BIO390: Introduction to Bioinformatics

Lecture I: What is Bioinformatics?

Michael Baudis | 2021-09-21

Course Information BIO390

- Tuesdays at 08:00; 2x45min
- 13 presentations by different lecturers
- (unchecked) homework / preparation exercises w/ focus on test topics
- course language is English
- course slides may/should be made available through the website
- written exam at end of course (== 14th course - December 21st)
- Organizer:

Prof. Dr. Michael Baudis

Department of Molecular Life Sciences (IMLS)

University of Zurich Campus Irchel, Y-13F-01

CH-8057 Zurich

email michael@baud.is

web info.baudisgroup.org

Please use website & OLAT for additional course information

https://compbiozurich.org/UZH-BIO390/

UZH BIO390

Introduction to Bioinformatics

UZH BIO390 - Introduction to Bioinformatics Lecture Series

This is a repository for materials related to the BIO390 *Introduction to Bioinformatics* lecture series at the University of Zürich.

Programme

Please see the listing of lectures here.

Time & Place

- 1 x 2h / week
- Tue 08:00-09:45
- UZH Irchel campus, **Y03-G-85**
- OLAT lecture recordings

Course Language

- English

2021 COVID19 Regulations

As per current rules for labs/block courses, BIO392 is held with “in person” attendance. Please follow the UZH Coronavirus Guidelines.

We provide lecture recordings through OLAT. In the case of changing situation with canceled on-site attendance or inability of lecturers to attend in person a ZOOM link will be provided.

Summary

The handling and analysis of biological data using computational methods has become an essential part in most areas of biology. In this lecture, students will be introduced to the use of bioinformatics tools and methods in different topics, such as molecular resources and databases, standards and ontologies, sequence and high performance genome analysis, biological networks, molecular dynamics, proteomics, evolutionary biology and gene regulation. Additionally, the use of low level tools (e.g. Programming and scripting languages) and specialized applications will be demonstrated. Another topic will be the visualization of quantitative and qualitative biological data and analysis results.

Learning Goals

The overall learning goals - especially the (limited) set necessary for passing the test - will be updated throughout the semester.

- Core **Learning Goals**

This list is not exhaustive; additional information about “need to know” topics will be provided during the individual lectures.

Literature and Resources

- Literature links and recommendations
- Resource links (browsers and online repositories)

University of Zurich[UZH]

https://compbiozurich.org/UZH-BIO390/

UZH BIO390

Introduction to Bioinformatics

News and Updates
- BIO390 HS21 Programme
- BIO390 HS20 Programme
- BIO390 HS19 Final Programme

Lectures

Teachers

Examples, Guides & FAQ

Related Sites

- CompbioZurich
- UZH392 course
- Baudisgroup at UZH

Github Projects

- compbiozurich
- progenetix

Tags

2020 2021 FAQ Markdown code days documentation exam teachers website

University of Zurich[UZH]

BIO390 HS21 Programme

In the following you will find the program for the 2020 Autumn lectures in the “Introduction to Bioinformatics” series at the University of Zurich.

For on-site access information or live streaming please see the relevant information in UZH’s “OLAT” platform (registered participants only):

- OLAT BIO390 HS21

Individual Lectures

2021-09-21

What is Bioinformatics? Introduction and Resources

Michael Baudis

The first day of the “Introduction to Bioinformatics” lecture series starts with a general introduction into the field and a description of the lecture topics, timeline and procedures.

@mbaudis 2021-09-21: more ...

2021-09-28

Biological Sequence Informatics

Christian von Mering

@mbaudis 2021-09-28: more ...

2021-10-05

Statistical Bioinformatics

Mark Robinson

@mbaudis 2021-10-05: more ...

2021-10-12

Machine Learning for Biological Use Cases

Valentina Boeva (ETHZ)

Brief note: In this lecture V. Boeva will cover the standard machine learning methods used in the analysis of biological data: dimensionality reduction, clustering, classification and regression.

2021-10-12: more ...

2021-10-19

Regulatory Genomics and Epigenomics

Izaskun Mallona

@mbaudis 2021-10-19: more ...

2021-10-26

https://compbiozurich.org/UZH-BIO390/

UZH BIO390 - Learning Goals

This page indicates some of the learning goals, as emphasised by the different lecturers. Some points will have been discussed in different lectures; accordingly, exam questions may not refer to information of one specific presentation.

Bioinformatics: Definition & Concepts

- definition of “Bioinformatics” (cf. Anna Tramontano)
- categories of informatics tools used in bioinformatics
- hypothesis versus data driven science
- areas of bioinformatics/bioinformaticians, in contrast to (“pure” modelling, statistics etc.)
- 3 main categories of biological data, and example resources
- definition of API
- common sequence related file formats
- hierarchies and relationships as 2 main principles of ontologies
- areas of “not-bioinformatics”, and why

Sequence Analysis

- substitution matrices
- BLAST

Statistical Bioinformatics

- statistical evidence for a change in the means
- usage of gene expression profiling
- dimensionality reduction
- central limit theorem
- multiple testing correction
- parameters for hierarchical clustering

Bioinformatics tools: Statistics & Graphics in R & BioConductor

- What is tidy data?
- ideas behind ggplot: components of a ggplot, arrangement of input data … (no actual code writing needed)
- interpret common types of plots, e.g. barplot, boxplot, histogram
- effect of data transformation (e.g. log) on common types of plots

Regulatory Genomics and Epigenomics

- secondary/tertiary human genome structure
- functional genome content
- transcription factors & genome interaction
- chemical genome modifications, their effectors and results
- Chip-Seq

University of Zurich[UZH]

https://compbiozurich.org/UZH-BIO390/

Some Recommended Books

- Anna Tramontano: Introduction to Bioinformatics
- Susan Holmes and Wolfgang Huber: Statistics for Biology
- Robert Gentleman: R Programming for Bioinformatics
- John Maindonald & W. John Braun: Data Analysis and Graphics Using R
- Andy Hector: The New Statistics with R
- Neil C. Jones & Pavel A. Pevzner: Bioinformatics Algorithms
- Edward Tufte: The Visual Display of Quantitative Information (& other works by Tufte)

BIO390: Course Schedule

- 2021-09-21: Michael Baudis - What is Bioinformatics? Introduction and Resources
- 2021-09-28: Christian von Mering - Sequence Bioinformatics
- 2021-10-05: Mark Robinson - Statistical Bioinformatics
- 2021-10-12: Valentina Boeva (ETHZ) - Machine Learning for Biological Use Cases
- 2021-10-19: Izaskun Mallona - Regulatory Genomics and Epigenomics
- 2021-10-26: Shinichi Sunagawa (ETHZ) - Metagenomics
- 2021-11-02: Katja Baerenfaller (SIAF) - Proteomics
- 2021-11-09: Puria Dasmeh - Biological Networks
- 2021-11-16: TBD - Text Mining topic
- 2021-11-23: Fabio Rinaldi - Text Mining
- 2021-11-30: Michael Baudis - Building a Genomics Resource
- 2021-12-07: Valérie Barbie (SIB) - Clinical Bioinformatics
- 2021-12-14: Michael Baudis - Genome Data & Privacy
- 2021-12-21: Exam (Multiple Choice)

Why Bioinformatics?

- **hypotheses** are the basis of biological experiments
- biological experiments produce **data**, the quantitative and/or qualitative read-outs of experiments
- both quantitative as well as qualitative data need to be **processed** for
  - **statistical significance**
  - **categorisation**
  - **communication**
- many datatypes are **beyond** the proverbial "**intuitive** understanding”
- analysis of data **confirms** or **refutes** initial **hypotheses** - or requires new hypotheses and new data

Biology is *impossibly complex* - But bioinformatics might help

Randall Munroe: https://xkcd.com/1605/

So, What is Bioinformatics?

- Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to formulate hypotheses about life." (Anna Tramontano)

What is Bioinformatics?

science

noun | sci·ence | \ˈsī-ən(t)s\

Popularity: Top 1% of lookups

- Bioinformatics is "the **science** that uses the instruments of informatics to analyze biological data in order to formulate hypotheses about life." (Anna Tramontano)

Merriam-Webster

a : knowledge or a system of knowledge covering general truths or the operation of general laws especially as obtained and tested through scientific method

b : such knowledge or such a system of knowledge concerned with the physical world and its phenomena : NATURAL SCIENCE

What is Bioinformatics?

Bioinformatics **uses** informatics tools for analyses

- Bioinformatics is "the science that **uses** the instruments of informatics to analyze biological data in order to formulate hypotheses about life." (Anna Tramontano)

➡ **software** (programming languages, statistics & visualisation, program and web APIs, databases, hardware drivers)
➡ **hardware** (HPC, data storage, signal measurement & processing)
➡ **algorithms** (modeling, encryption...)

What is Bioinformatics?

Bioinformatics **develops** informatics tools for analyses

- Bioinformatics is "the science that uses the **instruments of informatics** to analyze biological data in order to formulate hypotheses about life." (Anna Tramontano)

➡ **software** (statistics & visualisation packages, program and web APIs, file formats)
➡ **hardware** (drivers and procedures...)
➡ **algorithms** (modeling, encryption...)

What is Bioinformatics?

biological data

- Bioinformatics is "the science that uses the instruments of informatics to analyze **biological data** in order to formulate hypotheses about life." (Anna Tramontano)

sequences, graphs, high-dimensional data, spatial/geometric information, scalar and vector fields, patterns, constraints, images, models, prose, declarative knowledge ... *

adapted from "Catalyzing Inquiry at the Interface of Computing and Biology"; John C Wooley and Herbert S Lin. National Research Council (US) Committee on Frontiers at the Interface of Computing and Biology. Washington (DC): National Academies Press (US); 2005.

What is Bioinformatics?

analyze

verb | an·a·lyze | \ˈa-nə-ˌlīz\

Popularity: Top 1% of lookups

Bioinformatics **analyzes**

- Bioinformatics is "the science that uses the instruments of informatics to **analyze** biological data in order to formulate hypotheses about life." (Anna Tramontano)

1 : to study or determine the nature and relationship of the parts of (something) by analysis

What is Bioinformatics?

hypothesis

noun | hy·poth·e·sis | \hī-ˈpä-thə-səs\

Popularity: Top 1% of lookups

Merriam-Webster

- Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to **formulate hypotheses** about life." (Anna Tramontano)

b : an interpretation of a practical situation or condition taken as the ground for action

What is Bioinformatics?

hypothesis driven science

data driven science

hypothesis

noun | hy·poth·e·sis | \hī-ˈpä-thə-səs\

Popularity: Top 1% of lookups

Merriam-Webster

- Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to **formulate hypotheses** about life." (Anna Tramontano)

b : an interpretation of a practical situation or condition taken as the ground for action

42

- Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to formulate hypotheses about **life**." (Anna Tramontano)

Sources: nextprot | opentreeoflife | wikipedia | MacKay et al., Cancer Cell (2017) | original photos

Homework Assignment - How can Bioinformatics help with the *42* of Life Sciences?

42

- Bioinformatics is "the science that uses the instruments of informatics to analyze biological data in order to formulate hypotheses about **life**." (Anna Tramontano)

Sources: nextprot | opentreeoflife | wikipedia | MacKay et al., Cancer Cell (2017) | original photos

{bio_informatics_science}

{bio_informatics_science}

Randall Munroe - XKCD
https://xkcd.com/1319/

Bioinformatician	Bio**informatician**
strong biological knowledge	sufficient biological background
provides hypothesis and / or dataset	provides statistical, analysis methods
sufficient statistical and **computational** expertise to correctly use bioinformatics tools & develop workflows (scripting ...)	**sufficient biological** or **medical** background to understand problems presented and identify pitfalls and hidden biases arising from data generation
expert **user** of informatics tools	**developer** of informatics tools
may get a Nobel	may get rich

Bioinformatician	Bio**informatician**
strong biological knowledge	sufficient biological background
provides hypothesis and / or dataset	provides statistical, analysis methods
sufficient statistical and **computational** expertise to correctly use bioinformatics tools & develop workflows (scripting ...)	**sufficient biological** or **medical** background to understand problems presented and identify pitfalls and hidden biases arising from data generation
expert **user** of informatics tools	**developer** of informatics tools
may get a Nobel	may get rich

flux

Who is a Bioinformatician?

Randall Munroe: https://xkcd.com/1605/

bioinformatician

bioinformatician

What do Bioinformaticians work on?

- protein **structure** definition
- DNA/RNA/protein **sequence** analysis
- **quantitative** analysis of "-omics" and cytometry data
- **functional** enrichment of target data (e.g. genes, sequence elements)
- **evolutionary** reconstruction and "tree of life" questions
- **image processing** for feature identification and spatial mapping
- **statistical** analysis of measurements and observations
- **protocols** for efficient storage, annotation and retrieval of biomedical data
- **information extraction** from prose & declarative knowledge resources (think publications & data tables)
- **clinical** bioinformatics - risk assessment and therapeutic target identification
- ...

Bioinformatics: Data **Categories** & **Databases**

- biological data comes in **3 main categories**:
  - **sequence** data (nucleic acids, aminoacids)
  - **structural** data (DNA, RNA, proteins; intracellular organisation, tissues ...)
  - **functional** data (interactions in time and space)
- data storage & retrieval: importance of local and connected **databases**
  - **primary databases** - for deposition of original, raw data (e.g. SRA - sequence read archive; ENA - European Nucleotide Archive; GEO - NCBI Gene Expression Omnibus; EBI arrayExpress...)
  - **derived databases** - information resources providing agglomerated & **curated** data derived from primary sources (e.g. UniprotKB, nextProt, String, KEGG, arrayMap...)

SRA

arrayMap

ENA European Nucleotide Archive

nextprot

STRING

UniProt

DATA PIPELINE

arrayMap

progenetix

DATA PIPELINE

Bioinformatics & Data Curation - arrayMap data “Pipeline”

Bioinformatics: File Formats, Ontologies & APIs

- text or binary file formats, optimised for specific types of biological data
- examples from genomics:
  - **BAM** - compressed binary version of Sequence Alignment/ Map (SAM)
  - **BED** (Browser Extensible Data) -flexible way to define the data lines in an genome browser annotation tracks
  - **VCF** (Variant Call Format)

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration" itemRgb="On"
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7    127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7    127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7    127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7    127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7    127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7    127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

BED file example

File Formats: VCF

Genomic variant storage standard

- The VCF Variant Call Format is an example for a widely used file format with "built-in logic"
- has been essential to master the "genomics data deluge" through providing "logic compression" for genomic annotations which rely on the notion of "assessed variant in a population"
- very expressive, but complex interpretation
- mix of "observed" and "population" variant concepts confusing for some use cases
- no replacement in sight (but new versions)

The Variant Call Format (VCF) Version 4.2 Specification

25 Jun 2020

The master version of this document can be found at `https://github.com/samtools/hts-specs`.
This printing is version 09fbcec from that repository, last modified on the date shown above.

1 The VCF specification

VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position.

1.1 An example

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Bioinformatics: File Formats, Ontologies & APIs

- databases can be accessed through **A**pplication **P**rogramming **I**nterfaces
- ***API*** *: set of routines, protocols, and tools that specifies how software components interact, to exchange data and processing capabilities*
- web API example: implementing geographic maps, with parameters provided by the client (e.g. location coordinates, quantitative payload)
- web APIs provide a *machine readable* response to queries over HTTP
- bioinformatic applications frequently make use of web APIs for **data retrieval** or genome browser APIs for **data display**
- bioinformatics software libraries for API functionality are usually implemented in **Perl**, **Python** and/or **R**

Bioinformatics: File Formats, Ontologies & APIs

{“api_params":{"genome":"hg18", "db":"progenetix", "datatype":"sampledata", "count":20,
"call":"api_doctype=json&api_out=samples&db=progenetix&icdm_m=817&randno=20", "scope":"samples"},
"data":[{"ICDMORPHOLOGY":"Liver cell adenoma", "NCIT:CODE":"C3758", "ICDTOPOGRAPHYCODE":"C22", "_id":
{"$oid":"558e5c2ead9a82d95838f76c"}, "CLINICALGROUP":"Carcinomas: hepatic ca.", "PMID":"15765123",
"FOLLOWUP":"", "BIOSAMPLEID":"AM_BS__HKCI-C2-DOR", "ICDMORPHOLOGYCODE":"8170/0",
"NCIT:TERM":"Hepatocellular Adenoma", "UID":"HKCI-C2-DOR", "DIAGNOSISTEXT":"Hepatocellular carcinoma
[cell line, doxorubicin resistant subclone]", "DEATH":"", "ICDTOPOGRAPHY":"liver", "AGE":""},
{"FOLLOWUP":"", "PMID":"14578863", "ICDMORPHOLOGYCODE":"8170/3", "BIOSAMPLEID":"AM_BS__PHCC-30",
"ICDMORPHOLOGY":"Hepatocellular carcinoma, NOS", "_id":{"$oid":"558e5c36ad9a82d9583901bf"},
"CLINICALGROUP":"Carcinomas: hepatic ca.", "ICDTOPOGRAPHYCODE":"C22", "NCIT:CODE":"C3099",
"ICDTOPOGRAPHY":"liver", "DEATH":"", "DIAGNOSISTEXT":"Hepatocellular carcinoma", "AGE":"",
"NCIT:TERM":"Hepatocellular Carcinoma", "UID":"PHCC-30"}, {"DEATH":"",
"DIAGNOSISTEXT":"Hepatocellular carcinoma [chronic Hepatitis B]", "ICDTOPOGRAPHY":"liver",
"AGE":"P54Y2M", "NCIT:TERM":"Hepatocellular Carcinoma", "UID":"HCC-1997-14", "PMID":"8993981",
"FOLLOWUP":"", "BIOSAMPLEID":"AM_BS__HCC-1997-14", "ICDMORPHOLOGYCODE":"8170/3",
"ICDMORPHOLOGY":"Hepatocellular carcinoma, NOS", "NCIT:CODE":"C3099", "ICDTOPOGRAPHYCODE":"C22",
"_id":{"$oid":"558e5bfcad9a82d95838b62a"}, "CLINICALGROUP":"Carcinomas: hepatic ca."},
{"FOLLOWUP":"", "PMID":"11485905", "BIOSAMPLEID":"AM_BS__HCChypo-won-H18",
"ICDMORPHOLOGYCODE":"8170/3", "ICDMORPHOLOGY":"Hepatocellular carcinoma, NOS",
"CLINICALGROUP":"Carcinomas: hepatic ca.”, "_id":{"$oid":"558e5c48ad9a82d95839185a"},
"NCIT:CODE":"C3099", "SEX":"male" . . .

http://progenetix.org/api/?db=progenetix&api_out=samples&api_doctype=json&icdm_m=817&randno=20

BeaconAlleleRequest beacon ↗

{S}[B] Status [i]	implemented
Provenance	Beacon API
Used by	Beacon Progenetix database schema (Beacon+ backend)
Contributors	Marc Fiume Michael Baudis Sabela de la Torre Pernas Jordi Rambla Beacon developers...
Source (v1.1.0)	raw source [JSON] Github

Attributes

Type: object
Description: Allele request as interpreted by the beacon.

Properties

Property	Type
alternateBases	string
assemblyId	string
datasetIds	array of string
end	integer
endMax	integer
endMin	integer
mateName	https://schemablocks.org/schemas/beacon/v1.1.0/Chro [HTML]
referenceBases	string
referenceName	https://schemablocks.org/schemas/beacon/v1.1.0/Chro [HTML]
start	integer (int64)
startMax	integer
startMin	integer
variantType	string

alternateBases

- type: string

The bases that appear instead of the reference bases. Accepted values: [ACGTN]*. N is a wildcard, that denotes the position of any base, and can be used as a standalone base of any type or within a partially known sequence. For example a sequence where the first and last bases are known, but the middle portion can exhibit countless variations of [ACGT], or the bases are unknown: ANNT the Ns can take take any form of [ACGT], which makes both ACCT and ATGT (or any other combination) viable sequences.

Symbolic ALT alleles (DEL, INS, DUP, INV, CNV, DUP:TANDEM, DEL:ME, INS:ME) will be represented in `variantType`.

Optional: either `alternateBases` or `variantType` is required.

`alternateBases` **Value Example**

assemblyId

- type: string

Assembly identifier (GRC notation, e.g. GRCh37).

`assemblyId` **Value Example**

Curie sb-vr-spec ↗

{S}[B] Status [i]	implemented
Provenance	vr-spec
Used by	vr-spec
Contributors	Reece Hart Michael Baudis
Source (v1.0)	raw source [JSON] Github

Attributes

Type: string
Pattern: ^\w[^:]+:.+$
Description: A string that refers to an object uniquely. T sender.

VR does not impose any contraints on strings used as ids data, the VR Specification RECOMMENDS that implement

String CURIEs are represented as `prefix:reference` (W `namespace:accession` or `namespace:local id` colloqu

The VR specification also RECOMMENDS that `prefix` be

The `reference` component is an unconstrained string.

A CURIE is a URI. URIs may *locate* objects (i.e., specify wh VR uses CURIEs primarily as a naming mechanism.

Implementations MAY provide CURIE resolution mechan

Using internal ids in public messages is strongly discoura

`Curie` **Value Examples**

```
"ga4gh:GA.01234abcde"
```

```
"DUO:0000004"
```

```
"orcid:0000-0003-3463-0775"
```

```
"PMID:15254584"
```

Biosample sb-phenopackets ↗

{S}[B] Status [i]	implemented
Provenance	Phenopackets
Used by	Phenopackets
Contributors	GA4GH Data Working Group Jules Jacobsen Peter Robinson Michael Baudis Melanie Courtot Isuru Liyanage
Source (v1.0.0)	raw source [JSON] Github

Attributes

Type: object
Description: A Biosample refers to a unit of biological material from which the substrate molec genomic DNA, RNA, proteins) for molecular analyses (e.g. sequencing, array hybridisation, mas spectrometry) are extracted.

Examples would be a tissue biopsy, a single cell from a culture for single cell genome sequenci fraction from a gradient centrifugation.

Several instances (e.g. technical replicates) or types of experiments (e.g. genomic array as well experiments) may refer to the same Biosample.

FHIR mapping: Specimen.

Properties

Property	Type
ageOfIndividualAtCollection	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/Age.json [SRC] [HTML]
ageRangeOfIndividualAtCollection	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/AgeRange.json [SRC] [HTML]
description	string
diagnosticMarkers	array of https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/OntologyClass.json [SRC] [HTMI
histologicalDiagnosis	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/OntologyClass.json [SRC] [HTMI
htsFiles	array of https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/HtsFile.json [SRC] [HTML]
id	string
individualId	string
isControlSample	boolean
phenotypicFeature	array of https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/PhenotypicFeature.json [SRC] [HTML]
procedure	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/Procedure.json [SRC] [HTML]
sampledTissue	https://schemablocks.org/schemas/sb-

Checksum sb-checksum ↗

{S}[B] Status [i]	proposed
Provenance	GA4GH DRS (`develop` branch)
Used by	GA4GH DRS GA4GH TRS
Contributors	Susheel Varma
Source (v0.0.1)	raw source [JSON] Github

Attributes

Type: object
Description: Checksum

Properties

Property	Type
checksum	string
type	string

checksum

- type: string

The hexadecimal encoded (Base16) checksum for the data

`checksum` **Value Example**

```
"77af4d6b9913e693e8d0b4b294fa62ade6054e6b2f1ffb617ac955dd63fb0182"
```

type

- type: string

The digest method used to create the checksum. The value (e.g. `sha-256`) SHOULD be listed as `Hash Name String` in the GA4GH Hash Algorithm Registry. Other values MAY be used, as long as implementors are aware of the issues discussed in RFC6920.

GA4GH may provide more explicit guidance for use of non-IANA-registered algorithms in the future.

`type` **Value Example**

```
"sha-256"
```

Bioinformatics: File Formats, Ontologies & APIs

- ontologies in information sciences describe concrete and abstract **objects**, there precisely defined **hierarchies** and **relationships**
- ontologies in bioinformatics support the move from a descriptive towards an **analytical science** in describing biological data and relations among it

> "The widest use of ontologies within biology is for conceptual annotation – a representation of stored knowledge more computationally amenable than natural language."*

- Gene ontology (GO)
- NCIt Neoplasm Core
- Uberon anatomical structures
- Experimental Factor Ontology (EFO)
- Disease Ontology (DO)

Ontologies in bioinformatics. R Stevens, C Wroe, P Lord and C Goble, Department of Computer Science, University of Manchester. Retrieved from http://homepages.cs.ncl.ac.uk/phillip.lord/download/publications/handbook.pdf

Standardized Data

Data re-use depends on standardized, machine-readable metadata

- Multiple international initiatives (ELIXIR, GA4GH, MONARCH...) and resource providers (EBI, NCBI ...) work on the generation and implementation of data annotation standards
- emerging / established principles are the use of **hierarchical** coding systems where individual codes are represented as CURIEs
- other formats for non-categorical annotations based on international standards, e.g.
  - ISO (ISO 8601 time & period, ISO 3166 country codes ...)
  - IETF (GeoJSON ...)
  - W3C (CURIE ...)
- these standards become pervasive throughout GA4GH's ecosystem (e.g. Phenopackets ...)


```
"data_use_conditions" : {
  "label" : "no restriction",
  "id" : "DUO:0000004"
},

"provenance" : {
  "material" : {
    "type" : {
      "id" : "EFO:0009656",
      "label" : "neoplastic sample"
    }
  },
  "geo" : {
    "label" : "Zurich, Switzerland",
    "precision" : "city",
    "city" : "Zurich",
    "country" : "Switzerland",
    "latitude" : 47.37,
    "longitude" : 8.55,
    "geojson" : {
      "type" : "Point",
      "coordinates" : [
        8.55,
        47.37
      ]
    },
    "ISO-3166-alpha3" : "CHE"
  }
},
{
  "age": "P25Y3M2D"
}
```

Standardized Data

Data re-use depends on standardized, machine-readable metadata

```
"data_use_conditions" : {
  "label" : "no restriction",
  "id" : "DUO:0000004"
},

"provenance" : {
  "material" : {
    "type" : {
      "id" : "EFO:0009656",
      "label" : "neoplastic sample"
    }
  },
  "geo" : {
    "label" : "Zurich, Switzerland",
    "precision" : "city",
    "city" : "Zurich",
    "country" : "Switzerland",
    "latitude" : 47.37,
    "longitude" : 8.55,
    "geojson" : {
      "type" : "Point",
      "coordinates" : [
        8.55,
        47.37
      ]
    },
    "ISO-3166-alpha3" : "CHE"
  }
},
{
  "age": "P25Y3M2D"
}
```

Data Science: Meta-Studies of Metadata

Publication Landscape of Cancer CNV Profiling

Publication statistics for cancer genome screening studies. The graphic shows our assessment of publications reporting whole-genome screening of cancer samples, using molecular detection methods (chromosomal CGH, genomic array technologies, whole exome and genome sequencing).

For the years 1993-2018, we found 3'229 publications reporting 174'530 individual samples in single series from 1 to more than 1000 samples. Y-axis and size of the dots correspond to the sample number; the color codes indicate the technology used.

Map of the geographic distribution (by first author affiliation) of the 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets.

The numbers are derived from the 3'240 publications registered in the Progenetix database.

progenetix

But: What is not bioinformatics, though being "bio" and using computers?

- *"I do not think all biological computing is bioinformatics, e.g. **mathematical modelling** is not bioinformatics, even when connected with biology-related problems. In my opinion, bioinformatics has to do with management and the subsequent use of biological information, particular genetic information."* (Richard Durbin)
- **biologically-inspired computation** (neural networks etc.) - though their application may be part of bioinformatics
- **computational & systems biology**, where the emphasis is on **modelling** rather than on **data interpretation**

Bioinformatics OR **Computational / Systems Biology?**

- **Bioinformatics**

  Research, development, or **application** of computational **tools** and approaches to make the vast, diverse and complex **life sciences data** more understandable and useful

- **Computational biology**

  The development and application of **mathematical** and computational **approaches** to address **theoretical** and experimental questions in biology

source: Colin Dewey (UWisc): "Introduction to Bioinformatics Fall 2016"

xkcd

MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

Permanent link to this comic: https://xkcd.com/1987/

BIO390: Introduction to Bioinformatics

Lecture I: What are Bioinformaticians doing? Example of Developing "Federated Human Data" concepts

Michael Baudis | 2021-09-21

The trouble with human genome variation

Conclusions from the analysis of variation in the human genome

- 1. Humans are all very similar to each other
  - Two humans will show about 99.9% sequence identity with each other. In other words, only about 1 in 1’000 bp is different between two individuals.
  - Humans show about 98% sequence identity to chimps. So two humans are still much more similar to each other than either is to the monkey.

- 2. Humans are very different from each other
  - Two typical humans will likely have over 1’000’000 independent sequence differences in their genomes.

Slide from Michael Hengartner’s Human Genome Lecture

Finding Somatic Mutations In Cancer

Many Needles in a Large Haystack

- a typical humen genome (~3 billion base pairs) has ~5 million variants
- most of them are "**rare**"; i.e. can only be identified as recurring when sequencing thousands of people
- cancer cells accumulate additional variants, only **few** of which ("**drivers**") are relevant for the disease

* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

Graphic adapted from Mark Gerstein (GersteinLab.org; @markgerstein)

Federation of human genome data

- Many national datasets from human research participants needs to be stored locally
- ELIXIR developing a federation with shared metadata (FAIR) and local data store (secure)
- Linking local EGA to national clouds – and international access (ELIXIR-AAI)

GENOMICS

A federated ecosystem for sharing genomic, clinical data

Silos of genome data collection are being transformed into seamlessly connected, independent systems

The Global Alliance for Genomics and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

A federated data ecosystem. To share genomic data globally, this approach furthers medical research without requiring compatible data sets or compromising patient identity.

DNASTACK

Global Alliance for Genomics & Health

A ***Beacon*** answers a query for a specific genome variant against individual or aggregate genome collections

YES | **NO** | **\0**

ELIXIR Beacon Network

- developed under lead from ELIXIR Finland
- **authenticated access** w/ ELIXIR AAI
- **incremental extension**, starting with ELIXIR Beacon resources adhering to the **latest specification** (contrast to legacy networks)
- service details provided by individual Beacons, using **GA4GH service-info**
- **registration service**

➡**integrator** throughout ELIXIR Human Data

➡starting point for "**beyond ELIXIR**"

feature rich federated Beacon **services**

Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

Beacon *v2* API

The Beacon API v2 proposal opens the way for the design of a simple but powerful "**genomics API**".

Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

Beacon *v2* API

The Beacon API v2 proposal opens the way for the design of a simple but powerful "**genomics API**".

Vague but exciting ...

CERN DD/OC — Tim Berners-Lee, CERN/DD

Information Management: A Proposal — March 1989

Information Management: A Proposal

Abstract

This proposal concerns the management of general information about accelerators and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext sytstem.

Keywords: Hypertext, Computer conferencing, Document retrieval, Information management, Project control

Tim Berners-Lee: Information Management: A Proposal (CERN 1989) & WWW: First Page (1990)

World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

What's out there?
: Pointers to the world's online information, subjects , W3 servers, etc.

Help
: on the browser you are using

Software Products
: A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)

Technical
: Details of protocols, formats, program internals etc

Bibliography
: Paper documentation on W3 and references.

People
: A list of some people involved in the project.

History
: A summary of the history of the project.

How can I help ?
: If you would like to support the web..

Getting code
: Getting the code by anonymous FTP , etc.

BIO390: Course Schedule

- 2021-09-21: Michael Baudis - What is Bioinformatics? Introduction and Resources
- **2021-09-28: Christian von Mering - Sequence Bioinformatics**
- 2021-10-05: Mark Robinson - Statistical Bioinformatics
- 2021-10-12: Valentina Boeva (ETHZ) - Machine Learning for Biological Use Cases
- 2021-10-19: Izaskun Mallona - Regulatory Genomics and Epigenomics
- 2021-10-26: Shinichi Sunagawa (ETHZ) - Metagenomics
- 2021-11-02: Katja Baerenfaller (SIAF) - Proteomics
- 2021-11-09: Puria Dasmeh - Biological Networks
- 2021-11-16: TBD - Text Mining topic
- 2021-11-23: Fabio Rinaldi - Text Mining
- 2021-11-30: Michael Baudis - Building a Genomics Resource
- 2021-12-07: Valérie Barbie (SIB) - Clinical Bioinformatics
- 2021-12-14: Michael Baudis - Genome Data & Privacy
- 2021-12-21: Exam (Multiple Choice)

University of Zurich[UZH]

SIB

Prof. Dr. Michael Baudis
Institute of Molecular Life Sciences
University of Zurich
SIB | Swiss Institute of Bioinformatics
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

arraymap.org
progenetix.org
info.baudisgroup.org
sib.swiss/baudis-michael
imls.uzh.ch/en/research/baudis
beacon-project.io
schemablocks.org

Global Alliance
for Genomics & Health

elixir