_extension": {"github_URL": "https://github.com/common-workflow-language/hive-cwl-examples", "github_URI": ""}}

Identify treatment emergent amino acid substitutions [SO:0000048] that correlate with antiviral drug treatment failure.

Determine whether the treatment emergent amino acid substitutions [SO:0000048] identified correlate with treatment failure involving other drugs against the same virus.

GitHub CWL example: https://github.com/mr-c/hive-cwl-examples/blob/master/workflow/hive-viral-mutation-detection.cwl#L20
BCO applications for computational toxicology

- Create dose-response modeling extensions to capture both BMDS and tcpl pipelines
- For tcpl
  - Each BCO would be the platform-specific steps used to analyze the results
Search term

You can search by BCO ID, name, or contributor.

Base type for all BioCompute Objects

All BioCompute object types must adhere to this type in order to be compliant with BioCompute framework.

**bco_id**

1

A unique identifier that should be applied to each BCO instance, generated and assigned by a BCO database engine. IDs should never be reused.

**checksum**

A string-type, read-only value, protecting the object from internal or external alterations without proper validation generated with a SHA-256 hash function.

**bco_spec_version**

1.3.0

Version of the BCO specification used to define this document.

Provenance Domain

Structured text for tracking data through transformations, including contributors, reviewers, and versioning.

**name**

Public searchable name for BioCompute Object. This public field should take free text value using common biological research terminology supporting the terminology used in the usability_domain, external references (xref), and keywords sections.

version

https://www.biocomputeobject.org/bco_editor/
2019 Workshop

• Discussion panel
  • How is workflow communication currently handled?
  • What are the challenges?
  • FDA, academia, private sector, patient advocacy
• Use case examples
• Demos and hands on
• BioCompute App-a-thon on PrecisionFDA
Use Case Examples

• **Tuberculosis Detection**
  • Tuberculosis (TB) is top infectious killer in the world
  • WHO is adopting ReSeqTB pipeline to address the many challenges of detecting TB
  • Requires lineage identification, prediction of antibiotic resistance, recurrence of TB in previously treated patients

• **Test Submission**
  • HCV-1a use case using synthesized data
  • What data are necessary to make a regulatory decision?
  • Are summary data from one analysis pipeline sufficient?
  • How will the analysis pipeline be validated?

• **Embleema**
  • Embleema is a platform that allows users to take control of their own data
  • Marketplace for directly selling personal genome data
  • Aggregator for Real World Evidence
• BioCompute standalone Editor
• BioCompute integration into HIVE
• BioCompute integration into Galaxy
• BioCompute integration into Seven Bridges
Timeline of Major Events

1st BioCompute Workshop
March, 2014

1st BioCompute Publication
December, 2014

FDA Funding Begins
May, 2015

2nd BioCompute Workshop
March, 2017

Collaboration with FHIR/Harvard Begins
September, 2016

2nd BioCompute Workshop
March, 2018

Use Case Tests
June, 2017

3rd BioCompute Workshop
March, 2018

FDA Tests
October, 2018

2nd BioCompute Publication
December, 2018

Public Private Partnership Established
May, 2019

Today

4th Workshop
May, 2019

BioCompute IEEE Standard
Q3, 2019 (est.)

BioCompute ISO Standard
Q4, 2019 (est.)
Future Plans

• BCO writer integrated into more platforms
• BCO reader functionality
• BCO execute
• BCO cross-platform functionality
  • Requires containerization
• BCO databases
  • Cancer specific database
IEEE Standards Development Process

1. Initiating the Project
2. Mobilizing the Working Group
3. Drafting the Standard
4. Balloting the Standard
5. Gaining Final Approval
6. Maintaining the Standard

Now: Comment Resolution
International Standards Organization

- Increase reach
- IEEE and ISO joint agreement
  - 7 month estimated timeline, if accepted
  - SC32 WG2 currently reviewing
    - Convened by Denise Warzel
BioCompute Public-Private Partnership

• Develop a community of stakeholders with interests in creating a versatile data harmonization framework that allows the standardized definition of platform-independent bioinformatics pipelines for execution, and is easily read by humans and machines.

• Facilitate the development of tools and facilities implementing data typing, instantiation, deposition, storage, and distribution of validated BioCompute Objects through a BioCompute database, in order to enable reproducible scientific research and regulatory submissions of data and computations.

• Facilitate portability of pipelines for execution.
Executive Steering Committee

Raja Mazumder (Chair)
Dennis Dean
Carole Goble
Jonathon Keeney (Managing Director)

Vahan Simonyan
Gil Alterovitz
Jeremy Goecks
PrecisionFDA partnered with George Washington University and the FDA HIVE to launch the BioCompute Object (BCO) App-a-thon

- Participants are be given the opportunity to innovate and standardize BCO creation and conformance
- Beginning and advanced tracks will allow individuals from varying levels to participate

App-a-thon launching May 14th 2019!
BioCompute Object (BCO) App-a-thon

App-a-thon Goals:

• Attract novices and advanced programmers alike (beginner and advanced tracks)

• Provide exposure to those unfamiliar with BioCompute, and highlight its utility in communicating bioinformatics pipelines

• Build examples of ways that the BCO specification can be integrated into other tools (e.g., a platform)

• Create a database of BCOs that others can freely use
BioCompute Object (BCO) App-a-thon
May 14 through October 18

**Beginner Track:**
Create BCOs

**Advanced Track:**
Create BCO software tools

**App-a-thon Launch:**
May 14, 2019
Challenge details are made available

May 14 – October 18, 2019
Participants create BCOs and BCO tools

**BCOs and Tools**
- Complete BCO
- Conformance Tool
- Writer Tool

**Conformance Check***
- Ability of an FDA reviewer to understand the BCO
- Ability to generate accurate BCO; Ability to conduct conformance check on BCOs

**Deadline:**
October 18, 2019
Evaluation begins

---

* Complete conformance check details are available on challenge site
BioCompute Object (BCO) App-a-thon

App-a-thon Goals:

• Attract novices and advanced programmers alike (beginner and advanced tracks)

• Provide exposure to those unfamiliar with BioCompute, and highlight its utility in communicating bioinformatics pipelines

• Build examples of ways that the BCO specification can be integrated into other tools (e.g., a platform)

• Create a database of BCOs that others can freely use
BioCompute Object (BCO) App-a-thon

Submissions to the Beginner Track must:

• Documents a multi-step bioinformatics pipeline in JSON file format

• Conforms to current BioCompute Specifications
  o [https://github.com/biocompute-objects/BCO_Specification/tree/1.3.0](https://github.com/biocompute-objects/BCO_Specification/tree/1.3.0)

• Can be shown publicly
Submissions to the Advanced Track must:

• A tool with the ability to do the following:
  o Create a BCO
  o Check a BCO for conformance
  o Display a BCO

• A detailed User manual (README)

• The application may be uploaded to precisionFDA or submitted as code
App-a-thon site
https://precision.fda.gov/challenges/7

precisionFDA contact information:
Elaine Johanson, Acting Director
Office of the Chief Scientist/Office of Health Informatics
precisionfda@fda.hhs.gov