Bioinformatik

Wintersemester 2019 / 2020
Universität Potsdam

Nachricht von Prof. J. Eccard:

Liebe BBW Studierende Fachrichtung organismische Biologie:

Das Wahlmodul "Tierökologie und Humanbiologie" wird in PULS nicht vollständig abgebildet. Neben den Vorlesungen Tierökologie (Do, 1200) und dem Seminar Aktuelle Themen aus Tierökologie und Humanbiologie (Di 1400) gehört auch noch die Vorlesung Humanbiologie (Di 1015) mit dazu, Sie finden diese zur Anmeldung unter dem Fach Ernährungswissenschaften.

Bioinformatik

Bachelor Biowissenschaften (Prüfungsversion ab WiSe 2017/18)
Pflichtmodul für alle Spezialisierungsrichtung, 6LP
4 LP Bioinformatik V/Ü
2 LP Molekulare Evolution V

PULS!

Bachelor Biowissenschaften (Prüfungsversion ab WiSe 2010/11)
Pflichtmodul für Spezialisierungsrichtung Biochemie
Pflichtmodul für Spezialisierungsrichtung Molekularbiologie/Physiologie
4 LP Bioinformatik V/Ü

Bachelor Informatik/Compuational Science
Wahlpflicht, 6 LP

PULS!

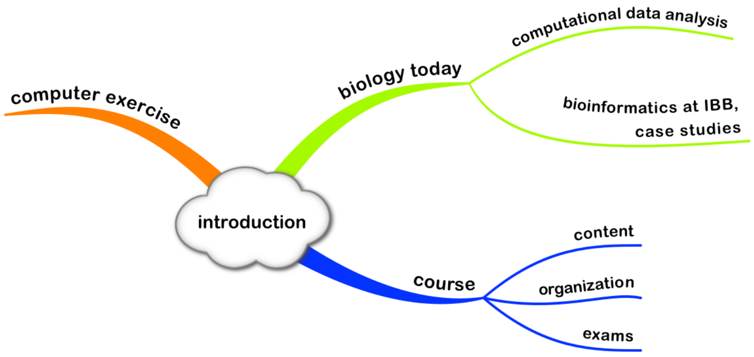
Bioinformatik

instructor: Stefanie Hartmann

teaching assistants:
Marie Gurke, Dennis Schlossarek, Viswa Bysani

Wintersemester 2019 / 2020
Universität Potsdam

Today's topics



introduction

computer exercise

biology today

computational data analysis

bioinformatics at IBB, case studies

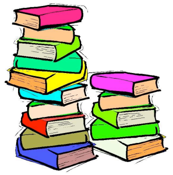
course

content

organization

exams

Science



- body of knowledge representing our current understanding of natural systems
- process that continually extends, refines, and revises that body of knowledge

...so you want to be a biologist! (I)

motivation

- you are interested in studying the natural world
- you want to do basic or applied science

you will need

- current content knowledge
  - biochemistry, molecular biology, physiology, genetics, ecology, evolution, ...
- skills
  - think critically, solve problems, collect data, analyze data, interprete data / results, collaborate and work independently, communicate results, read & analyze primary literature, and many many more!

Biology then

the roots

- organisms, properties
- observation & description
- taxonomy & classification
- evolutionary analysis

the first revolution: molecular biology

- provides detail about underlying mechanisms of life
  - DNA/heredity, ATP/energy, conserved sequences, ...
- reductionist approach
  - study phenomena at the level of molecules/genes

Biology now

the second revolution:

- high-throughput data acquisition
- information technology

bioinformatics: a set of tools

- use computers to manage & analyze data
- relevance for all biological disciplines

systems biology: a discipline

- study properties & interactions of biological systems in qualitative and quantitative manner
- experimental data + mathematical modelling

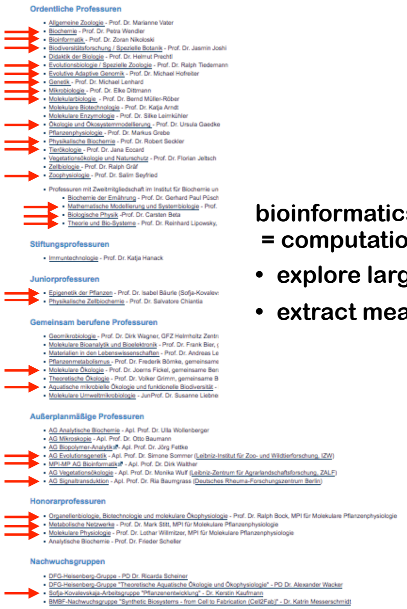
...so you want to be a biologist! (II)

biological data

- directly generated in electronic format
- converted into electronic format after collection
- available in electronic format online
- many different data types & formats exist!

computational data analysis

- which tools exist?
- which tools are appropriate for a data set / question?
- how do these tools work?
- what does the output look like?



Ordentliche Professuren

- Allgemeine Zoologie - Prof. Dr. Marianne Vater
- → Biochemie - Prof. Dr. Petra Wendler
- → Bioinformatik - Prof. Dr. Zoran Nikoloski
- → Biodiversitätsforschung / Spezielle Botanik - Prof. Dr. Jasmin Joshi
- Didaktik der Biologie - Prof. Dr. Helmut Prechtl
- → Evolutionsbiologie / Spezielle Zoologie - Prof. Dr. Ralph Tiedemann
- → Evolutive Adaptive Genomik - Prof. Dr. Michael Hofreiter
- → Genetik - Prof. Dr. Michael Lenhard
- → Mikrobiologie - Prof. Dr. Elke Dittmann
- → Molekularbiologie - Prof. Dr. Bernd Müller-Röber
- Molekulare Biotechnologie - Prof. Dr. Katja Arndt
- Molekulare Enzymologie - Prof. Dr. Silke Leimkühler
- → Ökologie und Ökosystemmodellierung - Prof. Dr. Ursula Gaedke
- Pflanzenphysiologie - Prof. Dr. Markus Grebe
- → Physikalische Biochemie - Prof. Dr. Robert Seckler
- → Tierökologie - Prof. Dr. Jana Eccard
- Vegetationsökologie und Naturschutz - Prof. Dr. Florian Jeltsch
- Zellbiologie - Prof. Dr. Ralph Gräf
- → Zoophysiologie - Prof. Dr. Salim Seyfried
- Professuren mit Zweitmitgliedschaft im Institut für Biochemie un
  - Biochemie der Ernährung - Prof. Dr. Gerhard Paul Püsch
  - → Mathematische Modellierung und Systembiologie - Prof.
  - → Biologische Physik -Prof. Dr. Carsten Beta
  - → Theorie und Bio-Systeme - Prof. Dr. Reinhard Lipowsky,

Stiftungsprofessuren

- Immuntechnologie - Prof. Dr. Katja Hanack

Juniorprofessuren

- → Epigenetik der Pflanzen - Prof. Dr. Isabel Bäurle (Sofja-Kovalev
- → Physikalische Zellbiochemie - Prof. Dr. Salvatore Chiantia

Gemeinsam berufene Professuren

- Geomikrobiologie - Prof. Dr. Dirk Wagner, GFZ Helmholtz Zentr
- Molekulare Bioanalytik und Bioelektronik - Prof. Dr. Frank Bier, g
- Materialien in den Lebenswissenschaften - Prof. Dr. Andreas Le
- Pflanzenmetabolismus - Prof. Dr. Frederik Börnke, gemeinsame
- → Molekulare Ökologie - Prof. Dr. Joerns Fickel, gemeinsame Ber
- Theoretische Ökologie - Prof. Dr. Volker Grimm, gemeinsame B
- → Aquatische mikrobielle Ökologie und funktionelle Biodiversität -
- Molekulare Umweltmikrobiologie - JunProf. Dr. Susanne Liebne

Außerplanmäßige Professuren

- AG Analytische Biochemie - Apl. Prof. Dr. Ulla Wollenberger
- AG Mikroskopie - Apl. Prof. Dr. Otto Baumann
- AG Biopolymer-Analytik- Apl. Prof. Dr. Jörg Fettke
- → AG Evolutionsgenetik - Apl. Prof. Dr. Simone Sommer (Leibniz-Institut für Zoo- und Wildtierforschung, IZW)
- → MPI-MP AG Bioinformatik - Apl. Prof. Dr. Dirk Walther
- AG Vegetationsökologie - Apl. Prof. Dr. Monika Wulf (Leibniz-Zentrum für Agrarlandschaftsforschung, ZALF)
- → AG Signaltransduktion - Apl. Prof. Dr. Ria Baumgrass (Deutsches Rheuma-Forschungszentrum Berlin)

Honorarprofessuren

- → Organellenbiologie, Biotechnologie und molekulare Ökophysiologie - Prof. Dr. Ralph Bock, MPI für Molekulare Pflanzenphysiologie
- → Metabolische Netzwerke - Prof. Dr. Mark Stitt, MPI für Molekulare Pflanzenphysiologie
- → Molekulare Physiologie - Prof. Dr. Lothar Willmitzer, MPI für Molekulare Pflanzenphysiologie
- Analytische Biochemie - Prof. Dr. Frieder Scheller

Nachwuchsgruppen

- DFG-Heisenberg-Gruppe - PD Dr. Ricarda Scheiner
- DFG-Heisenberg-Gruppe "Theoretische Aquatische Ökologie und Ökophysiologie" - PD Dr. Alexander Wacker
- → Sofja-Kovalevskaja-Arbeitsgruppe "Pflanzenentwicklung" - Dr. Kerstin Kaufmann
- BMBF-Nachwuchsgruppe "Synthetic Biosystems - from Cell to Fabrication (Cell2Fab)" - Dr. Katrin Messerschmidt

bioinformatics = computational biology
= computational data analysis

- explore large & complex data sets
- extract meaning from these data

Biological data: examples

sequences

- strings of nucleotides or amino acids

graphs (networks)

- metabolic pathways, regulatory networks, trees

high-dimensional data

- gene expression data, metabolites

geometric information

- molecular structure data

images

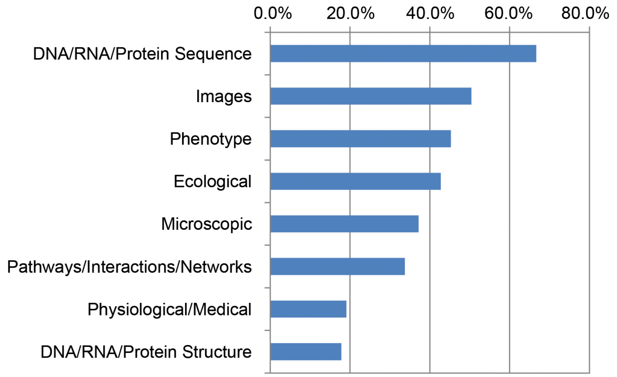
- microscopy, scanned images

text

- literature, annotations

Big data in biology

- ~90% of biology project leaders are currently or will soon be analyzing large data sets[1]
- most frequent data type: genotypic data



[1] Barone L, Williams J, Micklos D (2017) Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. PLoS Comput Biol 13(10): e1005755

Example: Genomics (M Hofreiter)

context

- ***Ursus spelaeus***
  - European cave bear
  - extinct for ~27,000 yrs

questions

- population sizes & structure?
- behavioral ecology?
- extinction?

data

- mitochondrial genome sequences

ATGCGCAGCGTA

Example: Genomics (M Hofreiter)

sampling & experimental work

- bones and teeth of ~40 cave bears and brown bears from caves in northern Spain
- DNA extraction and preparation for sequencing

data: 1,369,704,332 sequence reads generated

computational work

- quality filtering of data
- comparison to reference mtDNA, computing of one mtDNA sequence per sample
- molecular evolutionary analysis
  - relationships of bears found in different caves?

Example: Genomics (M Hofreiter)

results

- cave bears: one maternal lineage -- one cave (mostly)
- brown bears: mixing of lineages in different caves

implications: homing behavior in <u>cave</u> bears?

- bears return to their native caves for hibernation
- related to extinction of cave bears?
  - competition with humans, predictable prey, inability to colonize nearby empty caves

Cave bear study

DNA sequencing, data processing

- 9 different programs

Evolutionary analysis

- 9 different programs

18 different programs

Example: Animal physiology (S Seyfried)

context

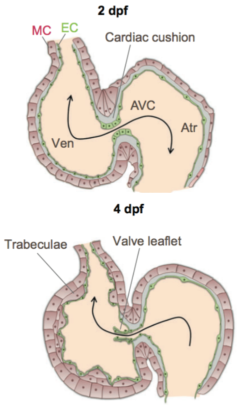
- cardiac development & function
- model system: zebrafish
- congenital heart defects

question

- mechanosensitive signalling during heart development

data

- sequences of genes that are expressed in beating and non-beating hearts

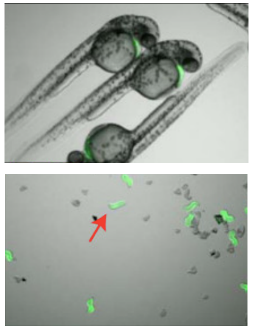


T Haak et al., Comp Biol 2016

Example: Animal physiology (S Seyfried)

sampling & experimental work

- extraction of functioning hearts (wt and mutant)
- mRNA extraction and preparation for sequencing



VA Lombardo et al., Jove 2015

data: 543,076,021 sequence reads

computational work

- quality filtering of data
- quantitative comparison to reference DNA: determination of expression level
- identification of differentially expressed genes

Example: Ecology (U Gaedke)

context

- investigation of invasive species
- *Cylindrospermopsis raciborskii*
- cosmopolitan, toxic, forms blooms

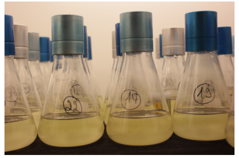
question

- factors for successful invasion?

data

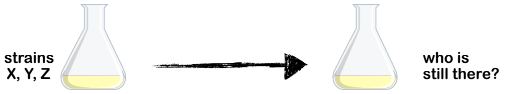
- whole genome sequences of different strains

Example: Ecology (U Gaedke)



sampling & experimental work

- 12 different strains
- grown in isolation
- competition experiments



- microscopy?
- genetic data?
  - sequence & compare genomes
  - find variable regions that can distinguish strains!

Example: Ecology (U Gaedke)

initial data: 48,446,846 sequence reads generated

computational work

- quality filtering of data
- assemble into genome sequences (fragments!)
- analysis & comparison of genome sequences

- comparative genomics
  - characterization, genetic similarity?
  - find variable regions for PCR & sequencing that can distinguish strains

Focus on sequences

- sequence data is abundant
- sequence data is relevant for
  - botany, zoology, microbiology, virology
  - evolution, ecology, genetics, cell & developmental biology, physiology, biochemistry, clinical diagnostics, pharmacology, agriculture, forensics, ...

- sequence data as an example to learn computational skills
- also a brief introduction to structure, networks

Bioinformatics in this course

- data and questions and motivation come from biology
- bioinformatics as a tool box for data analysis
  - algorithms (methods) to extract meaning from data
  - databases, data management tools to make data accessible, to integrate data
  - presentation tools to help comprehend large data sets

- I will introduce and cover many different tools
- I will show their relevance and application

Course goals

- Provide an overview of the concepts, applications, opportunities, and problems in bioinformatics research
  - mostly sequences
  - selected other data types (structures, networks)

- Introduce and allow practical experience with commonly used tools for biological research
  - linux operating system
  - locally installed and online tools
  - online databases
  - statistics software R

Learning goals

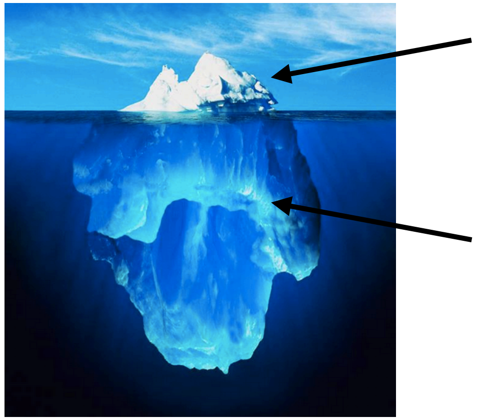
in February of next year, you will be able to:

- understand the principles and applications of important tools for biological data analysis
- search and retrieve genes and associated information (sequences, structures, domains, annotations, ...) from science online databases
- perform small-scale analyses of sequences and structures using locally installed and online tools
- statistically describe / explore biological data, using the software package R
- work under the Linux operating system, also in the terminal

Overview: course content

	lecture	lab
1	introduction	web resources
2	online resources	Genbank, Uniprot
3	linux	linux
4	online resources	linux
5	pairwise sequence comparisons	BLAST
6	multiple sequence alignments	MSA
7	phylogenetics	phylogenetics
8	(human) genomics	MSA, phylogenetics
9	metagenomics	BLAST
10	structure (RNA, protein)	BLAST
11	networks	KEGG
12	R	R (intro)
13	R	R (EDA)
14	R	R (cluster)

That’s just the tip of the bioinformatics iceberg!



what I can cover

- the basics
- selected topics

what I can’t cover

some topics

- in other BSc and MSc courses
- during your thesis work
- on your own when you’ll need it
- never

Lectures

bioinformatics

- new material, new methods, new way of working
- some facts, some biologial applications, many tools

end of lectures

- main terms / concepts

beginning of lectures

- review of previous week's material
- monitor your progress, adjust if necessary

lecture slides are in English, the science language

Computer exercises

- instructions for the computer exercises
  - are in English, the science language
  - are detailed: work through them on your own
  - you can leave if you have completed the exercise
  - some exercises contain optional sections
- finish the exercises before the next class
  - they will be discussed in the following lecture *
- computer exercises reinforce material covered in lectures and therefore are relevant for the exam

http://moodle2.uni-potsdam.de/

lecture slides, relevant literature, exercises will be posted:

- Thursday afternoon
  (usually, and usually up-to-date)

key to lab exercises will be posted:

- never

name: Bioinformatik
ID: bioinf_WS19
password: oregano

Assessment

exam: Fri, Feb 07, 2020 — 8:15 am, last lecture slot!
Haus 27, Raum 1.01

retake: Fri, Mar 20, 2020 — (TBA)

content knowledge

- define, identify, label, describe, summarize, ...

process skills

- predict, apply, compare, interpret, analyze, ...

Don't wait until it is too late!

- keep up with the material
  - attend lecture, complete computer exercises
- review lecture notes
- read material from credible sources
- work together in small groups
- take advantage of office hours
  - S. Hartmann:
    by appointment,
    Haus 29, room no. 2.54
    stefanie.hartmann@uni-potsdam.de

Literature and study aids

• lecture notes	helpful but not sufficient
• online resources	only some are appropriate
• scientific papers	I will assign some articles; esp. review articles are useful
• textbooks	(too) many available check the TOC!

Computer exercises

- get a Computer Pool account as soon as possible
  - https://www.chem.uni-potsdam.de/groups/pools/Studierende/studierende.html

	angemeldet	zugelassen
• 10h-10:45h, three pools (48 computers)	30	50
• 11h-11:45h, three pools (48 computers)	1	50
• 12h - 12:45h, two pools (16 computers)		2

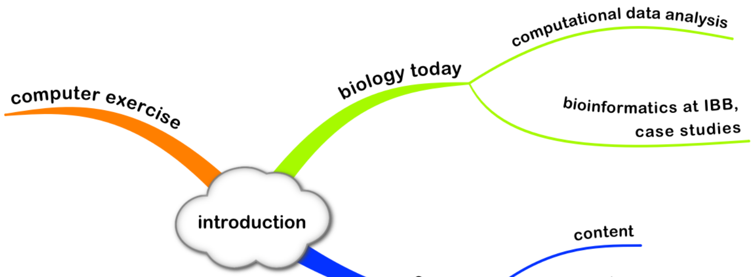
- bitte bei PULS ummelden / abmelden!

Today's computer exercise

if you don't have an account for the pools yet

- temporary computer pool access
  - computer name & IP: Curie, xxx.xxx.xxx.34
  - username: gast_34
  - password: 34-curie.

Today’s topics



introduction

computer exercise

biology today

computational data analysis

bioinformatics at IBB, case studies

course

content

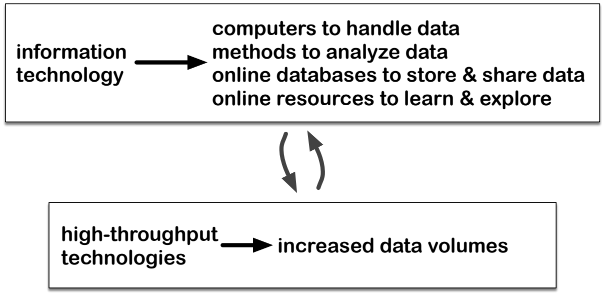
organization

exams

Biology today: an information science

- data-intensive!
- store & manage data
  - huge amounts, heterogeneous data
- make data accessible
  - data organization
  - hierarchical structure
  - appropriate formats & abstraction
- gain biological knowledge
  - learn about organisms, genomes, genes
  - use existing or new approaches
  - computational tools

The internet and biology



information technology → computers to handle data
methods to analyze data
online databases to store & share data
online resources to learn & explore

high-throughput technologies → increased data volumes

Resources

online resources require critical assessment!

- the internet is an electronic repository for text, graphic, and sound
  - much of it is accurate, unbiased, current, and appropriate for academic use
  - much of it is inaccurate, biased, out-of-date, and inappropriate for academic use

information literacy: the ability to
locate, understand, evaluate, and use information

- general-purpose search engines
- biological databases & tools

Today’s computer exercise

1. register for the course on moodle
   - instructions for today’s computer lab are on moodle
2. work through the exercise at your own pace
   - online resources
     - Google, Google Scholar
     - scientific databases
3. finish exercises before the next class

Key terms and concepts

- biological sciences: content knowledge, skills
- why is (almost) every biologists today also a bioinformaticst?
- search engine vs. searchable online database