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

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 прикладных исследованиях (коммуникационные системы, анализ больших биологических молекул; разработка систем акустической диагностики и систем автоматического распознавания речи и др.), так и в теоретических исследованиях (теория автоматов, теория формальных грамматик, алгоритмы поиска вхождений и т.п.). В 80-е годы прошлого столетия были осознаны связи между методами анализа символьных последовательностей в разных областях и сделаны первые попытки вычленить общие их основы.

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

Программа курса предусматривает лекции (30 часов) и практические занятия (30 часов).

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 в области анализа текстов, но построение которых представляется перспективным для анализа текстов.

Тематический план дисциплины «Робастные методы в статистике»

№	Название темы	Всего часов по дисциплине	Аудиторные часы		Самостоятельная работа
			Лекции	Сем. и практика	
1	Тема 1. Парное сравнение последовательностей	60	10	10	44
2	Тема 2.Множественное сравнение последовательностей	60	10	10	44
3	Тема 3.Вероятностные модели семейств последовательностей.	70	10	10	42
	Итого	190	30	30	130

I. 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.[Есть русский перевод].

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. "Компьютера" №36 (413), 24.09.2001 (http://offline.computerra.ru/2001/413/12801/)

II. 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, $O_{аудиторная,}$, 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:

$$O_{текущий} = 0{,}6 \cdot O_{к/р} + 0{,}4 \cdot O_{аудиторна} ;$$

Результирующая оценка за итоговый контроль в форме экзамена выставляется по следующей формуле, где $O_{зач}$ – оценка за работу непосредственно на зачете:

$$O_{итоговый1} = 0{,}4 \cdot O_{зач} + 0{,}6 \cdot O_{текущий}.$$

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

$$O_{текущий} = 0{,}6 O_{дз} + 0{,}4 \cdot O_{к/р};$$

Результирующая оценка за итоговый контроль в форме экзамена выставляется по следующей формуле, где $O_{экзамен}$ – оценка за работу непосредственно на экзамене:

$$O_{итоговый} = 0{,}4 \cdot O_{экзамен} + 0{,}3 \cdot O_{текущий} + 0{,}3 \cdot O_{итоговый1}.$$

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 – неудовлетворительно 2 – очень плохо 3 – плохо	неудовлетворительно – 2
4 – удовлетворительно 5 – весьма удовлетворительно	удовлетворительно – 3
6 – хорошо 7 – очень хорошо	хорошо – 4
8 – почти отлично 9 – отлично 10 – блестяще	отлично – 5

III. 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» сопоставления каждой пары позиций сравниваемых последовательностей. Обобщение ДП-алгоритма. Нелинейные весовые функции для удалений и вставок фрагментов. Почему такие весовые функции осмысленны. ДП-алгоритм построения оптимального выравнивания, его сложность.

*Разреженное динамическое программирование.

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

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 затравочно есходство. Достоинства (быстрота) и недостатки (возможность потери локальных сходств).Затравки. Пример затравки: серия совпадений. Характеристики затравок: избирательность(вероятность появления затравочного сходства в случайной последовательности) и чувствительность (вероятность того, что локальное сходство интересующего нас вида содержит 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.[Есть русский перевод].

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. "Компьютера" №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 двух последовательностей – функциональный параметр процедуры поиска. Требования к алгоритму (быстрота, адекватность предметной области). Пример: семейство алгоритмов BLAST (базовый алгоритм BLAST и его обобщения – G-BLAST иPSI-BLAST. Препроцессинг базы последовательностей при поиске, основанном на программе BLAST. Оценка значимости найденных сходств. Понятие P-value. Знакомство с теорией Карлина-Альтшуля.

2.2. Множественное выравнивание последовательностей. Одновременное сопоставление (выравнивание) нескольких последовательностей. Выравнивание нескольких последовательностей как объект. Способы описания семейства последовательностей 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 эволюционным деревом. Поэтапное множественное выравнивание с одновременным построением эволюционного дерева. Достоинства и недостатки этих алгоритмов.

2.4. Профили последовательностей. Описание семейства последовательностей с помощью позиционно-зависимых матриц весов замен (профилей). Выравнивание последовательности и профиля. Поиск локальных сходств для профиля и последовательностей. Базы данных, содержащие семейства сходных функционально 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 найденных сходств.

2.6. меры множественного сравнения последовательностей

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

Анализ множественных выравниваний. Что дает одновременный анализ нескольких последовательностей? Консервативные позиции и участки. Точечные мутации (SNP). *Построение профиля по множественному выравниванию.

* Филогенетические деревья. Понятие о филогенетическом дереве. Построение филогенетического дерева по семейству последовательностей. Надежность построения филогенетического дерева

Основнаялитература

1. Durbin, R., Eddy, S., Krogh, A., Mitchison, G. BiologicalSequenceAnalysis: Probabilistic Models of Proteins and Nucleic Acids.[Есть русский перевод].

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. "Компьютера" №36 (413), 24.09.2001 (http://offline.computerra.ru/2001/413/12801/)

Тема 3. Вероятностные модели семейств последовательностей

3.1. 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 семейства последовательностей.

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

3.3. Примеры использования СММ. Сегментация последовательностей Участки статистической однородности последовательности. Описание различных участков различными вероятностными моделями. СММ, описывающая неоднородную последовательность. Сведение задачи о сегментации к задаче построения оптимальной ПС СММ.

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

*Понятия об алгоритме Ахо-Корасик. СММ и распределение вероятностей на словах фиксированной длины. Понятие об алгоритме вычисления вероятности конечно автоматного множества последовательностей фиксированной длины относительно заданной СММ.

* Распознавание кодирующих последовательностей. Варианты СММ, описывающих прокариотический геном, их состояния. Учет консенсуса 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.[Есть русский перевод].

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. "Компьютера" №36 (413), 24.09.2001 (http://offline.computerra.ru/2001/413/12801/)

IV. 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.> /