



ИНСТИТУТ
БИОИНФОРМАТИКИ

Учебные программы курсов Института биоинформатики 2021/22

Программы переподготовки “Биоинформатика для биологов” и
“Алгоритмическая биоинформатика”

[Направление “Биоинформатика для биологов”](#)

[Осенний семестр](#)

[Linux и командная строка для биоинформатики, осень](#)
[Программирование на Python, осень \(Санкт-Петербург\)](#)
[Программирование на Python, осень \(Онлайн\)](#)
[Дискретная математика, осень](#)
[Статистика и анализ данных в R, осень \(Онлайн\)](#)
[Статистика и анализ данных в R, осень \(Санкт-Петербург\)](#)
[Практикум по биоинформатике, осень](#)

[Весенний семестр](#)

[Программирование на Python, весна \(Онлайн\)](#)
[Программирование на Python, весна \(Санкт-Петербург\)](#)
[Машинное обучение, весна](#)
[Практикум по биоинформатике, весна](#)
[Молекулярная эволюция, весна](#)
[Молекулярная филогенетика, весна](#)
[Журнальный клуб, весна \(Санкт-Петербург\)](#)
[Журнальный клуб, весна \(Онлайн\)](#)

[Направление “Алгоритмическая биоинформатика”](#)

[Осенний семестр](#)

[Статистика, осень \(Онлайн\)](#)
[Теория вероятностей, осень \(Санкт-Петербург\)](#)
[Молекулярная биология, осень](#)
[Алгоритмы в биоинформатике, осень](#)

[Практикум по биоинформатике, осень](#)

[Весенний семестр](#)

[Статистика, весна \(Онлайн\)](#)

[Статистика, весна \(Санкт-Петербург\)](#)

[Машинное обучение, весна](#)

[Алгоритмы в NGS, весна](#)

[Практикум по биоинформатике, весна](#)

[Молекулярная эволюция, весна](#)

[Молекулярная филогенетика, весна](#)

[Журнальный клуб, весна \(Санкт-Петербург\)](#)

[Журнальный клуб, весна \(Онлайн\)](#)

Направление “Биоинформатика для биологов”

Осенний семестр

Курс по статистике и программированию на Python идет параллельно у студентов Санкт-Петербурга и онлайн-направления. Дискретная математика и работа в командной строке – общие для всех.

Linux и командная строка для биоинформатики, осень

Преподаватель: Геннадий Захаров, EPAM Systems

Описание курса и план занятий

Общая цель курса – научиться комфортно работать в командной строке Linux/MacOsX и научиться эффективно обрабатывать данные при помощи инструментов командной строки.

Мы познакомимся с ключевыми понятиями, необходимыми для работы в командной строке, наиболее полезными командами, их опциями и примерами использования. Освоим удалённую работу на сервере при помощи командной

строки – особенности работы, обмен данными между собственным компьютером и сервером, полезные приемы.

Изучаемые примеры направлены в первую очередь на анализ данных биоинформатики – секвенированных последовательностей ДНК, аннотаций геномов и т. д. Помимо стандартных команд Linux (zcat, grep, sed, awk) рассмотрим утилиты для работы с геномными данными (bwa, samtools, bedtools).

Практика будет полезна всем, кто ожидает столкнуться с анализом больших объёмов текстовых или табличных данных, в том числе с анализом геномов и транскриптомов. Также практика будет полезна всем планирующим заниматься высокопроизводительными вычислениями на компьютерных кластерах и облачных системах.

Итоги обучения

По окончании практики участники смогут использовать командную строку для анализа данных в тех случаях, где это эффективнее, чем при помощи визуальных утилит или написания собственных программ. Также участники смогут комфортно работать в системах, где единственным доступным интерфейсом является командная строка (например компьютерные кластеры, виртуальные сервера и другие окружения для высокопроизводительных расчётов).

Требования к ПО

Установка окружения Linux

Для выполнения заданий в рамках практики потребуется любой дистрибутив Linux (в виртуальной машине или «на железе»), командная строка MacOSX или дистрибутив, установленный в режиме WSL под Windows 10. Работа всех команд в реальном дистрибутиве Linux и в WSL идентична. Ожидается, что все описанные команды без изменений будут работать в окружении MacOSX, однако 100% применимость всех команд не проверялась.

Полезные инструкции:

1. [Инструкция по установке Linux в VirtualBox \(может дополняться по мере подготовки курса\)](#)

2. [Инструкция по установке «на железо»](#). Актуальной на данный момент является версия Kubuntu 20.04. Установка выполняется аналогично версии 18.04.
3. [Инструкции по активации WSL в Windows 10](#). На данный момент актуальная версия – Ubuntu 20.04. Все остальное делается аналогично.

Установка утилит

При помощи Conda

На сегодняшний день существует множество различных приложений для биоинформатических задач. Многие из этих программ требуют определённых зависимостей и дополнительных библиотек. Поскольку эти программы часто обновляются, версии, поставляемые вместе с дистрибутивом (установленные через apt, yum или аналогичный менеджер пакетов), часто бывают устаревшими. Пакетный менеджер Conda – кроссплатформенное решение (подходит для любой системы) интегрирующее более 2000 различных утилит. На сегодня это оптимальный способ установить и поддерживать в актуальном состоянии софт для биоинформатики.

Детальная установка Conda описана [на сайте](#) для каждой конкретной операционной системы. В нашем случае нужны только утилиты командной строки, так что процесс установки очень прост:

Для Linux

1. Скачиваем [установочный файл Miniconda wget](#)
2. Запускаем установку: `bash Miniconda3-latest-Linux-x86_64.sh`
3. На вопрос про инициализацию отвечаем `yes`
4. После нужно закрыть и снова открыть терминал.

Для MacOSX

Скачать и запустить установочный файл согласно инструкции на сайте [MiniConda](#).

Установка пакетов через Conda

Нам надо подключить дополнительные коллекции программ (каналы), создать наше собственное окружение и установить в него дополнительные пакеты. Все это выполняется следующими командами:

`conda config --add channels conda-forge`

conda config --add channels bioconda
conda update --all

conda install bwa samtools bedtools snakemake tmux sra-tools mamba
conda install может выполняться довольно долго, в зависимости от скорости компьютера и интернет-соединения.

При помощи пакетного менеджера системы

Если, как в нашем случае, требуется установить всего несколько утилит, и нет необходимости иметь самые последние версии, то это можно сделать из пакетного менеджера системы:

Для Ubuntu/Debian потребуется всего 2 команды. Первая перечисляет список доступных программ на сервере, вторая собственно выполняет установку.

sudo apt update

sudo apt install bwa samtools bedtools snakemake tmux

sudo потребует ввести пароль пользователя. По традиции UNIX-систем при вводе пароль никак не отображается (вдруг за вами наблюдает злоумышленник, он не должен знать даже число символов в пароле), так что вводить придётся вслепую.

Аттестация

Итоговая оценка будет складываться из результатов тестов после каждой практики. Тесты будут проходить в виде набора задач на платформе Stepik.

Дополнительные материалы и полезные ссылки

1. <https://habr.com/ru/post/47163/>
2. <https://losst.ru/shpargalka-po-tmux>
3. <https://linuxconfig.org/bash-scripting-tutorial-for-beginners>
4. <https://ryanstutorials.net/bash-scripting-tutorial/>
5. <https://habr.com/ru/post/126996/>
6. <https://www.biostars.org/p/335903/>

Программирование на Python, осень (Санкт-Петербург)

Преподаватели: Слепченков Александр, Evil Martians; Сытник Екатерина, Mislab

Описание курса и план занятий

В курсе будут рассмотрены основы программирования на языке Python: синтаксис, типы данных, функции и классы, работа с исключениями, ввод-вывод. Кроме этого мы научимся работать с инструментами, которые упрощают процесс разработки программ, и обсудим подходы, которые позволяют создавать более качественный код. Упор будет сделан на применение языка, знакомство с экосистемой Python и практические навыки разработки ПО.

Занятия планируется проводить в форме семинаров, большая часть времени которых будет отведена практике и обсуждению сложных вопросов, значительную часть теории предполагается изучать в форме домашней подготовки. Существенное место будет отведено решению задач и ревью кода.

Содержание курса предварительное и может меняться в зависимости от уровня подготовки учащихся и прогресса. Порядок перечисления тем не отражает порядок их изложения.

Python (язык)

1. Переменные и операции
2. Ввод-вывод
3. Коллекции
4. Функции и исключения
5. Классы и типы
6. Модули и пакеты

Python (экосистема)

1. PEP
2. Интерпретатор (поверхностно)
3. Средства работы с окружениями и менеджеры пакетов: virtualenv, pip, pipenv

Разработка ПО

1. Git / GitHub
2. Дебаггинг
3. Тестирование
4. Подходы к разработке ПО

Computer Science

1. Регулярные выражения
2. Системы счисления, бинарные операции

3. Структуры данных.

- Linked list
- Queues
- Graphs
- Hash tables

Алгоритмы

1. Сортировки
2. “Графовые алгоритмы”
3. Бинарный поиск
4. Динамическое программирование

Чего с большой вероятностью не будет в курсе:

- Подрбностей устройства интерпретатора.
- Многопоточного программирования. В питоне нет потоков в общепринятом понимании в силу внутренних особенностей устройства интерпретатора и в большинстве биоинформатических задач можно обойтись без этого.
- Анализа данных и изучения документации библиотек.

Кроме того, по этим темам достаточно книжек и докладов, и если они вам понадобились, то, скорее всего, уровень достаточен для того, чтобы их понять

Итоги обучения

По итогам обучения на курсе студенты научатся:

- Писать программы на python (и не писать, когда это не надо)
- Читать чужой код на python
- Тестировать и отлаживать код на python
- Проводить ревью чужого кода на python
- Самостоятельно разбираться в документации python и его библиотек
- Пользоваться поисковыми системами / [stackoverflow](#)
- Работать с системами контроля версий в лице Git
- Получить общее представление о самых простых и распространённых структурах данных и алгоритмах

Требования к ПО

Для выполнения курса на вашем компьютере должны быть установлены:

1. [Python 3.7](#) и выше
2. [git](#) (любой вариант, позволяющий работать из командной строки)
3. [PyCharm Community](#)
4. [virtualenv](#)

У вас должен быть аккаунт на [GitHub](#)

Аттестация

Для получения зачета в первой части курса необходимо будет набрать 70% от максимального количества баллов. Баллы будут начисляться за домашние задания, проверочные работы и работу на семинарах.

Дополнительные материалы и полезные ссылки

- [Python 3 documentation](#)
- [PEP 8 – Style Guide for Python Code](#)
- [Git](#)

Программирование на Python, осень (Онлайн)

Преподаватели: Александр Ильин, Skoltech; Роман Кругликов

Описание курса и план занятий

В курсе будут рассмотрены основы программирования на языке Python: синтаксис, типы данных, функции и классы, работа с исключениями, ввод-вывод. Кроме этого мы научимся работать с инструментами, которые упрощают процесс разработки программ, и обсудим подходы, которые позволяют создавать более качественный код. Упор будет сделан на применение языка, знакомство с экосистемой Python и практические навыки разработки ПО. Занятия планируется проводить в форме семинаров, большая часть времени которых будет отведена практике и обсуждению сложных вопросов, значительную часть теории предполагается изучать в форме домашней подготовки. Существенное место будет отведено решению задач и ревью кода.

Содержание курса предварительное и может меняться в зависимости от уровня подготовки учащихся и прогресса. Порядок перечисления тем не отражает порядок их изложения.

Python (язык)

1. Переменные и операции
2. Ввод-вывод
3. Коллекции
4. Функции и исключения
5. Классы и типы
6. Модули и пакеты

Python (экосистема)

1. PEP
2. Интерпретатор (поверхностно)
3. Средства работы с окружениями и менеджеры пакетов: virtualenv, pip, pipenv
4. Библиотеки pandas/numpy/biopython

Разработка ПО

1. Git / GitHub
2. Дебаггинг
3. Тестирование
4. Подходы к разработке ПО

Computer Science

4. Регулярные выражения
5. Структуры данных: Graphs, Hash tables и тд
6. Алгоритмы
 - Сортировки
 - “Графовые алгоритмы”
 - Бинарный поиск
 - Динамическое программирование

Чего с большой вероятностью не будет в курсе:

- Подробностей устройства интерпретатора.
- Многопоточного программирования. В питоне нет потоков в общепринятом понимании в силу внутренних особенностей устройства интерпретатора и в большинстве биоинформатических задач можно обойтись без этого.

Кроме того, по этим темам достаточно книжек и докладов, и если они вам понадобились, то, скорее всего, уровень достаточен для того, чтобы их понять

Итоги обучения

По итогам обучения студенты научатся:

- Писать программы на python (и не писать, когда это не надо)
- Читать чужой код на python
- Тестировать и отлаживать код на python
- Проводить ревью чужого кода на python
- Самостоятельно разбираться в документации python и его библиотек
- Пользоваться поисковыми системами / stackoverflow
- Работать с системами контроля версий в лице Git
- Получить общее представление о самых простых и распространённых структурах данных и алгоритмах

Требования к ПО

Для выполнения курса на вашем компьютере должны быть установлены:

5. [Python 3.7](#) и выше
6. [git](#) (любой вариант, позволяющий работать из командной строки)
7. [PyCharm Community](#)
8. [virtualenv](#)

У вас должен быть аккаунт на [GitHub](#)

Аттестация

Для получения зачета в первой части курса необходимо будет набрать 70% от максимального количества баллов. Баллы будут начисляться за домашние задания, проверочные работы и работу на семинарах.

Дополнительные материалы и полезные ссылки

1. [Python 3 documentation](#)
2. [PEP 8 – Style Guide for Python Code](#)

3. [Git](#)

Дискретная математика, осень

Преподаватели: Артём Васильев, Университет ИТМО; Дарья Балашова (ассистент), PhD student, Amsterdam UMC / University of Amsterdam

Описание курса и план занятий

В этом курсе мы познакомимся с основными понятиями и алгоритмами дискретной математики: основная математическая нотация, комбинаторика, динамическое программирование и алгоритмы теории графов.

Темы занятий:

1. Логика, доказательства, булевы функции
2. Теория множеств
3. Комбинаторика
4. Комбинаторная генерация
5. Асимптотический анализ алгоритмов
6. Алгоритмы сортировки
7. Динамическое программирование
8. Теория графов: основные понятия
9. Теория графов: алгоритмы на графах

Итоги обучения

1. Познакомятся с основными понятиями дискретной математики и алгоритмов: теории множеств, комбинаторики, динамического программирования, теории графов.
2. Научатся строить строгие математические рассуждения и доказательства.
3. Научатся оценивать время работы алгоритма, понимать, какое время работы приемлемо для каких объемов данных.
4. Научатся формулировать и решать задачи с помощью динамического программирования.

Требования к ПО

Требования к предустановленным программам нет

Аттестация

При получении финальной оценки учитываются баллы за домашние задания и тесты.

При получении $\geq 90\%$ баллов ставится оценка "5".

При получении $\geq 75\%$ баллов ставится оценка "4".

При получении $\geq 60\%$ баллов ставится оценка "3".

Статистика и анализ данных в R, осень (Онлайн)

Преподаватели: Лаврентий Данилов, СПбГУ; Данил Литвинов, МГУ

Описание курса и план занятий

Статистика

Правила постановки эксперимента, формулирование гипотез. Случайные и достоверные события в биологии. Случайные и систематические ошибки при постановке экспериментов. Понятие случайной величины. Ее основные свойства. Понятие генеральной совокупности. Мода, медиана и среднее как меры центральной тенденции.

Математические свойства среднего. Методы создания выборок -- рандомизация, попарный отбор (для исследований взаимосвязанных событий), стратометрический отбор (кластерная выборка), приближенное моделирование. Планирование эксперимента: Определение минимально необходимого объема выборки для проведения эксперимента. Меры изменчивости данных. Размах, дисперсия и стандартное отклонение как метрики для описания выборки. Свойства дисперсии и стандартного отклонения.

Основные понятия теории вероятности. Формула Бернулли. Байесовская статистика. Теорема Байеса. Априорная и апостериорная вероятности. Применение в филогенетике. Влияние результатов предыдущих экспериментов на текущий.

Экспоненциальное распределение, t-распределение, распределение хи-квадрата, нормальное распределение. Характеристика нормального распределения. Правило трех сигм. Центральная предельная теорема и следствие из неё. Формулирование нулевой и альтернативной гипотезы. Двухсторонняя и односторонняя альтернативная гипотеза. Случайность и достоверность различий. Уровень значимости (p-value). Ошибки первого и второго рода. Мощность тестов.

Описание данных, которые могут быть непараметрическими (социологические опросы, расщепление при скрещивании). Методика Хи-квадрата, расчет

хи-квадрата для 2 признаков. Многопольный Хи-квадрат, ограничение метода. Точный критерий Фишера. Условия применения точного критерия Фишера.

Т-распределение, свойства t-распределения. Т-критерий Стьюдента. Ограничение использования t-критерия Стьюдента. Использование критерия для сравнения средних двух выборок. Проверка распределения на нормальность. Сравнение нескольких выборок. Введение поправок на сравнение.

Корреляция между двумя переменными, понятие корреляции. Коэффициент корреляции и его доверительный интервал. Корреляций Пирсона, Спирмена и Кендалла. Правила применимости. Сравнение тестирование гипотез и построение моделей как 2 подхода в статистике. Состав простой линейной модели. Обозначение коэффициентов. Математическое обоснование коэффициентов интерсепт и бета. Графическая интерпретация коэффициентов.

Условия применимости линейных моделей. Трансформация данных. Мультиколлинеарность предикторов.

Анализ изменчивости параметров: Дисперсионный анализ. Модели дисперсионного анализа. F-значение, его расчет и интерпретация. Распределение F-значения. Расчет межгруппового среднего квадрата и внутригруппового среднего квадрата. Проблема множественных сравнений. Введение поправки Бонферрони и Холма, их сравнение друг с другом. Использование ANOVA и ANCOVA. Общая изменчивость и факторная изменчивость. Условия применимости. Анализ главных компонент (PCA). Матричная реализация PCA через SVD, UMAP и TSNE. Факторные нагрузки и ординации. Свойства компонент. Интерпретация результатов PCA.

Логистическая регрессия. Логистическая регрессия для дискретных данных. Логит-преобразования. Метод Максимального-Правдоподобия. Тест Вальда. Анализ девиансы и G2 критерий. Проверка на избыточность дисперсии. Множественная логистическая регрессия. Удаление предикторов и упрощение модели.

Программирование в R

Типы переменных -- непрерывные, дискретные, ранговые, категориальные, логические. Создание пользовательских функций. Глобальное и локальное окружение. Встроенный git и Rproject. Использование функций из пакетов, загрузка и установка пакетов (ggplot2, readxl, dplyr). Сабсетинг данных. Функции семейства apply. Пользовательские функции. Создание отчетов в .Rmd.

Итоги обучения

1. Знакомство с основами программирования на языке R. Выполнение базовых операций и работа с функциями из различных пакетов.

2. Проведение первичного анализа данных. Навык выдвижения и проверки гипотез в рамках проводимого исследования.
3. Подбор адекватных статистических методов в соответствии с задачей исследования. Проверка условий применимости статистических тестов.
4. Презентация полученных результатов статистически грамотным языком в специализированном формате.

Требования к ПО

1. Язык программирования [R версии](#) 4.0.X
2. Среда для работы с R - [Rstudio](#) Desktop 1.3.X с установленными пакетами из коллекции tidyverse (вся коллекция).
3. [Git](#) (любой вариант, позволяющий работать из командной строки)

Аттестация

При получении оценки учитываются домашние задания и самостоятельные работы.

Статистика:

- 65% — 3
- 75%-85% — 4
- >85% + по всем заданиям есть хотя бы 1 балл — 5

Программирование на R — зачет от 70% баллов

Дополнительные материалы и полезные ссылки

1. С. Гланц Медико-биологическая статистика Пер. с англ. под ред. Н. Е.Бузикашвили и Д. В. Самойлова М., «Практика», 1999 г
2. Gareth James, Daniela Witten, Trevor Hastie and Rob Tibshirani “An Introduction to Statistical Learning”
3. А. Б. Шипунов, Е. М. Балдин, П. А. Волкова, А. И. Коробейников, С. А. Назарова, С. В.Петров, В. Г. Суфиянов “Наглядная статистика. Используем R!” 2014.
4. Alain F. Zuur | Elena N. Ieno | Erik H.W.G. Meesters A Beginner's Guide to R. 2009

5. David S. Moore, George P. McCabe, Bruce A. Craig, "Introduction to the Practice of Statistics", 2010

Статистика и анализ данных в R, осень (Санкт-Петербург)

Преподаватели: Алексей Забелкин, Университет ИТМО; Дарья Халенова, СПбГУ

Описание курса и темы занятий

1. Introduction to statistics: variable types, basic distributions and hypotheses examples
Понятие случайной величины и различные их типы: численные (непрерывные и дискретные), ординальные, классовые. Основные вопросы статистики. Примеры построения распределений и их визуализации. Примеры построения статистических гипотез: goodness of fit, равенство средних, независимость переменных. Практика – знакомство с R и встроенными типами данных.
2. Calculating data statistics: central and variance tendency measures, correlation
Описательные выборочные статистики по данным. Меры центральной тенденции: средние, медиана, мода. Асимметрия как мера "скоса". Меры разброса: межквартильный размах, дисперсия, стандартное отклонение. Ковариация и корреляция. Практика – применения данных методов в языке R.
3. Data visualization: distribution visualization technics, multidimensional data, heatmaps, Venn diagrams
Методы визуализации распределений: ящик с усами, скрипичная диаграмма, гистограмма. Их плюсы и минусы в разных ситуациях. Тепловые карты и примеры их применения. Методы визуализации пересечения множеств: диаграммы Венна, UpSet.
4. Data visualization practice: visualizing experiment data, explanatory data analysis (EDA)
Разведочный анализ данных на примере датасета о результатах Иммунотерапии. Визуализация данных множественного повторения эксперимента. Разбор типичных ошибок при визуализации данных.
5. Testing mean equality one-sample hypotheses with simulations
Практика. Проверка гипотезы об эффективности разработанного лекарства на основании использования эмуляций. Введение понятий p-value, а также нулевой и альтернативной гипотез.
6. Testing goodness of fit hypothesis with simulations

Практика. Проверка гипотезы goodness of fit на примере подсчета соответствия GC-контента в полученных данных литературным.

7. Testing independency hypothesis with simulations

Практика. Понятие независимых событий и условной вероятности. Проверка гипотезы о независимости переменных в случае классовых переменных. Проверка гипотезы о независимости переменных в случае численных переменных. Сведение задачи проверки независимости численных переменных к классовым при помощи бининга. Разбор случая смешанных переменных.

8. Sampling methods, sources of bias

Разница наблюдательных и экспериментальных подходов. Возможные отношения между скоррелированными переменными, примеры. Различные методы построения выборок (вероятностные и невероятностные) и типичные источники ошибок (bias). Дизайн эксперимента.

9. Distributions: Bernoulli, binomial, geometrical, uniform, normal, exponential, Poisson

Понятие дискретного распределения. Примеры дискретных распределений: Бернулли, биномиальное, геометрическое, распределение Пуассона. Примеры их встречаемости в реальности. Понятие непрерывного распределения. Задание непрерывного распределения через кумулятивную функцию вероятности. Примеры непрерывных распределений: равномерное, нормальное, экспоненциальное

10. Testing Hypotheses using Theory

Применение теоретических знаний о распределениях для построения статистических тестов. Проверка гипотез goodness of fit, независимости переменных и равенства средних с помощью теории. Сравнение с реализованными тестами.

11. Types of errors and power of tests

Типы ошибок при тестировании гипотез: ошибки первого и второго рода. Особенности данных ошибок и их интерпретация. Концепция компромисса между типами ошибок.

12. Central limit theorem, t-tests

Центральная предельная теорема. Примеры ее применения и встречаемости в реальной жизни. Различные типы t-test для проверки равенства средних: одновыборочный, двухвыборочный (зависимый и независимый случай).

13. Multiple comparisons problem

Проблема множественного сравнения. Различные типы поправок на множественное сравнение. Тест ANOVA для множественной проверки

равенства средних и принципы его работы. Концепция односторонней и двусторонней альтернативной гипотезы.

14. Regression analysis: linear regression using least squares

Обзор более сложных статистических методов. Задача уменьшения размерности. Задача линейной регрессии и различные способы ее решения. Тесты и статистики основанные на рангах: корреляция Спирана, тест Манна-Уитни.

Итоги обучения

1. Знакомство с основами программирования на языке R. Выполнение базовых операций и работа с функциями из различных пакетов.
2. Проведение первичного анализа данных. Навык выдвижения и проверки гипотез в рамках проводимого исследования.
3. Подбор адекватных статистических методов в соответствии с задачей исследования. Проверка условий применимости статистических тестов.
4. Презентация полученных результатов статистически грамотным языком в специализированном формате.

Требования к ПО

1. Язык программирования [R версии](#) 4.X.X
2. Среда для работы с R - [Rstudio](#) Desktop 1.4.X с установленными пакетами из коллекции tidyverse (вся коллекция).
3. [Git](#) (любой вариант, позволяющий работать из командной строки)

Аттестация

При получении оценки учитываются Домашние задания и самостоятельные работы.

60-74% — 3

74-90% — 4

>= 90% — 5

Дополнительные материалы и полезные ссылки

Книги:

1. Diez DM, Cetinkaya-Rundel M, and Barr CD. OpenIntro Statistics. Fourth edition. OpenIntro, 2019.
 - *have great case studies and cover most statistics topics*
2. Gareth James, Daniela Witten, Trevor Hastie and Rob Tibshirani “An Introduction to Statistical Learning”
 - *cover topics about statistical learning and models construction*
3. David S. Moore, George P. McCabe, Bruce A. Craig, “Introduction to the Practice of Statistics”, 2010
 - *more about the practical part of statistics*

Курсы:

1. [Introduction to Probability and Data with R, Duke University](#)
 - a. *Basic, but the well-explained course about data. Also have part with introduction to probabilistic theory.*

Практикум по биоинформатике, осень

Преподаватели: Михаил Райко, ЦАБ СПбГУ; Юрий Барбитов, Институт биоинформатики; Алиса Моршнева (ассистент), ЦИН РАН

Описание курса и план занятий

Курс состоит из нескольких практических заданий, которые выполняются в смешанных группах. Задания будут основаны на реальных данных, и посвящены различным аспектам биоинформатики (в основном анализ данных NGS).

Темы проектов:

1. Выравнивание на референс и поиск SNP
2. Сборка геномов de novo
3. Аннотация геномов – структурная и функциональная
4. Филогенетика
5. РНК-секвенирование и дифференциальная экспрессия генов
6. Аннотация сипов у человека
7. Метагеномика

Итоги обучения

В рамках курса студенты научатся:

1. Использовать различные биоинформатические инструменты для решения конкретных биологических задач
2. Выбирать подходящие тулы под задачи, задавать осмысленные вопросы и тестировать гипотезы.
3. Писать связные осмысленные тексты, что было сделано и зачем

Требования к ПО

Все проекты подразумевают запуск тулов, поэтому потребуется доступ к компьютерам с Linux. Данные будут не очень большие, мощности среднего ноутбука должно хватить.

Аттестация

По каждому проекту надо написать отчёт. Отчёт представляет собой мини-статью, на 1-2 страницы, с введением, методами, результатами и т.п.

За каждый отчёт можно получить максимум 10 баллов. Можно добрать 2 дополнительных балла за “лабораторный журнал”. Один проект с минимальным числом баллов будет отброшен. Еще 1 дополнительный балл, если отчёт написан на английском.

Для зачёта нужно получить к концу семестра не менее 70% от максимального количества баллов.

Весенний семестр

Программирование на Python, весна (Онлайн)

Преподаватели: Александр Ильин, Skoltech; Роман Кругликов, ИБХ РАН

Описание курса и план занятий

В курсе будут рассмотрены основы программирования на языке Python: синтаксис, типы данных, функции и классы, работа с исключениями, ввод-вывод. Кроме этого мы научимся работать с инструментами, которые упрощают процесс разработки программ, и обсудим подходы, которые позволяют создавать более качественный код. Упор будет сделан на применение языка, знакомство с экосистемой Python и практические навыки разработки ПО.

Занятия планируется проводить в форме семинаров, большая часть времени которых будет отведена практике и обсуждению сложных вопросов, значительную часть теории предполагается изучать в форме домашней подготовки. Существенное место будет отведено решению задач и ревью кода.

Темы:

1. ООП
2. Pandas
3. Работа с интернетом, веб-сервера
4. Библиотека requests, парсинг веб-страниц, работа с API
5. Итераторы и генераторы
6. Декораторы
7. Логирование
8. Стил ь кода, аннотации типов, документация и PEP8
9. Параллельное программирование
10. Групповой проект
11. Новые возможности в Python 3.7+
12. Тестирование кода
13. SQL

Чего с большой вероятностью не будет в курсе:

- Подробностей устройства интерпретатора.
- Изучения редко используемых на практике особенностей языка (напр. контекстные менеджеры, метаклассы и т.д.)

Кроме того, по этим темам достаточно книжек и докладов, и если они вам понадобились, то, скорее всего, уровень достаточен для того, чтобы их понять.

Итоги обучения

По итогам обучения на курсе студенты научатся:

- Писать программы на python (и не писать, когда это не надо)
- Читать чужой код на python
- Тестировать и отлаживать код на python
- Проводить ревью чужого кода на python
- Самостоятельно разбираться в документации python и его библиотек
- Пользоваться поисковыми системами / stackoverflow
- Работать с системами контроля версий в лице Git
- Получить общее представление о самых простых и распространённых структурах данных и алгоритмах

Аттестация

Баллы начисляются за выполнение домашних заданий и контрольных. Оценка выставляется на основании баллов:

[60; 70)% — 3

[70; 80)% — 4

[80; +∞)% — 5

Программирование на Python, весна (Санкт-Петербург)

Преподаватели: Слепченков Александр, Evil martians; Екатерина Сытник, Mislab

Описание курса и план занятий

В курсе будут рассмотрены основы программирования на языке Python: синтаксис, типы данных, функции и классы, работа с исключениями, ввод-вывод. Кроме этого мы научимся работать с инструментами, которые упрощают процесс разработки программ, и обсудим подходы, которые позволяют создавать более качественный код. Упор будет сделан на применение языка, знакомство с экосистемой Python и практические навыки разработки ПО. Занятия планируется проводить в форме семинаров, большая часть времени которых будет отведена практике и обсуждению сложных вопросов, значительную часть теории предполагается изучать в форме домашней подготовки. Существенное место будет отведено решению задач и ревью кода.

Темы:

1. Функциональное программирование
2. ООП. Общие принципы
3. ООП. Классы, наследование

4. ООП. Разбор примеров
5. Тестирование
6. Контекстные менеджеры
7. Декораторы
8. Docker
9. SQL
10. Пайплайны
11. Сеть. Форматы данных для передачи по сети

Чего с большой вероятностью не будет в курсе:

- Подробностей устройства интерпретатора.
- Многопоточного программирования. В питоне нет потоков в общепринятом понимании в силу внутренних особенностей устройства интерпретатора и в большинстве биоинформатических задач можно обойтись без этого.

Кроме того, по этим темам достаточно книжек и докладов, и если они вам понадобились, то, скорее всего, уровень достаточен для того, чтобы их понять

Итоги обучения

По итогам обучения на курсе студенты научатся:

- Писать программы на Python (и не писать, когда это не надо)
- Читать чужой код на Python
- Тестировать и отлаживать код на Python
- Проводить ревью чужого кода на Python
- Самостоятельно разбираться в документации python и его библиотек
- Пользоваться поисковыми системами / [stackoverflow](#)
- Работать с системами контроля версий в лице Git
- Получить общее представление о самых простых и распространённых структурах данных и алгоритмах

Аттестация

Баллы начисляются за выполнение домашних заданий и контрольных. Оценка выставляется на основании баллов:

- [60; 70)% — 3
- [70; 80)% — 4
- [85; +∞)% — 5

Машинное обучение, весна

Преподаватели: Лаврентий Данилов, аспирант СПбГУ, ЕРАМ; Даниил Литвинов, ImmunoMind

13 занятий: 13-14 лекций-семинаров, 12 квизов, 3-4 домашних задания

Описание курса и темы занятий

1. Введение. О чем наш курс? Что из себя представляет машинное обучение и где оно применяется? Различные типы задач, решаемые при помощи машинного обучения. Алгоритм k ближайших соседей (KNN). Что такое k и как правильно его выбрать. Общий план действий при работе с методами машинного обучения. Как правильно измерить качество полученной модели?
2. Линейные методы: линейная и логистическая регрессия. Как алгоритмы машинного обучения “учатся” (градиентный спуск). Еще способы оценить качество модели (ROC AUC, PR AUC etc).
3. Кластеризация.
4. Деревья решений. Алгоритмы, основанные на деревьях решений (начало).
5. Ансамбли алгоритмов машинного обучения.
6. Отбор и создание признаков.
7. Введение в нейронные сети. Перцептрон. Знакомство с библиотекой PyTorch.
8. Обучение нейронных сетей. Алгоритм обратного распространения ошибки.
9. Нейронные сети для работы с картинками. Операция свертки.
10. Продвинутое обучение нейронных сетей. Борьба с переобучением.
11. Обучение на малом объеме данных. Как взять готовую модель и применить ее для своих целей?
12. Задачи компьютерного зрения: сегментация и детекция.
13. Explainability и interpretability или как понять мотивацию вашей модели и попытаться интерпретировать результаты.
14. Нейронные сети для работы с последовательностями (текст, звуки, видео).

Итоги обучения

После прохождения курса студентами будут приобретены следующие компетенции:

1. Умение работать с библиотекой scikit-learn

2. Выбор и оптимизация алгоритмов машинного обучения для различных задач.
3. Умение работать с нейронными сетями.

Требования к ПО

Язык программирования Python не ниже версии 3.7. Потребуется установка cuda и фреймворков для работы с нейросетями.

Аттестация

За выполнение домашних заданий, прохождение онлайн курса на Stepik и решение квизов выставляются баллы. Оценка выставляется на основании баллов:

- от 65-70% — 3
- от 71-85% — 4
- больше >85% — 5

Практикум по биоинформатике, весна

Преподаватели: Михаил Райко, ЦАБ СПбГУ; Юрий Барбитов, Институт биоинформатики; Алиса Моршнева (ассистент), ИНЦ РАН

Описание курса и темы занятий

Курс состоит из нескольких практических заданий, которые выполняются в смешанных группах. Задания будут основаны на реальных данных, и посвящены различным аспектам биоинформатики (в основном, анализ данных NGS).

Темы проектов весеннего семестра:

1. Человеческая генетика, аннотация SNP, ассоциации
2. РНК-секвенирование и дифференциальная экспрессия генов
3. Метагеномика (ампликонная и shotgun)
4. Иммуногеномика, RepSeq
5. Протеомика

Итоги обучения

В рамках курса студенты научатся:

1. Использовать различные биоинформатические инструменты для решения конкретных биологических задач
2. Выбирать подходящие тулы под задачи, задавать осмысленные вопросы и тестировать гипотезы.
3. Писать связные осмысленные тексты, что было сделано и зачем

Требования к ПО

Все проекты подразумевают запуск тулов, поэтому потребуется доступ к компьютерам с Linux. Данные будут не очень большие, мощности среднего ноутбука должно хватить.

Аттестация

По каждому проекту надо написать отчёт. Отчёт представляет собой мини-статью, на 1-2 страницы, с введением, методами, результатами и т.п.

За каждый отчёт можно получить максимум 10 баллов. Можно добрать 2 дополнительных балла за “лабораторный журнал”. Один проект с минимальным числом баллов будет отброшен. Ещё один дополнительный балл, если отчёт написан на английском.

Для зачёта нужно получить к концу семестра не менее 70% от максимального количества баллов.

Молекулярная эволюция, весна

Преподаватель: Юрий Барбитов, Институт биоинформатики

Описание курса и план занятий

- Введение. Теория эволюции. Генетическое разнообразие популяций.
- Генетика популяций. Закон Харди-Вайнберга.
- Эффекты случайного мутагенеза и потока генов на частоты аллелей в популяции.
- Дрейф генов. Модель Райта-Фишера. Естественный отбор и его математическое описание.
- Геномный анализ популяционной структуры. Генетическое смешение. F-статистики и ABBA/BABA тест. ADMIXTURE.
- Методы оценки давления отбора на последовательности генов и геномов. dN/dS, Tajima's D, FHH.
- Анализ демографической истории популяций. Коалесцентная теория. Skyline-методы. Группа методов на основе SMC.
- Геномные тренды в эволюции. Селективное давление против новых мутаций в простых и сложных геномах.

Онлайн-часть:

1. [Курс “Молекулярная филогенетика”](#)
2. [Курс “Генетика и геномика популяций”](#)
3. Методы построения и проверки филогенетических деревьев. Методы расстояний и дискретные методы. Байесовский подход.

Молекулярная филогенетика, весна

Преподаватель: Полина Дроздова, НИИ Биологии ИГУ

Описание курса и темы занятий

Курс состоит из онлайн-курса (самостоятельное изучение) и практикумов, на которых мы сосредоточимся на применении программ. По результатам практикумов нужно будет написать мини-отчёты по предоставленному плану.

План занятий

1. Введение. Работа с деревьями. Как построить дерево в R & Python. Будем работать с: R (ape, ggtree); python (Biopython, etc3). HW1: *мини-отчёт №1 + модуль 1 онлайн-курса.*

2. Откуда брать данные, базы данных, встроенные инструменты, blast и всё такое. Учимся работать с: ncbi-blast+, ncbi eutils. HW2: *мини-отчёт №2 + модуль 2 онлайн-курса.*
3. Выравнивание нуклеотидных и белковых последовательностей. Будем работать с: clustalw, clustalO, muscle, mafft, kalign, tcoffee, prank. HW3: *мини-отчёт №3 + модуль 3 онлайн-курса, уроки 1-4.*
4. Выбор модели замен и построение дерева. Учимся работать с: PartitionFinder, jModelTest2, RAxML, IQtree. HW4: *мини-отчёт №4 + модуль 3 онлайн-курса, уроки 5-9.*
5. Байесовские методы и датировка. Учимся работать с: *BEAST. HW5: *модуль 4 онлайн-курса + мини-отчёт №5.*
6. Маркерные гены. Филогеномика. Обсуждение итогов курса.

Самостоятельное изучение

- [Онлайн-курс “Молекулярная филогенетика”](#)

Итоги обучения

Участники курса:

- узнают, как работает филогенетический анализ:
 - где хранятся данные последовательностей и как извлекать данные из баз;
 - какие данные стоит использовать для филогенетического анализа и почему;
 - как работает выравнивание;
 - что нужно, чтобы построить дерево на основе выравнивания;
 - какие бывают способы построения деревьев;
 - как оценить качество полученной филогении.
- на практике научатся работать с основным набором программ, популярных в современной филогенетике:
 - извлечения данных из NCBI;
 - выравнивания нуклеотидных и белковых последовательностей;
 - подбора оптимальной модели эволюции последовательностей;
 - построения филогенетических деревьев;
 - отрисовки деревьев.

Аттестация

- Общий балл будет складываться из балла за курс (50 %) и баллов за мини-отчёты (в сумме дают 50 %).

- При получении $\geq 90\%$ баллов ставится оценка “5”.
- При получении $\geq 75\%$ баллов ставится оценка “4”.
- При получении $\geq 60\%$ баллов ставится оценка “3”.
- При сдаче отчёта после дедлайна, но до конца курса оценка за него будет снижена в два раза. То же верно про сдачу теоретической части (онлайн-курса).

Журнальный клуб, весна (Санкт-Петербург)

Преподаватель: Александр Ткаченко, Университет ИТМО

Описание курса и темы занятий

Научный семинар, который объединяет потоки биологов и информатиков для рассмотрения актуальных научных статей по биоинформатике и связанным темам.

В течение этого семестра каждый из студентов выступает с докладом по мотивам научной статьи. Темы занятий и статьи выбирает и распределяет между участниками куратор. В подготовке доклада принимает участие два человека: обычно биолог и информатик, но не всегда. Вместе им необходимо прочитать и понять предложенную статью, а затем донести ваше понимание до аудитории.

Темы занятий:

- Single-cell sequencing
- Metagenomics
- Protein design
- Proteomics
- Population genomics
- Nanopore sequencing and genome assembly

Аттестация

Критерий зачета: один доклад за семестр.

Журнальный клуб, весна (Онлайн)

Преподаватель: Полина Дроздова, НИИ Биологии ИГУ

Описание курса и темы занятий

Научный семинар, который объединяет потоки биологов и информатиков для рассмотрения актуальных научных статей.

- Семинары **не будут** проводиться на английском языке, но мы будем много внимания обращать на язык: читать вслух, обсуждать особенности научного английского, термины, произношение и принципы написания научных статей.
- За эти 6 занятий мы рассмотрим 3 статьи: одну экспериментальную статью, один обзор и одну статью про программу или алгоритм.
- Ваши предложения статей на обсуждение всегда приветствуются! Статью будем выбирать голосованием (окончательное решение остаётся за преподавателем). Критерии подходящей статьи: она должна быть англоязычной, интересной, желательно цитируемой, желательно с открытым доступом (или легко доступна на другом ресурсе, если журнал нас не любит).

План занятий:

1. Вводное занятие. О том, как улучшать свой английский и что нужно для свободного общения в науке. Начинаем читать и обсуждать экспериментальную статью. ДЗ: дочитать статью самостоятельно, можно начать заполнять заметки по ней.
2. Заканчиваем обсуждение экспериментальной статьи. Говорим о том, как написать статью. ДЗ: закончить оформлять заметки и ответить на вопросы по статье (написать abstract).
3. Начинаем читать статью про тул. ДЗ: дочитать статью самостоятельно и начать заполнять заметки.
4. Заканчиваем читать и обсуждать статью про тул. Говорим о том, как писать такие статьи. ДЗ: закончить оформлять заметки и ответить на вопросы по статье.
5. Начинаем читать обзор. ДЗ: дочитать статью самостоятельно, начать заполнять заметки и опросник.
6. Заканчиваем читать и обсуждаем обзор. Обсуждаем, что должно быть в хорошем обзоре и вообще подводим итоги курса. Обсуждаем, кто как вёл

заметки по статьям и чему мы можем друг у друга поучиться. ДЗ: закончить заметки и опросник.

Итоги обучения

Участники курса:

- научатся или прокачают свой навык вдумчиво читать англоязычные статьи и извлекать из них максимум информации;
- научатся вести заметки по статьям;
- узнают общие принципы написания разных типов статей;
- потренируются писать короткие англоязычные тексты.

Аттестация

Общий балл складывается из работы на занятиях (36%), заметок по статьям (30%), написания небольших текстов (34%) и онлайн-курса (см. ниже; можно получить бонусом до 36,5%). Для получения зачёта нужно набрать не менее 70 %. Работы, сданные после дедлайна, но до окончания курса, засчитываются со штрафом 50%.

Другое

[Онлайн-курс по английскому языку для биологов](#) не требуем (он довольно длинный), но рекомендуем. Частичное или полное прохождение курса будет бонусом к общему баллу.

Направление “Алгоритмическая биоинформатика”

Осенний семестр

Курс по статистике и теории вероятностей идет параллельно у студентов Санкт-Петербурга и онлайн-направления. Молекулярная биология и алгоритмы – общие занятия для всех.

Статистика, осень (Онлайн)

Преподаватели: Юрий Белоусов, НИУ ВШЭ; Владимир Пирогов, Университет ИТМО

Описание курса и темы занятий

Курс посвящен базовым понятиям математической статистики. Наибольшее внимание будет уделено точечным оценкам и построению критериев для проверки статистических гипотез.

1. Понятие выборки. Генеральные и выборочные характеристики. Связь между ними. Эмпирическое распределение.
2. Точечные оценки. Требования к статистическим оценкам. Методы построения. Примеры оценок.
3. Интервальное оценивание. Доверительный интервал, асимптотический доверительный интервал. Точные доверительные интервалы для среднего и дисперсии в нормальной модели. Построение асимптотических доверительных интервалов для среднего.
4. Проверка статистических гипотез. Мощность, ошибка первого и второго рода, p -значения, консервативные и радикальные критерии. Связь проверки статистических гипотез и построения доверительных интервалов. Асимптотические критерии. Проверка сложных гипотез. Непараметрические критерии. Равномерно наиболее мощные критерии.
5. Моделирование распределений (дискретных и непрерывных).
6. Корреляционный анализ. Различные коэффициенты корреляции. Проверка гипотез о значимости коэффициента корреляции.

Требования к ПО

Знаний определенных языков программирования не требуется, задания можно будет делать на любом удобном ЯП.

Аттестация

За выполнение домашних заданий ставятся баллы. Итоговая оценка выставляется по доле от максимально возможного количества баллов:

- до 65% - неудовлетворительно,
- от 65% до 80% - удовлетворительно,
- от 80% до 90% - хорошо,
- от 90% - отлично.

Дополнительные материалы и полезные ссылки

[Онлайн-курс А. И. Храброва на Stepik](#)

Рекомендации по дополнительной литературе будут выдаваться во время семестра.

Теория вероятностей, осень (Санкт-Петербург)

Преподаватели: Антон Коробейников, ЦАБ и кафедра статистического моделирования СПбГУ; Владислав Страшко, BIOCAD

Описание курса и темы занятий

1. Элементарные исходы, пространство элементарных исходов, события. Сигма-алгебры, их свойства. Случайная величина. Сигма-алгебра случайной величины. Понятие измеримости. Мера и вероятность
2. Распределения случайных величин. Виды распределений: дискретное, абсолютно непрерывное. Свойства плотности. Функция распределения, ее свойства. Теорема существования случайной величины с заданной функцией распределения. О (не-)существовании равномерной вероятности.
3. Примеры распределений
4. Многомерные распределения. Совместное распределение. Свойства.
5. Примеры многомерных распределений
6. Независимость случайных величин
7. Числовые характеристики распределений
 - a. Математическое ожидание случайной величины
 - b. Свойства математического ожидания
 - c. Дисперсия и моменты старших порядков
 - d. Свойства дисперсии
 - e. Числовые характеристики зависимости
8. Марковские цепи
9. Сходимость последовательностей случайных величин

- a. Сходимости “почти наверное” и “по вероятности”
 - b. Неравенства Чебышева
 - c. Законы больших чисел
10. Центральная предельная теорема и все-все-все
- a. Слабая сходимость (“по распределению”)
 - b. Центральная предельная теорема
 - c. Предельная теорема Муавра-Лапласа
 - d. Теорема Пуассона

Итоги обучения

1. Знакомство с базовыми понятиями классической статистики: выборки, оценки, статистические гипотезы
2. Изучение свойств базовых понятий, их границ применимости
3. Умение применять методы классической статистики к решению прикладных задач
4. Умение проверять теоретико-вероятностные и статистические факты и свойства моделированием

Аттестация

- Общий балл будет складываться из баллов за выполнение самостоятельных заданий: домашних работ (45% от общего балла), квизов (15% от общего балла) и контрольных работ (40% от общего балла).
- Работы с дедлайном (домашние) присланные после дедлайна не проверяются и не оцениваются.
- Пересчет баллов за курс в оценки осуществляется по следующей шкале:
 - При получении $\geq 85\%$ баллов ставится оценка “5”.
 - При получении $\geq 65\%$ баллов ставится оценка “4”.
 - При получении $\geq 40\%$ баллов ставится оценка “3”.

Молекулярная биология, осень

Преподаватель: Юрий Барбитов, Институт биоинформатики

Описание курса и темы занятий

Общая цель курса: познакомить студентов с базовыми понятиями молекулярной биологии, научить ориентироваться в научной литературе и понимать биологическую мотивацию и основные выводы из научных исследований. Курс будет посвящен общим аспектам строения клетки и различных процессов в ней. Помимо вопросов организации и функционирования генома, несколько занятий

будет посвящено работе белков, их взаимосвязям и разнообразным молекулярным процессам в клетке.

Курс будет построен вокруг [онлайн-курса “Молекулярная биология клетки” на Stepik](#)

Каждое занятие будет посвящено одной из тем из онлайн-уроков с разбором дополнительных подтем. На каждом занятии будет проводиться небольшая проверочная работа по пройденному материалу (в отдельном курсе на Stepik). Также предусмотрено несколько семинаров, на которых студентам будет предложено сделать короткое сообщение по теме.

Темы занятий:

1. Введение в молекулярную биологию, основные понятия.
2. Структура ДНК и ее воспроизведение.
3. Стабильность генома и мутагенез (+ семинар)
4. 3Д-структур генома.
5. Жизненный цикл РНК и экспрессия генов.
6. Трансляция и структура белков.
7. Генетика сложных признаков (семинар)
8. Клеточный обмен веществ.
9. Организация тканей и межклеточные взаимодействия.
10. Клеточный цикл и онкогенез. Наследственные болезни человека (+ семинар)

Итоги обучения

Общая цель курса: познакомить студентов с базовыми понятиями молекулярной биологии, научить ориентироваться в научной литературе и понимать биологическую мотивацию и основные выводы из научных исследований. Курс будет посвящен общим аспектам строения клетки и различных процессов в ней. Помимо вопросов организации и функционирования генома, несколько занятий будет посвящено работе белков, их взаимосвязям и разнообразным молекулярным процессам в клетке.

Требования к ПО

Требуется ноутбук для прохождения заданий на Stepik

Аттестация

Итоговая оценка будет складываться из следующих компонентов:

1. Прохождение онлайн-курса на Stepik – 30 баллов. Балл со Stepik пересчитывается по пропорции 1:5.
2. Проверочные работы на занятиях – 28 баллов. Необходимо написать все проверочные вне зависимости от балла.
3. Домашние задания – 22 балла;
4. Итоговая работа – 20 баллов. Ответы на вопросы по статье в письменном и устном виде.

Итоговая оценка выставляется следующим образом? 90 и более баллов – 5, 80 и более баллов – 4, 60 и более баллов – 3. При получении менее, чем 60 баллов, зачет по курсу не выставляется.

Алгоритмы в биоинформатике, осень

Преподаватель: Екатерина Носкова, Университет ИТМО

Описание курса и темы занятий

Курс создан для введения слушателей в курс основных алгоритмов и принципов в биоинформатике.

Курс состоит из двух частей:

1. Введение в биоинформатику – история исследования ДНК и белковых структур.
2. Основные алгоритмы, используемые в биоинформатике.

План занятий:

1. (11.09) История - 1 - история исследования протеинов
2. (18.09) История - 2 - история исследования ДНК
3. (25.09) Выравнивание 1
 - а. Центральная догма молекулярной биологии
 - б. Что такое выравнивание последовательностей.

- c. Алгоритм Нидлмана-Вунша
 - d. Типы матриц весов (BLOSUM и PAM)
 - e. Домашнее задание 1 (3 задачи)
- 4. (02.10) Выравнивание 2
 - a. Квиз 1
 - b. Алгоритм Смита-Ватермана
 - c. Аффинные штрафы. Алгоритм 3-level Manhattan
 - d. Алгоритм Хиршберга
 - e. Домашнее задание 2 (2 задачи)
- 5. (16.10) Поиск в базе
 - a. Квиз 2
 - b. Понятие n-gram
 - c. Алгоритм FASTA
 - d. Алгоритм BLAST
 - e. Алгоритмы поиска подстрок: Рабина-Карпа, Кнута-Морриса-Пратта, Ахо-Корасика
 - f. Суффиксное дерево, суффиксный массив
 - g. Домашнее задание 3 (1 задача)
- 6. (23.10) РНК (методы работы с вторичной структурой)
 - a. Квиз 3
 - b. Что такое РНК и какие они бывают
 - c. Понятие вторичной структуры
 - d. Алгоритм максимизации спаренных пар (Нуссинов)
 - e. SOTA алгоритмы
 - f. Домашнее задание 4 (1 задача)
- 7. (30.10) Деревья, множественные выравнивания
 - a. Квиз 4
 - b. Понятие филогенетических деревьев, способы их представления.
 - c. Основные метрики качества филогенетических деревьев
 - d. Алгоритм WPGMA
 - e. Алгоритм UPGMA
 - f. Алгоритм Neighbor Joining
 - g. Глобальное выравнивание n последовательностей: n-мерные матрицы
 - h. Понятие профиля выравнивания. Понятие консенсусной последовательности. Жадный алгоритм.
 - i. Алгоритм ClustalW.
 - j. Домашнее задание 5 (3 задачи)
- 8. (06.11) Промежуточный тест, Практика/вопросы
- 9. (13.11) НММ

- a. Квиз 5
 - b. Задача о кривой монете
 - c. Марковский процесс
 - d. Скрытая марковская модель
 - e. Алгоритм Витерби
 - f. Алгоритм Forward-Backward
 - g. Алгоритм Баума-Велча
 - h. HMM профиль
 - i. HNAAlign (Clustal Omega)
 - j. Домашнее задание 6 (3 задачи)
10. (20.11) Сборка геномов
- a. Квиз 6
 - b. Задача сборки геномов
 - c. Жадный алгоритм
 - d. Алгоритм Overlap Layout Consensus
 - e. Графы де Брюйна
11. (27. 11) Финальный тест

Итоги обучения

По окончании курса студенты будут понимать принципы работы основных базовых алгоритмов биоинформатики и мотивацию стоящую за дизайном этих алгоритмов, с биологической точки зрения.

Требования к ПО

Знаний определенных языков программирования не требуется, задания можно будет делать на любом удобном ЯП. Предпочтительно – Python.

Аттестация

Квизы (6 квизов по 5 вопросов): **60 баллов** + немного дополнительных

Домашние задания (6): **240 баллов** (+ дополнительные (пока их нет)). У каждого задания – свой мягкий и жесткий дедлайн. После мягкого дедлайна – половина баллов, после жесткого - задание не принимается.

1. 3 задачи (20 + 20 + 20 баллов)

2. 2 задачи (10 + 20 баллов)
3. 1 задача (10 баллов)
4. 1 задача (20 баллов)
5. 3 задачи (20 + 20 + 20 баллов)
6. 3 задачи (20 + 20 + 20 баллов)

Промежуточный и финальный тесты: **100 баллов**

Мах: 400 баллов

5 - 330 (83%)

4 - 280 (71%)

3 - 200 (50%)

Практикум по биоинформатике, осень

Преподаватели: Михаил Райко, ЦАБ СПбГУ; Юрий Барбитов, Институт биоинформатики; Алиса Моршнева (ассистент), ЦИН РАН

Описание курса и темы занятий

Курс состоит из нескольких практических заданий, которые выполняются в смешанных группах. Задания будут основаны на реальных данных, и посвящены различным аспектам биоинформатики (в основном анализ данных NGS).

Темы проектов

1. Выравнивание на референс и поиск SNP
2. Сборка геномов de novo
3. Аннотация геномов - структурная и функциональная
4. Филогенетика
5. РНК-секвенирование и дифференциальная экспрессия генов
6. Аннотация снипов у человека
7. Метагеномика

Итоги обучения

В рамках курса студенты научатся:

1. Использовать различные биоинформатические инструменты для решения конкретных биологических задач
2. Выбирать подходящие тулы под задачи, задавать осмысленные вопросы и тестировать гипотезы.
3. Писать связные осмысленные тексты, что было сделано и зачем

Требования к ПО

Все проекты подразумевают запуск тулов, поэтому потребуется доступ к компьютерам с Linux. Данные будут не очень большие, мощности среднего ноутбука должно хватить.

Аттестация

По каждому проекту надо написать отчёт. Отчёт представляет собой мини-статью, на 1-2 страницы, с введением, методами, результатами и т.п.

За каждый отчёт можно получить максимум 10 баллов. Можно добрать 2 дополнительных балла за “лабораторный журнал”. Один проект с минимальным числом баллов будет отброшен. Ещё 1 дополнительный балл, если отчёт написан на английском.

Для зачёта нужно получить к концу семестра не менее 70% от максимального количества баллов.

Весенний семестр

Статистика, весна (Онлайн)

Преподаватели: Юрий Белоусов, НИУ ВШЭ; Владимир Пирогов, Университет ИТМО; Антон Ершов, НИУ ВШЭ

Описание курса и темы занятий

Курс является логическим продолжением прошлого семестра и посвящен чуть более продвинутым, но все-таки базовым понятиям математической статистики.

Темы:

1. Регрессионный анализ: оценка метода наименьших квадратов, обобщенный метод наименьших квадратов, проверка гипотез для коэффициентов в модели линейной регрессии, доверительный эллипсоид, шаговые регрессии, ридж-оценки, автокоррелированные ошибки, гетероскедастичные ошибки, мнк оценки в случае вырожденных матриц плана (псевдообратная матрица и обобщенно обратные матрицы).
2. Однофакторный и многофакторный дисперсионный анализ, F-тест.
3. Проблема множественных тестов: поправки Бонферрони и Зидака, алгоритм Хольма, критерий Тьюки.
4. Введение в байесовскую статистику.

Аттестация

За выполнение домашних заданий ставятся баллы. Оценка выставляется по доле от максимально возможного количества баллов:

- до 65% - неудовлетворительно,
- от 65% до 80% - удовлетворительно,
- от 80% до 90% - хорошо,
- от 90% - отлично.

Статистика, весна (Санкт-Петербург)

Преподаватели: Антон Коробейников, ЦАБ СПбГУ; Влад Страшко, BIOCAD

Описание курса и план занятий

1. Понятие выборки. Генеральные и выборочные характеристики. Связь между ними. Эмпирическое распределение.
2. Выборочные характеристики как оценки генеральных. Свойства.
3. Точечные оценки. Состоятельность. Несмещенность. Метод моментов (выборочных характеристик). Метод максимума правдоподобия.
4. Разложение дисперсии. Состоятельность выборочной дисперсии. Примеры оценок по методу моментов.
5. Интервальное оценивание. Понятие доверительного интервала. Распределения, связанные с нормальными. Построение доверительных интервалов через центральные статистики. Точные доверительные

интервалы для среднего и дисперсии в нормальной модели.

Асимптотические доверительные интервалы. Построение асимптотических доверительных интервалов для среднего: в общем случае, частный случай выборки из биномиального распределения (вероятность успеха в испытаниях Бернулли).

6. Неравенство Рао-Крамера. Эффективность. Сверхэффективные оценки. Доверительные интервалы. Примеры.
7. Общие подходы к построению статистических критериев. Понятие мощности, ошибок первого и второго рода. Понятие р-значения. Связь проверки статистических гипотез и построения доверительных интервалов.
8. Одновыборочный и двухвыборочный t-критерий (критерий Уэлча). Критерий Фишера. Асимптотические критерии. Проверка гипотез относительно значения среднего (одновыборочный и двухвыборочный z-критерии, критерий пропорций).
9. Критерии согласия: критерий Колмогорова-Смирнова, критерий Крамера-Смирнова-фон Мизеса. Критерий хи-квадрат. Критерий хи-квадрат и оценки параметров (проверка сложных гипотез). Критерий хи-квадрат и выборка из абсолютно-непрерывного распределения. Критерии согласия и проверка сложных гипотез.
10. Непараметрические критерии: критерий знаков, критерий Манна-Уитни, критерий Вилкоксона. Проблема множественных сравнений: FWER, FDR, схемы критериев.

Машинное обучение, весна

Преподаватели: Елена Картышева и Алиса Аленичева, JetBrains Research

Описание курса и план занятий

1. (08.02) Лекция 1. Introduction. EDA.
HW1 (EDA)
2. (15.02) Лекция 2. kNN. Clustering. Active Learning.
Q1. HW2 (Clustering)
3. (22.02) Лекция 3. Decision Trees.
Q2. HW3 (Decision Tree)
4. (01.03) Лекция 4. Ensembles: bagging (RF), boosting (AdaBoost, gradient boosting), stacking.
Q3. HW4 (Ensembles)
5. (08.03) Лекция 5. SVM.
Q4. HW5 (SVM)

6. (15.03) Лекция 6. Regression: linear, polynomial. Bias and variance. Regularization: L1, L2. Ridge regression, LARS, LASSO, SVR. Outliers: Theil-Sen, RANSAC, Huber Regressor.
Q5, P1
7. (22.03) Лекция 7, Perceptron. Neural Nets.
Практика. PyTorch.
Q6, HW6 (Perceptron)
8. (29.03) Лекция 8, DL for Images.
Практика. Работа с изображениями.
Q7, P2
9. (05.04) Лекция 9. Генеративные модели (GANs).
Мини-проект (в группах)
10. (12.04) Лекция 10. Снижение размерности.
Практика. Снижение размерности
11. (19.04) Лекция 11. DL for text.
Чек-пойнт по проекту
Практика. работа с текстом.
P3
12. (26.04) Лекция 12. Reinforcement learning.
13. (3.05) Лекция 13. Защита мини-проектов по GAN.

Аттестация

Оценка складывается из трех элементов:

1. Квизы – 7
2. Домашки – 6
3. Практики – 3
4. Мини-проект – 1

+ 10 баллов за каждое домашнее задание (60 баллов)

+ 10 баллов за каждую практику (30 баллов)

+ 4 балла за каждый квиз (28 баллов)

+ 12 баллов за проект (аккуратность репозитория, выступление)

Всего за курс можно набрать 130 баллов:

>=110 - отлично

>=95 - хорошо

>=85 - удовлетворительно

Алгоритмы в NGS, весна

Преподаватель: Андрей Пржибельский, ЦАБ СПбГУ

Описание курса и план занятий

1. Введение. Краткая история секвенирования от Сангера до наших дней. Ключевые характеристики различных технологий секвенирования. Области применения секвенирования и актуальные задачи в биоинформатике.
2. Выравнивание коротких ридов. Программы Bowtie2 и BWA. Суффиксный массив, преобразование Борроуза-Виллера, FM-индекс и их применение в задачи выравнивания коротких ридов.
3. Выравнивание длинных ридов с высоким уровнем ошибки. Программы BLASR, hybridSPAdes, MHAP и minmap. Быстрый поиск k-меров, минимальные хеши и минимайзеры.
4. Исправление ошибок в риды Illumina. Программы Quake и BayesHammer. Исправление ридов с равномерным и неравномерным покрытием.
5. Сборка геномов. Подход Overlap-Layout-Consensus. Использование графа де Брюйна для сборки коротких ридов. Упрощение графа де Брюйна. Разрешение геномных повторов.
6. Оценка качества геномных сборок. Утилита QUAST. Основные метрики оценки качества. Альтернативные методы.
7. Транскриптомика. Сборка транскриптомов de novo и по референсному геному. Алгоритмы выравнивания ридов РНК. Задача оценки дифференциальной экспрессии генов.
8. Метагеномика. Основные задачи и проблемы в сборке метагеномов. Биннинг и классификация результатов.

Аттестация

- В курсе предполагается 7 домашних заданий. Для допуска к зачету необходимо сдать все.
- Сдача домашнего задания подразумевает его выполнение полностью.
- У каждого задания есть дедлайн - 15 дней от пары, на которой оно задано. В случае опоздания, на зачете будет задан дополнительный вопрос на эту тематику, но домашнее задание необходимо выполнить в любом случае.
- Если к заданию есть замечания, их надо исправить до зачета.
- В курсе также будут два теста. Для допуска к зачету необходимо написать оба минимум на 66%.
- Если все задания и тесты сданы, студент допускается к устному зачету.

Практикум по биоинформатике, весна

Преподаватели: Михаил Райко, ЦАБ СПбГУ; Юрий Барбитов, Институт биоинформатики; Алиса Моршнева (ассистент), ИНЦ РАН

Описание курса и план занятий

Курс состоит из нескольких практических заданий, которые выполняются в смешанных группах. Задания будут основаны на реальных данных, и посвящены различным аспектам биоинформатики (в основном анализ данных NGS).

Темы проектов второго семестра:

1. Человеческая генетика, аннотация SNP, ассоциации
2. РНК-секвенирование и дифференциальная экспрессия генов
3. Метагеномика (ампликонная и shotgun)
4. Иммуногеномика, RepSeq
5. Протеомика

Итоги обучения

В рамках курса студенты научатся:

1. Использовать различные биоинформатические инструменты для решения конкретных биологических задач
2. Выбирать подходящие тулы под задачи, задавать осмысленные вопросы и тестировать гипотезы.
3. Писать связные осмысленные тексты, что было сделано и зачем

Требования к ПО

Все проекты подразумевают запуск тулов, поэтому потребуется доступ к компьютерам с Linux. Данные будут не очень большие, мощности среднего ноутбука должно хватить.

Аттестация

По каждому проекту надо написать отчёт. Отчёт представляет собой мини-статью, на 1-2 страницы, с введением, методами, результатами и т.п.

За каждый отчёт можно получить максимум 10 баллов. Можно добрать 2 дополнительных балла за “лабораторный журнал”. Один проект с минимальным числом баллов будет отброшен. Ещё один дополнительный балл, если отчёт написан на английском.

Для зачёта нужно получить к концу семестра не менее 70% от максимального количества баллов.

Молекулярная эволюция, весна

Преподаватель: Юрий Барбитов, Институт биоинформатики

Описание курса и план занятий

1. Введение. Теория эволюции. Генетическое разнообразие популяций.
2. Генетика популяций. Закон Харди-Вайнберга.
3. Эффекты случайного мутагенеза и потока генов на частоты аллелей в популяции.
4. Дрейф генов. Модель Райта-Фишера. Естественный отбор и его математическое описание.
5. Геномный анализ популяционной структуры. Генетическое смешение. F-статистики и ABBA/BABA тест. ADMIXTURE.
6. Методы оценки давления отбора на последовательности генов и геномов. dN/dS , Tajima's D, EHH.
7. Анализ демографической истории популяций. Коалесцентная теория. Skyline-методы. Группа методов на основе SMC.
8. Геномные тренды в эволюции. Селективное давление против новых мутаций в простых и сложных геномах.

Онлайн-часть:

1. Курс “Молекулярная филогенетика”, <https://stepik.org/course/2054/>
2. Курс “Генетика и геномика популяций”, <https://stepik.org/course/9182/>
3. Методы построения и проверки филогенетических деревьев. Методы расстояний и дискретные методы. Байесовский подход.

Молекулярная филогенетика, весна

Преподаватель: Полина Дроздова, НИИ Биологии ИГУ

Описание курса и план занятий

Курс состоит из онлайн-курса (самостоятельное изучение) и практикумов, на которых мы сосредоточимся на применении программ. По результатам практикумов нужно будет написать мини-отчёты по предоставленному плану отчета.

Темы занятий:

1. Введение. Работа с деревьями. Как построить дерево в R & Python. Будем работать с: R (ape, ggtree); python (Biopython, etc3). HW1: *мини-отчёт №1 + модуль 1 онлайн-курса.*
2. Откуда брать данные, базы данных, встроенные инструменты, blast и всё такое. Учимся работать с: ncbi-blast+, ncbi eutils. HW2: *мини-отчёт №2 + модуль 2 онлайн-курса.*
3. Выравнивание нуклеотидных и белковых последовательностей. Будем работать с: clustalw, clustalO, muscle, mafft, kalign, tcoffee, prank. HW3: *мини-отчёт №3 + модуль 3 онлайн-курса, уроки 1-4.*
4. Выбор модели замен и построение дерева. Учимся работать с: ParitionFinder, jModelTest2, RAxML, IQtree. HW4: *мини-отчёт №4 + модуль 3 онлайн-курса, уроки 5-9.*
5. Байесовские методы и датировка. Учимся работать с: *BEAST. HW5: *модуль 4 онлайн-курса + мини-отчёт №5.*
6. Маркерные гены. Филогеномика. Обсуждение итогов курса.

Самостоятельное изучение:

- [Онлайн-курс “Молекулярная филогенетика”](#)

Итоги обучения

Участники курса:

- узнают, как работает филогенетический анализ:
 - где хранятся данные последовательностей и как извлекать данные из баз;

- какие данные стоит использовать для филогенетического анализа и почему;
- как работает выравнивание;
- что нужно, чтобы построить дерево на основе выравнивания;
- какие бывают способы построения деревьев;
- как оценить качество полученной филогении.
- на практике научатся работать с основным набором программ, популярных в современной филогенетике:
 - извлечения данных из NCBI;
 - выравнивания нуклеотидных и белковых последовательностей;
 - подбора оптимальной модели эволюции последовательностей;
 - построения филогенетических деревьев;
 - отрисовки деревьев.

Аттестация

- Общий балл будет складываться из балла за курс (50 %) и баллов за мини-отчёты (в сумме дают 50 %).
- При получении $\geq 90\%$ баллов ставится оценка “5”.
- При получении $\geq 75\%$ баллов ставится оценка “4”.
- При получении $\geq 60\%$ баллов ставится оценка “3”.
- При сдаче отчёта после дедлайна, но до конца курса оценка за него будет снижена в два раза. То же верно про сдачу теоретической части (онлайн-курса).

Журнальный клуб, весна (Санкт-Петербург)

Преподаватель: Александр Ткаченко, Университет ИТМО

Описание курса и темы занятий

Научный семинар, который объединяет потоки биологов и информатиков для рассмотрения актуальных научных статей по биоинформатике и связанным темам.

В течение этого семестра каждый из студентов выступает с докладом по мотивам научной статьи. Темы занятий и статьи выбирает и распределяет между участниками куратор. В подготовке доклада принимает участие два человека:

обычно биолог и информатик, но не всегда. Вместе им необходимо прочитать и понять предложенную статью, а затем донести ваше понимание до аудитории.

Темы занятий:

- Single-cell sequencing
- Metagenomics
- Protein design
- Proteomics
- Population genomics
- Nanopore sequencing and genome assembly

Аттестация

Критерий зачета: один доклад за семестр.

Журнальный клуб, весна (Онлайн)

Преподаватель: Полина Дроздова, НИИ Биологии ИГУ

Описание курса и темы занятий

Научный семинар, который объединяет потоки биологов и информатиков для рассмотрения актуальных научных статей.

- Семинары **не будут** проводиться на английском языке, но мы будем много внимания обращать на язык: читать вслух, обсуждать особенности научного английского, термины, произношение и принципы написания научных статей.
- За эти 6 занятий мы рассмотрим 3 статьи: одну экспериментальную статью, один обзор и одну статью про программу или алгоритм.
- Ваши предложения статей на обсуждение всегда приветствуются! Статью будем выбирать голосованием (окончательное решение остаётся за преподавателем). Критерии подходящей статьи: она должна быть англоязычной, интересной, желательно цитируемой, желательно с открытым доступом (или легко доступна на другом ресурсе, если журнал нас не любит).

План занятий:

7. Вводное занятие. О том, как улучшать свой английский и что нужно для свободного общения в науке. Начинаем читать и обсуждать экспериментальную статью. *ДЗ: дочитать статью самостоятельно, можно начать заполнять заметки по ней.*
8. Заканчиваем обсуждение экспериментальной статьи. Говорим о том, как написать статью. *ДЗ: закончить оформлять заметки и ответить на вопросы по статье (написать abstract).*
9. Начинаем читать статью про тул. *ДЗ: дочитать статью самостоятельно и начать заполнять заметки.*
10. Заканчиваем читать и обсуждать статью про тул. Говорим о том, как писать такие статьи. *ДЗ: закончить оформлять заметки и ответить на вопросы по статье.*
11. Начинаем читать обзор. *ДЗ: дочитать статью самостоятельно, начать заполнять заметки и опросник.*
12. Заканчиваем читать и обсуждаем обзор. Обсуждаем, что должно быть в хорошем обзоре и вообще подводим итоги курса. Обсуждаем, кто как вёл заметки по статьям и чему мы можем друг у друга поучиться. *ДЗ: закончить заметки и опросник.*

Итоги обучения

Участники курса:

- научатся или прокачают свой навык вдумчиво читать англоязычные статьи и извлекать из них максимум информации;
- научатся вести заметки по статьям;
- узнают общие принципы написания разных типов статей;
- потренируются писать короткие англоязычные тексты.

Аттестация

Общий балл складывается из работы на занятиях (36%), заметок по статьям (30%), написания небольших текстов (34%) и онлайн-курса (см. ниже; можно получить бонусом до 36,5%). Для получения зачёта нужно набрать не менее 70 %. Работы, сданные после дедлайна, но до окончания курса, засчитываются со штрафом 50%.

Другое

[Онлайн-курс по английскому языку для биологов](#) не требуем (он довольно длинный), но рекомендуем. Частичное или полное прохождение курса будет бонусом к общему баллу.