**Правительство Российской Федерации**

**Федеральное государственное автономное образовательное учреждение высшего профессионального образования
«Национальный исследовательский университет
«Высшая школа экономики»**

Факультет Компьютерных наук

Департамент больших данных и информационного поиска

Базовая кафедра Яндекс

УТВЕРЖДАЮ

Академический руководитель

образовательной программы

«Науки о данных»

по направлению 01.04.02

 «Прикладная математика и информатика»

С.О. Кузнецов

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_ 2014 г.

**Программа дисциплины** «Анализ символьных последовательностей»

для направления 01.04.02 "Прикладная математика и информатика" подготовки магистра

для магистерской программы "Науки о данных"

**Автор программы:**

Ройтберг М.А., д.ф.-м.н. (mroytberg@lpm.org.ru)

Одобрена на заседании базовой кафедры Яндекс «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_ 2014  г.

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_\_\_\_\_ М.А. Бабенко

Рекомендована Академическим советом образовательной программы

«Науки о данных» «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_ 2014  г.

Менеджер базовой кафедры Яндекс \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Е.Ф. Баулин

Москва, 2014

*Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения подразделения разработчика программы.*

Пояснительная записка

## Автор программы

Ройтберг М.А., д.ф.-м.н.

## Требования к студентам

## Изучение курса «Анализ символьных последовательностей» требует предварительныхзнаний в следующих областях: информатика, дискретная математика. Необходимые сведения из биологии будут сообщены в ходе курса. Желательно знание основ молекулярной биологии и генетики в объеме курса средней школы

## Аннотация

Дисциплина «Анализ символьных последовательностей» предназначена для подготовки магистров по направлению 01.04.02 «Прикладная математика и информатика».

Методы анализа символьных последовательностей развивались как в прикладныхисследованиях (коммуникационные системы, анализ больших биологических молекул;разработка систем акустической диагностики и систем автоматического распознавания речии др.), так и в теоретических исследованиях (теория автоматов, теория формальныхграмматик, алгоритмы поиска вхождений и т.п.). В 80-е годы прошлого столетия былиосознаны связи между методами анализа символьных последовательностей в разныхобластях и сделаны первые попытки вычленить общие их основы.

Изучение аналогий между проблемами и методами анализа символьныхпоследовательностей из различных прикладных областей часто является весьмапродуктивным. В частности, для анализа текстов на естественных языках, которые являютсяодним из главных объектов анализа в информационных технологиях, полезно знакомство сметодами и подходами к анализу символьных последовательностей, выработанными биоинформатике. Это связано как с аналогиями между биологическими и естественнымитекстами, так и, прежде всего, с тем, что именно в биоинформатике разработка методованализа последовательностей в последние годы происходила наиболее интенсивно.

Программа курса предусматривает лекции (32часа) и практические занятия (32 часа).

## Учебные задачи курса

Цели курса:

* дать представление о методах анализа символьных последовательностей,

применяемых в биоинформатике;

* проследить аналогии между проблемами и методами анализа символьных

последовательностей в области анализа текстов и в биологии;

* выделить эффективные методы биоинформатики, аналоги которых отсутствуют вобласти анализа текстов, но построение которых представляется перспективным дляанализа текстов.

Тематический план дисциплины «Робастные методы в статистике»

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№** | **Название темы** | **Всего часов по дисциплине** | **Аудиторные часы** | **Самосто-ятельная работа** |
| **Лекции** | **Сем. и практика занятия** |
| 1 | Тема 1. Парное сравнение последовательностей  | 50 | 10 | 10 | 30 |
| 2 | Тема 2.Множественное сравнение последовательностей  | 52 | 10 | 10 | 32 |
| 3 | Тема 3.Вероятностные модели семействпоследовательностей.  | 60 | 12 | 12 | 36 |
|  | Итого | 162 | 32 | 32 | 98 |

# Источники информации

## Список литературы

### Основная литература

### 1. Durbin, R., Eddy, S., Krogh, A., Mitchison, G. Biological Sequence Analysis: Probabilistic

### Models of Proteins and Nucleic Acids.[Естьрусскийперевод].

### 2. Gusfield, D. Algorithms on Strings, Trees and Sequences. Computer Science and

### Computational Biology.[Есть русский перевод].

### Дополнительнаялитература

1. Sankoff, D. and Kruskal, J. (eds.) Time Warps, String Edits, and Macromolecules: TheTheory and Practice of Sequence Comparision. Addison-Wesley Publishing Co.,

Reading, Massachusetts, 1983.А.Н. Вероятность.М.:Наука. 1980.

### 2. Ройтберг М.А. Биоалгоритмика. "Компьютерра" №36 (413), 24.09.2001

(http://offline.computerra.ru/2001/413/12801/)

# Формы контроля и структура итоговой оценки

• Текущий контроль: - письменная аудиторная контрольная работа (60 мин.) и индивидуальное домашнее задание.

• Итоговый контроль – письменный экзамен (120 мин.)

Формирование оценки.

Оценка работы студентов на семинарских и практических занятиях, *Оаудиторная,*, формируется по десятибалльной шкале и выставляется рабочую ведомость перед итоговым контролем. При формировании оценки учитывается: активность на семинарских занятиях, правильность решения задач на семинаре, результаты письменных тестовых опросов.

Результирующая оценка за текущий контроль в первом модуле учитывает результаты студента по текущему контролю следующим образом:

*Отекущий* = 0,6*·Ок/р+ 0,4·Оаудиторна* ;

Результирующая оценка за итоговый контроль в форме экзамена выставляется по следующей формуле, где *Озач* – оценка за работу непосредственно на зачете:

*Оитоговый1 =0,4·Озач +0,6·Отекущий·*

Результирующая оценка за текущий контроль во втором модуле учитывает результаты студента по текущему контролю следующим образом:

*Отекущий* = 0,6*Одз+ 0,4·Ок/р*;

Результирующая оценка за итоговый контроль в форме экзамена выставляется по следующей формуле, где *Оэкзамен* – оценка за работу непосредственно на экзамене:

*Оитоговый =0,4·Оэкзамен +0,3·Отекущий +0,3·Оитоговый1.*

В диплом ставится оценка за итоговый контроль, которая является результирующей оценкой по учебной дисциплине.

### Таблица соответствия оценок по десятибалльной и системе зачет/незачет

|  |  |
| --- | --- |
| **Оценка по 10-балльной шкале** | **Оценка по 5-балльной шкале** |
| 1 | Незачет |
| 2 |
| 3 |
| 4 | Зачет |
| 5 |
| 6 |
| 7 |
| 8 |
| 9 |
| 10 |

### Таблица соответствия оценок по десятибалльной и пятибалльной системе

|  |  |
| --- | --- |
| **По десятибалльной шкале** | **По пятибалльной системе** |
| 1 – неудовлетворительно2 – очень плохо3 – плохо | неудовлетворительно – 2 |
| 4 – удовлетворительно5 – весьма удовлетворительно | удовлетворительно – 3 |
| 6 – хорошо7 – очень хорошо | хорошо – 4 |
| 8 – почти отлично9 – отлично10 – блестяще | отлично – 5 |

# Программа дисциплины «Многомерный статистический анализ»

**Тема 1. Парное сравнение последовательностей**

1.1. Основные задачи анализа последовательностей и их отражение в биоинформатике:

- сопоставление в целом (парное, множественное),

- определение количественной меры сходства последовательностей в целом;

- поиск общих мотивов (в двух и многих последовательностях);

- поиск в БД;

- поиск и выделение функционально значимых участков;

- разбиение последовательности на «статистически однородные» участки.

1.2. Простейшая постановка задачи парного выравнивания. Определение выравнивания последовательностей. Удаления и вставки фрагментов.Замена символов. Вес выравнивания. Оптимальное выравнивание. Алгоритм динамического программирования (ДП) для построения оптимального выравнивания для случая посимвольного удаления и вставки. Его сложность.

 1.3. Эволюционная адекватность алгоритмического выравнивания. Сравнение выравниваний. Меры качества алгоритмического выравнивания: точность (какая часть эталонного выравнивания воспроизведена алгоритмически) и достоверность (какая часть алгоритмического выравнивания воспроизводит эталонное выравнивание). Ограниченность простейшей постановки задачи выравнивания.

1.4. Уточнение постановки задачи парного выравнивания: учет специфики последовательностей. Матрица замен. Линейная весовая функция для удаления и вставок фрагментов. Алгоритм. Поиск оптимального глобального и оптимального локального выравнивания;односторонне-глобальное выравнивание. Алгоритмы.

1.5. Другие методы сопоставления последовательностей в целом Анализ множества всех возможных выравниваний.Неоднозначность оптимального выравнивания. Оценка «вероятности» сопоставлениякаждой пары позиций сравниваемых последовательностей. Обобщение ДП-алгоритма. Нелинейные весовые функции для удалений и вставок фрагментов.Почему такие весовые функции осмысленны. ДП-алгоритм построения оптимальноговыравнивания, его сложность.

 \*Разреженное динамическое программирование.

\* Многокритериальный подход к построению выравниваний. Примеры векторных (многокритериальных) оценок выравнивания. Парето-оптимальное выравнивание. Алгоритм построения множества Парето-оптмальныхвыравниваний.

1.6. Задача поиска всех локальных сходств. Биологические предпосылки постановки задачи. Общее определение локального сходства. Примеры.Двух-этапный приближенный алгоритм решения.Идея алгоритма: 1) предварительный поиск относительно сильных и легко находимых«затравочных» сходств; 2) поиск тех локальных сходств, которые содержат затравочноесходство. Достоинства (быстрота) и недостатки (возможность потери локальных сходств).Затравки.Пример затравки: серия совпадений. Характеристики затравок: избирательность(вероятность появления затравочного сходства в случайной последовательности) ичувствительность (вероятность того, что локальное сходство интересующего нас видасодержит затравку). «Разрывные» затравки, их преимущества.

\*Алгоритм вычисления чувствительности затравки.

\*. Сравнение геномов – пример сравнения очень длинных последовательностей. Постановка задачи. Представление о строении геномов и сходстве геномов различных организмов. Более и менее консервативные участки генома. Различные подходы к выравниванию геномов. Интернет-ресурсы геномных выравниваний. Иерархический подход к выравниванию геномов. Неоправданность оптимизации целевой функции.

### Основная литература

### 1. Durbin, R., Eddy, S., Krogh, A., Mitchison, G. BiologicalSequenceAnalysis: Probabilistic

### Models of Proteins and Nucleic Acids.[Естьрусскийперевод].

### 2. Gusfield, D. Algorithms on Strings, Trees and Sequences. Computer Science and

### Computational Biology.[Есть русский перевод].

### Дополнительнаялитература

1. Sankoff, D. and Kruskal, J. (eds.) Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparision. Addison-Wesley Publishing Co., Reading, Massachusetts, 1983.А.Н. Вероятность. М.:Наука. 1980.

### 2. Ройтберг М.А. Биоалгоритмика. "Компьютерра" №36 (413), 24.09.2001. (http://offline.computerra.ru/2001/413/12801/)

## Тема 2.Множественное сравнение последовательностей

2.1. Постановка задачи поиска в базе символьных последовательностей. Алгоритм вычисления меры сходства двухпоследовательностей – функциональный параметр процедуры поиска. Требования калгоритму (быстрота, адекватность предметной области). Пример: семейство алгоритмов BLAST (базовый алгоритм BLAST и его обобщения – G-BLAST иPSI-BLAST.Препроцессинг базы последовательностей при поиске, основанном напрограмме BLAST.Оценка значимости найденных сходств. Понятие P-value. Знакомство с теориейКарлина-Альтшуля.

2.2. Множественное выравнивание последовательностей. Одновременное сопоставление (выравнивание) нескольких последовательностей.Выравнивание нескольких последовательностей как объект. Способы описания семействапоследовательностей как целого.ДП-алгоритм построения множественного выравнивания.Описание алгоритма. Сложность. Недостатки в практическом использовании.

\*Возможности ускорения для случая близких последовательностей.

2.3. Поэтапный алгоритм построения множественного выравнивания. Понятие об эволюционном (филогенетическом) дереве. Выравнивание выравниваний.Алгоритм поэтапного множественного выравнивания с предопределенным эволюционнымдеревом. Поэтапное множественное выравнивание с одновременным построениемэволюционного дерева. Достоинства и недостатки этих алгоритмов.

2.4. Профили последовательностей.Описание семейства последовательностей с помощью позиционно-зависимых матрицвесов замен (профилей). Выравнивание последовательности и профиля. Поиск локальныхсходств для профиля и последовательностей. Базы данных, содержащие семейства сходныхфункционально значимых фрагментов и их профили.

2.5. Поиск множественных локальных сходств заранее неизвестного вида.Постановка задачи. Подходы к решению. Алгоритм Gibbs-sampler. Иерархический подход к поиску фрагментов, общих для группы последовательностей.Случай неоднородного набора последовательностей. Участки, содержащиеся «почти во всех» последовательностях. Классификация набора последовательностей, отбраковка «лишних». Значимость найденныхсходств.

2.6. Примеры множественного сравнения последовательностей

Множественное сравнение геномов. Постановка задачи, трудности, связанные с размером задачи. Алгоритмы и программымножественного сравнения геномов.

Анализ множественных выравниваний. Что дает одновременный анализ нескольких последовательностей? Консервативныепозиции и участки. Точечные мутации (SNP). \*Построение профиля по множественномувыравниванию.

\* Филогенетические деревья. Понятие о филогенетическом дереве. Построение филогенетического дерева посемейству последовательностей. Надежность построения филогенетического дерева

### Основнаялитература

### 1. Durbin, R., Eddy, S., Krogh, A., Mitchison, G. BiologicalSequenceAnalysis: Probabilistic

### Models of Proteins and Nucleic Acids.[Естьрусскийперевод].

### 2. Gusfield, D. Algorithms on Strings, Trees and Sequences. Computer Science and

### Computational Biology.[Есть русский перевод].

### Дополнительнаялитература

1. Sankoff, D. and Kruskal, J. (eds.) Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparision. Addison-Wesley Publishing Co.,

Reading, Massachusetts, 1983.А.Н. Вероятность. М.:Наука. 1980.

### 2. Ройтберг М.А. Биоалгоритмика. "Компьютерра" №36 (413), 24.09.2001

(http://offline.computerra.ru/2001/413/12801/)

## Тема 3. Вероятностные модели семейств последовательностей

* 1. Что такое вероятностная модель (генератор) семейства последовательностей. Примеры:бернуллиевская модель, марковская модель порядка n; скрытая марковская модель (СММ).Вероятность последовательности относительно модели. Адекватность модели для данной выборки.

СММ и автоматы. Построение СММ по обучающей выборке.Выбор графа СММ. Подбор параметров(вероятностей перехода) для СММ. СММ как средство описания семействапоследовательностей.

* 1. Определение последовательностей состояний СММ по символьной последовательности.Символьная последовательность и порождающая ее последовательность состоянийСММ (ПС СММ). Наиболее вероятная ПС СММ для данной символьнойпоследовательности. ДП-алгоритм построения ПС СММ для данной символьнойпоследовательности. СММ как средство описания структуры последовательности.

3.3. Примеры использования СММ. Сегментация последовательностей Участки статистической однородности последовательности. Описаниеразличных участков различными вероятностными моделями. СММ, описывающаянеоднородную последовательность. Сведение задачи о сегментации к задаче построенияоптимальной ПС СММ.

3.4. Вычисление вероятности множества последовательностей.Множество последовательностей данной длины, содержащих подслово данного вида.Биологические примеры. Конечноавтоматность таких множеств.

\*Понятия об алгоритмеАхо-Корасик. СММ и распределение вероятностей на словах фиксированной длины.Понятие об алгоритме вычисления вероятности конечноавтоматного множествапоследовательностей фиксированной длины относительно заданной СММ.

\* Распознавание кодирующих последовательностей.Варианты СММ, описывающих прокариотический геном, их состояния. Учетконсенсуса границ. Качество предсказания при использовании различных моделей.

\* Обогащенные символьные последовательности

Символьные последовательности и дуговые структуры. Понятие о скобочной разметке символьной последовательности. Связь с деревьями.Примеры из лингвистики и биологии.Сравнение обогащенных символьных последовательностей. Постановка задачи о выравнивании обогащенных символьных последовательностей.Варианты постановок: глобальное и локальное выравнивание. Виды весовых функций.Понятие об алгоритмах построения оптимального выравнивании для различных постановок.

\*Предсказание вторичной структуры РНК.

### Основнаялитература

### 1. Durbin, R., Eddy, S., Krogh, A., Mitchison, G. BiologicalSequenceAnalysis: Probabilistic

### Models of Proteins and Nucleic Acids.[Естьрусскийперевод].

### 2. Gusfield, D. Algorithms on Strings, Trees and Sequences. Computer Science and

### Computational Biology.[Есть русский перевод].

### Дополнительнаялитература

1. Sankoff, D. and Kruskal, J. (eds.) Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparision. Addison-Wesley Publishing Co.,

Reading, Massachusetts, 1983.А.Н. Вероятность. М.:Наука. 1980.

### 2. Ройтберг М.А. Биоалгоритмика. "Компьютерра" №36 (413), 24.09.2001

(http://offline.computerra.ru/2001/413/12801/)

# Методические указания студентам

Самостоятельная работа студента предусматривает выполнение теоретических заданий, направленных на овладение техникой построения алгоритмов сравнения символьных последовательностей, а также практическое использование программ, реализующих эти алгоритмы.

Автор программы: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/ <РойтбергМ.А.> /