Рис. 1

Тема лекции
Компьютерный анализ информационного содержания геномов

Лектор Юрий Львович Орлов

Рис.2
Структура лекции
Обзор задач анализа информационного содержания геномов
Определение генетического текста и классификация повторов
Сложность текста. Сложностные разложения нуклеотидных последовательностей по Лемпелю-Зиву
Анализ полных геномов с помощью сложностных разложений
Контекстные деревья-источники (Марковские модели с переменной памятью) для анализа геномных последовательностей
Ссылки, литература

Рис.3.
Данная лекция посвящена
Компьютерному анализу информационного содержания геномов
Задачи исследования информационных процессов в молекулярной генетике:

Обзор задач анализа текстов ДНК:
Поиск и предсказание генов, функциональная аннотация
Генетическое рестрикционное картирование
Физическое картирование
Секвенирование
Поиск сходства
Предсказание генов
Анализ мутаций
Множественное выравнивание
Поиск сигналов в ДНК

Данная классификация задач взята из
Pevzner, Pavel. Computational Molecular Biology: an algorithmic approach. MIT Press, 2000, 311 pages.
(Павел Певзнер работал ранее в Москве, сейчас в США, это один из последних учебников по биоинформатике в мире)

Рис.4.
К числу основных задач компьютерного анализа генетических текстов в целом можно отнести:
1) предсказание кодирующих участков генов и открытых рамок считывания;
2) предсказание функциональных сигналов (функциональных сайтов и регуляторных районов);
3) статистический анализ генетических текстов, исследование структуры повторов и модели порождения символьных последовательностей, сегментация геномов;
4) анализ вторичной структуры РНК и сигналов трансляции;
5) анализ аминокислотных последовательностей глобулярных белков, предсказание вторичной структуры, функциональных сайтов и доменов глобулярных белков по их аминокислотным последовательностям.

6) поиск гомологии и выравнивание генетических текстов, филогенетические сравнения;
7) задачи оперирования с большими массивами информации и управления (Интернет-навигации) разрозненными специализированными базами данных, содержащими информацию о первичных последовательностях биополимеров и их функциональную аннотацию.

Данная, более подробная классификация, сделана на основе из сборника "Комепьютерный анализ генетических текстов" под редакцией Д.Франк-Каменецкого, 1990, Москва - давнее, но наиболее полное пособие по биоинформатике на русском языке.

Рис.5.
Пример – линейное расположение функциональных участков гена.
Вы видите на схеме промоторный район, точку старта транскрипции, чередование экзонов и интронов, сигнал терминации транскрициии.

Рис.6.
До начала эпохи массового секвенирования многим исследователям казалось, что функциональные участки будут закодированы однозначными последовательностями, скорее всего изменяющими локальные физические свойства ДНК. В действительности, проблема определения функции по последовательности ДНК гораздо сложнее, что связано с неоднозначностью кодирования генетической информации [Франк-Каменецкий М.Д., 1990]. Лишь участки действия некоторых ферментов имеют одинаковое строение (например, сайты действия рестриктаз II типа).
Пример приведен на рисунке - сайт рестрикции HindII CTGCAC, GTTAAC

Участки, которые узнают на ДНК белки-регуляторы работы генов, не столь однозначны. Их поиск в нуклеотидной последовательности - гораздо более сложная задача. Сайты начала и окончания работы РНК-полимераз, участки начала трансляции РНК, участки сплайсинга имеют весьма сложное строение. Они состоят из нескольких блоков, находящихся на варьирующих расстояниях, часто включают элементы вторичной структуры, что требует разработки специальных методов их распознавания.

Рассмотрим формальное
Понятие генетического текста
Основными структурно-функциональными компонентами живой клетки являются три класса генетических макромолекул: ДНК, РНК и белки. Генетические макромолекулы образованы из стандартных наборов мономеров - 4 типа нуклеотидов для ДНК и РНК и 20 типов аминокислот для белков. Генетическим текстом называется представление реальной макромолекулы в виде конечной последовательности символов соответствующего алфавита. Анализ таких последовательностей, представляющих иерархические уровни организации наследственной информации приведен в соответствующих главах данной работы, при описании применения разработанных компьютерных методов.
Рассмотрение молекул ДНК, РНК и белков как генетических текстов является наиболее простым представлением соответствующих реальных макромолекул. Однако такое представление отражает наиболее существенные особенности генетических макромолекул, а именно – наличие в них генетической информации и линейного способа ее записи и хранения.

Рис.7.
Определение ДНК как гетерогенной полимерной молекулы, обеспечивающей хранение и передачу генетической информации в ряду поколений сразу же ставит вопрос о важности степени гетерогенности (разнородности чередующихся символов) для надежности и

адекватности передачи такой наследственной информации. Таким образом, встает вопрос анализа сложности генетических текстов.
Кратко рассмотрим самые общие свойства ДНК в геномах. Молекулы ДНК выполняют в первую очередь функцию материальных носителей наследственной памяти, т.е. составляют структурную основу геномов. Наиболее типичны двухнитевые кольцевые или линейные молекулы геномной ДНК (двойная спираль), хотя имеются также однонитевые формы [Зенгбуш П., 1982].
Рассматривая последовательность ДНК как генетический текст, предполагается, что это последовательность ДНК от 5' к 3'-концу, соответствующая ей комплементарная последовательность в базах данных не записывается.
Особенности организации геномов и генов эукариот и прокариот были изложены в предыдущих лекциях.

Рис.8.
Свойства молекул РНК как генетических текстов

Молекулы РНК выполняют ряд ключевых функций: выступают в качестве носителей генетической информации (геномные РНК вирусов и фагов, мРНК); обеспечивают процесс трансляции (тРНК); являются компонентами клеточных субструктур, в частности рибосом (5S РНК, 16S РНК, 23S РНК), РНП-частиц (U-РНК). Для молекул РНК характерно наличие вторичной структуры, образованной взаимно комплементарными участками, формирующими двойные спирали.
Свойства белков как генетических текстов
Белки – нерегулярные полимеры, состоящие из мономеров 20 канонических типов, отличающихся по строению боковых групп. Потенциально разнообразие белков безгранично, поскольку каждому белку свойственна своя особая аминокислотная последовательность, закодированная в ДНК клетки, вырабатывающей данный белок. При биосинтезе белка основные цепи аминокислот "сшиваются" пептидными связями в единую полипептидную цепь. Расположение аминокислот в полипептидной цепи называют первичной структурой белка. Будем рассматривать ее при представлении белка как генетического текста.

Рис.9.
Молекулярно-генетические процессы как операции над генетическими текстами
(1) удвоение или мультипликация текста (репликация);
(2) перекодирование текста из одного алфавита в другой (транскрипция, обратная транскрипция, трансляция);
(3) перекомбинирование частей текстов, инверсии, дупликации, делеции, транслокации его частей, замены отдельных символов и т.д. (что соответствует законным и незаконным рекомбинациям ДНК, мутациям, процессингу, сплайсингу РНК, процессингу белков, модификациям ДНК, РНК и белков);
(4) установление парных соответствий между элементами одного или нескольких текстов (что соответствует пространственному сближению мономеров при самоорганизации);
(5) стирание текста или его фрагмента (соответствует деградации биополимера);
(6) разбиение текстов на группы (соответствует сегрегации генетических макромолекул, т.е. распределению по дочерним клеткам или компартментам одной клетки).

Рис.10.
Общими характеристиками функциональных сайтов являются:
(1) Специфичность, т.е. способность к выполнению строго определенного набора функций.
(2) Блочность, т.е. наличие в большинстве функциональных сайтов не одной, а нескольких линейно непрерывных зон, необходимых для проявления функциональной активности.

При этом положение зон относительно друг друга может варьировать в определенных пределах.

(3) Структурная (посимвольная) вырожденность функциональных сайтов.

Сайты с одинаковой функцией не идентичны по первичной структуре. Они могут отличаться как по расстоянию между блоками, так и по последовательностям отдельных блоков. В ряде случаев функциональные сайты, выполняющие идентичную функцию, могут полностью отличаться по первичной структуре. Они могут отличаться как по расстоянию между блоками, так и по последовательностям отдельных блоков. В ряде случаев функциональные сайты, выполняющие одну и ту же функцию, могут полностью отличаться по первичной структуре. Это относится как к нуклеотидным сайтам, так и к активным сайтам белков.
Представление в форме консенсуса -распространенная форма представления консервативных элементов - была описана на предыдущих лекциях.

Рис.11.
Пример представление в форме консенсуса и анализа функционального сайта:
Согласно [Bucher, 1990], строгий консенсус TATA-бокс связывающего белка (TBP) имеет вид TATAAAAA, тогда как его расширенный аналог в вырожденном 15-буквенном коде имеет вид STWTAWADRSSSSSS Анализ нуклеотидных последовательностей, обладающих повышенным сродством к TBP [Ponomarenko et al., 1997], вывил еще более длинный расширенный консенсус, имевший вид
NNNCNGSSSSCCCTTTWWWWAAAGSSSSSSSCNNG.
Заметно, что вырожденные консенсусы содержат много дополнительной информации о сайте связывания. В частности: (A+T)-богатое "ядро" консенсуса (подчеркнуто внизу) окружено G+C богатыми флангами. Правомерно предположить, что наблюдаемые в вырожденном коде закономерности важны для правильного взаимодействия ДНК TATA-бокса с белком TBP.

Рис.12.
Классификация функциональных участков генома как текстов
1) сайты линейной разметки;
2) сайты точечной разметки;
3) регуляторные элементы.
Сайты линейной разметки определяют границы протяженного участка в первичной структуре макромолекулы, вовлеченного в реализацию данного процесса. На уровне ДНК типичными сайтами этого типа являются начала репликации Ori-участки ДНК, в области которых инициируется процесс репликации. Ori про- и эукариот – это линейные сайты разметки, как правило, двухстороннего действия, так как репликация в них часто инициируется одновременно в двух направлениях.
Сайты линейной разметки одностороннего действия – промоторы и терминаторы транскрипции. На уровне РНК сайтами линейной разметки являются, например, инициирующий и терминирующий кодоны, определяющие границы области трансляции.
Сайты точечной разметки определяют границы локального участка в пределах макромолекулы, вовлеченного в реализацию данного процесса. В ДНК к ним относятся, например, сайты топоизомеразы II, осуществляющей двухнитевые разрывы ДНК и ее воссоединение в строго определенных позициях, а также рестрикционные сайты ДНК. На уровне РНК к сайтам этого типа относится, например, сигнальная последовательность Шайно-Дельгарно, которая за счет комплементарных взаимодействий с 3'-концом 16S РНК обеспечивает правильное позиционирование мРНК относительно рибосомы при инициации трансляции. В ДНК прокариот к числу таких регуляторных сайтов относятся, например, операторы, которые расположены, как правило, рядом с промоторами или

перекрываются с ними. Взаимодействуя с молекулами репрессоров или активаторов, они могут модулировать скорость транскрипции [Зенгбуш П., 1982; Стент Г., Кэлиндар Р., 1981]. Для эукариот структура регуляторных сайтов описывается более сложным образом.

Рис.13.
Взаимная совместимость генетических сообщений
Реальные генетические тексты содержат множество различных генетических сообщений [Трифонов Э.Н., 1997]. Например, в первичной структуре белка представлена информация о его пространственной структуре и локализации функциональных сайтов.
В первичной структуре мРНК, помимо информации о белке в виде последовательности кодонов, имеется следующая информация о структурно-функциональной организации мРНК:
(1) информация об особенностях инициации трансляции в виде последовательности сайта связывания рибосомы;
(2) о скорости трансляции в виде частот использования кодонов;
(3) о вторичной структуре мРНК в виде расположения взаимно комплементарных участков;
(4) об интенсивности нуклеазной деградации в виде стабильности шпилек мРНК.

Рис.14.
На уровне гена, кодирующего эту мРНК, кроме того закодирована информация о локальной конформации ДНК в виде взаимного расположения пуриновых и пиримидиновых пар [Зенгер В., 1987], а также информация о локализации нуклеосом в виде участков специфического связывания с гистонами [Trifonov E.N., 1982].
Таким образом, в пределах реального генетического текста записано, как правило, несколько генетических сообщений, определяющих различные аспекты структурно-функциональной организации макромолекул [Трифонов Э.Н., 1997]. Расположение элементов алфавита в тексте, определяемое одним генетическим сообщением может противоречить их расположению, задаваемому другим генетическим сообщением. Одновременная запись множества генетических сообщений в пределах одного текста возможна лишь в случае, если эти генетические сообщения совместимы, т.е. взаимно не искажают друг друга. Совместимость существенно зависит от такого ключевого свойства генетических сообщений, как их синонимичность. Синонимичность позволяет из множества эквивалентных по функциональной значимости вариантов выбирать взаимно совместимые.

Рис.15.
Упорядоченность структуры генетических текстов. Теоретическая классификация типов повторов
Повторы могут быть совершенными (полное совпадение) и несовершенными (допускается частичное несовпадение). Степень вырожденности (несовершенности) повтора фиксированной длины вычисляется (1) по числу несовпадений и (2) по вероятности получить такое число несовпадений по случайным причинам.
Наглядный пример повторов, связанных с информационным содержанием генетических текстов, - группировка нуклеотидов в триплеты, соответствующие кодонам. При этом взаимное расположение кодонов и их частотные характеристики определяются аминокислотной последовательностью белка. Несовершенными повторами являются также различные функциональные сайты, выполняющие идентичные функции, в том числе и промоторы.

Рис.16.

Повторы в нуклеотидных последовательностях подразделяются на прямые, инвертированные, симметрические, а также палиндромы и комплементарные палиндромы.
(Примеры приведены на рисунке)

Рис.17.
Для аминокислотных последовательностей нет отношения комплементарности. в целом для них из-за большего размера алфавита в целом характерны повторы меньшей длины, чем в нуклеотидных последовательностях.
Вероятность повтора длиной в N символов составляет 1/20**N). Таким образом, для произвольно взятой пары аминокислотных остатков (а.о.), N=2, грубая оценка вероятности равна 1/400. Если принять средний размер белка в 300 а.о., то повтор лишь двух заданных остатков в аминокислотной последовательности уже можно считать не случайным.

Рис.18.
Генетическая классификация повторов. Тандемные и диспергированные повторы
Эволюционные пути возникновения повторов можно разделить на:
(1) дупликативный;
(2) конвергентный.
Дупликативный путь появления повторов связан с удвоением генетического материала при репликации и сохранением двух копий участков ДНК.
Конвергентный путь предполагает сходные структурно-функциональные ограничения на порядок и состав элементов в определенных участках текста.
Повторы в геноме можно разделить на два класса:
1)Тандемные повторы, к которым относятся разные виды сателлитной ДНК, гены рРНК [Heslop-Harrison, 2000].
2) Диспергированные повторы, распределённые в геноме по принципу чередования с уникальными последовательностями [Smit, 1996]. К этому классу относятся, в частности, различные типы перемещающихся (мобильных) элементов [Kumor and Bennetzen, 1999].
Более подробная классификация повторов была дана в лекции Левицкого В.Г.
Следует отметить, что насыщенность геномов прокариот повторяющимися последовательностями довольно низка [Kolchanov and Lim, 1994], в то время как для геномов эукариот высокая насыщенность повторами - одно из их характерных свойств [Heslop-Harrison, 2000].

Рис.19.
О взаимосвязи с другими лекциями данного курса.
Планируемый цикл "Основы компьютерной геномики"
Комбинаторика символьных последовательностей
Теория кодирования информации
Компьютерные методы анализа генетических текстов
Большой практикум по компьютерной геномике

Перекрывание вопросов с уже прочитанными лекциями:

Лекция «Контекстный анализ и распознавание сайтов связывания транскрипционных факторов»
(Транскрипция у эукариот, строение 5'-регуляторных районов генов, задачи предсказания сайтов, использование консенсуса и весовой матрицы.)

Лекция «Механизмы регуляции транскрипции: описание в компьютерных базах данных»
(Механизмы транскрипции, классификация транскрипц. факторов по типу экспресии и механизмам активации)

Рис.20.
В связи с необходимостью придерживаться во всех лекциях единой терминологии будет полезен следующий

Список общепринятых сокращений в работах по анализу генетических текстов

ТФ - транскрипционный фактор
ССТФ - сайты связывания транскрипционных факторов
РГП - регуляторные геномные последовательности
ОРФ - открытая рамка считывания
КЭ - композиционный элемент
п.о. - пара оснований ДНК
а.о. – аминокислотный остаток
нт - нуклеотид
Распространенные англоязычные термины
HMM (Hidden Markov models) - скрытые марковские модели
PST (Probabilistic Suffix Tree) - вероятностные суффиксные деревья
VMM (Variable Memory Markov model) - марковская модель с переменной памятью

Рис.21.
Сложность генетических текстов.

Важной характеристикой символьной последовательности является ее сложность. На интуитивном уровне она отражает возможность компактного представления последовательности, основанную на выявлении тех или иных ее структурных особенностей.
Пример простой последовательности ДНК (микросателлит):
TATATATATATATATATATA
Пример реальной последовательности
caggagtTCAAGGTAAtaagg
Общий подход к определению сложности символьных последовательностей (текстов) был предложен А.Н. Колмогоровым [Колмогоров А.Н., 1965]. Существуют различные конструктивные реализации этого подхода применительно к последовательностям конечной длины для произвольных (не обязательно генетических) текстов [LempelA.&Ziv J., 1976; Rissanen J., 1986; Ebeling W. & Jimenes-Montano M.A., 1980].

Рис.22.
Сложность генетических текстов.
Подход Колмогорова предполагает существование некоторой оптимальной порождающей характеристики (алгоритма) для генерации текста из некоторой кодирующей последовательности меньшего размера.
Число операций кодирования последовательности с помощью такого алгоритма называется алгоритмической сложностью (сложностью по Колмогорову).
Еще раз подчеркнем -
Не существует рациональных алгоритмов для определения такой сложности – есть лишь различные реализации в рамках заданных ограничений.
Список международных компьютерных ресурсов по анализу сложности генетических текстов:
International resources:
http://csweb.haifa.ac.il/library/#complex Linguistic Complexity (Haifa University)
http://wwwdbl.dei.unipd.it/Verbumculus/ Verbumculus (Padova University)
http://www.biochem.ucl.ac.uk/bsm/SIMPLE/index.html SIMPLE 3.0

http://monod.uwaterloo.ca/downloads/gencompress/ GenCompress

Рис.23.
Наибольшее распространение получила мера сложности, введенная Лемпелем и Зивом (мера LZ). Каждый фрагмент имеет прообраз в последовательности и кодируется указателем на этот прообраз.
Она лежит в основе многих алгоритмов сжатия текстовой информации [Гусев В.Д. и др., 1991; Gusev V.D. et al., 1999].
Применительно к генетическим текстам (ДНК– и АМ–последовательностям) нас интересует не столько возможность их сжатия (как правило, они достаточно сложны и плохо сжимаются), сколько связанная с вычислением сложности процедура разложения (факторизации) последовательности на фрагменты, которые можно интерпретировать как элементы ее структуры (см.схему).

Основной задачей является исследование свойств факторизаций, возникающих при различных обобщениях меры сложности LZ, учитывающих специфику генетических текстов. Знание этих свойств позволяет выделять в последовательности структуры, играющие важную роль в регуляции основных генетических процессов.

Рис.24.
Сложность текста и сложностные разложения.
Сложностные разложения связаны с теорией сжатия данных.
Терминология:
Декомпозиция = разложение = факторизация
Факторизация – представление последовательности ДНК S в форме набора неперекрывающихся фрагментов S1,S2 и т.д.

Рис.25.
Лемпель и Зив предложили измерять сложность последовательности числом шагов порождающего ее процесса.
Допустимыми операциями при этом являются: генерация нового символа и копирование "готового" фрагмента из уже синтезированной части текста. Схема порождения последовательности S может быть представлена в виде конкатенации фрагментов

Схема Лемпеля и Зива: Каждый фрагмент является копией другого участка последовательности
Число фрагментов m минимально, когда длины Si максимальны-
(Числа шагов минимально при выборе максимального прототипа Si для копирования на каждом шаге.)
Расширение схемы Лемпеля и Зива для последовательностей ДНК:
Каждый фрагмент является копией или комплементарной копией 5’-района

Рис.26.
Классификация повторов в связи с декомпозицией.
Возможные операции: прямое и обратное копирование и прямое и обратное комплементарное копирование
таким образом четыре типа операций соответствуют четырем типам повторов - Прямой, Симметричный, Инвертированный, Прямой комплементарный.

Рис.27.
Рассмотрим пример сложностного разложения последовательности.
Рассмотрим модельную последовательность.

Сложностное разложение последовательности выполняется слева-направо
Сначала определяем новые символы, для которых нет прообразов слева -
G T и A.
Затем со следующей позиции находим повтор GT, это прямой повтор (Direct repeat)
Затем CT, это инвертированный повтор (Inverted repeat) участка AG
Затем симметричный (Symmetric repeat) GATG
и прямой комплементарный (Complementary repeat) CA
При разложении данной искусственной (приведенной для примера) последовательности встретились все виды повторов

Общее число операций копирования, т.е. сложность последовательности - семь.
Средняя длина компоненты в сложностном разложении - 1.85

Интерес представляет определение сложности и средней длины компонент для протяженных нуклеотидных последовательностей.

Рис.28.
Метод реализован в ИМ и ИЦиГ.
Приведен Интернет-адрес, интерфейс программы сложностных разложений.

Рис.29.
Основные применения метода сложностных разложений.
Полное сложностное разложение (декомпозиция) последовательности на повторяющиеся элементы;
Профиль последовательности в скользящем окне;
Разложение последовательности на повторяющиеся фрагменты другой последовательности; и
Анализ выборки последовательностей: Разложение последовательности в выборке на повторяющиеся фрагменты из всех остальных последовательностей

Цели использования сложностных разложений
Поиск точных повторов всех типов: прямые, симметричные, прямые и инвертированные комплементарные;
Поиск районов низкой сложности в нуклеотидных последовательностях; и
Сравнение последовательностей: поиск наибольших общих фрагментов

Область применимости:
Анализ геномных последовательностей на различных уровнях:
Короткие функциональные сайты (10-100 по);
Последовательности генов, регуляторных районов (100-10,000 по)
Полные геномы (100,000-10,000,000 по)

Рис.30.
Профиль сложности в скользящем окне позволяет найти районы низкой сложности в последовательности
Рассмотрим технику построения профилей с помощью скользящего окна.

На данном рисунке представлены профили для промотора от минус трехсот до плюс ста оснований относительно старта транскрипции [-300;+100].

Профили приведены для размера окна в десять, двадцать, сорок и семьдесят пять нуклеотидов. Красной линией отмечены районы, выделяемые соответствующими окнами.

Видно, что район пониженной сложности от минус ста до минус пятидесяти пар оснований относительно старта транскрипции выделяется только для бОльших размеров окна.

Рис.31.
На этом слайде представлены соответствующая последовательность промотора и найденный участок пониженной сложности.
[–300;+100] промоторный район гена MMIFNBG.
TATA-бокс выделен красным.
Район низкой сложности выделенный с использованием скользящего окна показан синим цветом (см. предыдущий рисунок). Меньший размер скользящего окна не смог выделить этот район низкой сложности.

Рис.32.
Рассмотрим задачу поиска районов пониженной сложности для полностью секвенированных геномов.
На рисунке представлен профиль сложности для генома Borrelia burgdorferi скользящим окном 1000 нт. Выявленный район низкой сложности отмечен красной стрелкой.
Отмеченный сегмент соответствует гену BB0210, кодирующему поверхностный трансмембранный белок 1 (surface-located membrane protein 1, lmp1). Он содержит два прямых повтора, 234 нт и 315 нт длиной, расположенные в кодирующем районе.

Рис.33.
На данном слайде представлены профили сложности генома Bacillus subtilis.
Профили построены для разных размеров скользящего окна - в тысячу, пять тысяч и десять тысяч нуклеотидов, соответственно.
Вы можете видеть, что районы пониженной сложности выявляются только для бОльшего размера окна.
Стрелкой указаны районы пониженной сложности.
Район 90,000-100,000 п.о. Содержит кластер генов рибосомальных РНК (16S,23S, и 5S) и транспортных РНК для Val, Thr, Lys, Leu, Gly, Arg, Pro, и Ala.
Район 160,000-175,000 содержит другой кластер рибосомальных и транспортных РНК
Райн 3,665,000-3,675,000 соответствует белку связывающемуся с мембраной, относящемуся к синтезу галактозамин содержащих кислот.

Рис.34.
На данном слайде представлен
Профиль сложности генома Bacillus subtilis совмещенный с длинами компонент разложения.
Профили в скользящем окне различных размеров
Коричневая линия (пики) соответствует длинам компонент в полном сложностном разложении.
Стрелками отмечены длинные фрагменты не выявленные техникой скользящего окна (отсутствие пиков вниз на всех профилях).
Таким образом, только полное сложностное разложение выявляет все повторы.

Рис.35.
Рассмотрим статистику полного сложностного разложения того же генома Bacillus subtilis (EMBL:AL009126|BSUB)
В данной таблице представлено количество повторов в зависимости от типа повторов в полном сложностном разложении этого генома.
Можно видеть, что для больших длин повторов в сложностном разложении встречаются только прямые (стрелка) и инвертированные (стрелка) повторы.

Общее распределение числа повторов в разложенн представлено на этой диаграмме.
Видна большая встречаемость прямых и инвертированных повторов.

Доля каждого типа повторов в разложении указана в круговой диаграмме справа.

Рис.36.
На данном рисунке представлено продолжение таблицы (для компонент длины больше чем 20 nt)
Видно, что для больших длин вообще не встречаются симметричные и комплементарные повторы.
На диаграмме представлены пропорции повторов каждого типа в разложении (>20 нт)

Видно преобладание числа прямых повторов

На этой диаграмме представлена суммарная длина длинных повторов в сложностном разложении (>20 нт). Прямые чаше встречаются и имеют бОльшую длину

Рис.37.
Распределение длин компонент в разложении генома B. subtilis
представлено графически

Видно преобладание прямых и инвертированных повторов, отмеченных темно-синим и светло-голубым цветом.
Рассмотрим хвост распределения более детально.
Прямые и инвертированные повторы могут быть очень протяженными,
максимальный повтор составляет 2952 nt

На диаграмме приведена суммарная длина всех повторов длины больше 20 нуклеотидов

Общая длина таких повторов в сложностном разложении составляет около 5 процентов для генома Bacillus subtilis (круговая диаграмма).

Рис.38.
Здесь показано чему соответствует максимальный компонент сложностного разложения данного бактериального генома
Это Gene 23S-RNA
Фрагмент 2952 п.о.
Расположение (позиции относительно Ori точки начала репликации)
Копия: 32150-35102
Другая копия: 92226-95178.
Этот повтор не выявляется с помощью техники скользящего окна.

Рис.39.
На данном слайде представлены результаты сложностных разложений для нескольких бактериальных геномов и фрагментов хромосом эукариот.
Для каждого генома выполнено полное сложностное разложение с использованием повторов всех типов.
Средняя длина фрагмента в сложностном разложении может служить в качестве меры относительной сложности геномов.

На графике показана корреляция между длиной последовательности и средней длиной компонента разложения для геномов, приведенных в таблице.

Теоретически такая зависимость для случайных последовательностей имеет логарифмический характер.

Рис.40.
Посмотрим как характеризуются средние длины повторов различных типов в сложностном разложении
Для тех же геномных последовательностей показана средняя длина компонент для прямых, симметричных, комплементарных и инвертированных повторов, соответственно.

Обозначено буквами D, S, C, I
Отдельно показана длина максимального повтора и его тип.
Видно, что все точные (exact) повторы максимальной длины всегда прямые или инвертированные.

Для всех последовательностей видно преобладание прямых и инвертированных повторов над симметричными и комплементарными во всех случаях, без исключений.
В то же время для случайной, сгенерированной последовательности такого преобладания не выявляется.

Рис.41.
Более точно.
Результаты результаты компьютерной симуляции для случайной последовательности длиной 1Мб (один мегабайт или одна мегабаза пар оснований) с равными частотами нуклеотидов заключаются в том, что:
нет никаких повторов длины более 21 нуклеотида,
нет разницы между повторами различных типов и преобладания прямых повторов.

Те же результаты получаются при computer simulation последовательности в точности той же длины, что и реальный геном, например Borrelia burgdorferi,
и с теми же частотами нуклеотиов.
Хотя средняя длина повтора при этом несколько возрастает, качественно результаты остаются теми же.
- нет длинных повторов длины более 21 нуклеотида,
- нет разницы между повторами различных типов.

Заключение по анализу распространенности повторов:
Прямые и инвертированные повторы происходят в эволюции в результате дупликаций и дупликаций с последующими инверсиями.
Для симметричных и прямых комплементарных повторов аналогичных молекулярно-генетических механизмов происхождения не существует.

Рис.42.
В продолжение анализа информационного содержания и сложности рассмотрим методику моделирования последовательностей с помощью марковских моделей.

Порождающие деревья-источники или марковские модели с переменной памятью для анализа генетических текстов

Рассмотрим разложение нуклеотидной последовательности на перекрывающиеся фрагменты

Рис.43.

Здесь Вы можете видеть пример разложения на перекрывающиеся фрагменты небольшой длины.
Пошаговая генерация последовательности ДНК на основе коротких предшествующих контекстов:
Можно рассматривать ее как декомпозицию с перекрывающимимися фрагментами (в данном случае динуклеотидами).
Если мы рассматриваем зависимость только от короткого предшествующего контекста, можно рассматривать нуклеотидную последовательность как марковскую модель.
Для поиска функциональных сайтов и предсказания структуры генов широко используются скрытые марковские модели (HMM, Hidden Markov models).

Рис.44.
Рассмотрим расширение набора предшествующих контекстов.
Мы можем использовать тринуклеотиды, в некоторых случаях - только нуклеотиды (контекст длины 1).
Такой набор контекстов удобно представлять в форме суффиксного дерева - см.рисунок.
Модель использует суффиксное дерево для генерации последовательности.
Программа доступна по адресу: http://wwwmgs.bionet.nsc.ru/programs/complexity/.

Используя данную программу можно оценивать близость последовательности к исходной модели, определяя ее вероятность.

Рис.45.
Приведено графическое представление порождающего дерева-источника и компьютерная выдача для реальных данных.
Заметим, что такое дерево является удобным и наглядным, однозначным визуальным представлением всего набора коротких предшествующих контекстов.
Дерево построено с использованием программы Complexity для нуклеотидных последовательностей ССТФ АР-1. Каждая связь от листьев к корню соответствует контексту в последовательности ДНК и имеет свой набор вероятностей сгенерировать следующий символ (см.таблицу слева).
Например в дереве, показанном на рисунке, 12 контекстов длины 3 нуклеотида (NCC, NAC, и NGC); 9 контекстов длины 2 нт (AA, TA, GA, AG, TG, GG, CG, AC, TC, GC, и CC); и 1 контекст длины 1 нуклеотид (T). Всего 22 предшествующих контекста. Каждый контекст определяет четыре числа – вероятности появления символа непосредственно справа от этого контекста. Чтобы определить эти значения, мы используем частоты соответствующих олигонуклеотидов на единицу большей длины: всего 4 22 = 88 значений.

Рис.46.
Связь между марковскими моделями и порождающими деревьями-источниками.
Маркоские модели являются частным случаем порождающих деревьев.

Примеры полных суффиксных деревьев, соответствующих полным наборам
4 нуклеотидов,
16 динуклеотидов,
64 тринуклеотидов.

Заметим, что порождающее дерево-источник может тем не менее быть дополнено и включено в марковскую модель большего порядка.

Рис.47.

Приведены представления для марковской модели (полное дерево) и контекстного дерева-источника.
Марковский источник 3-го порядка.
Контекстное дерево-источник, полученное из полного дерева.

Вероятность порождения последовательности Xn=X1X2…Xn определяется вероятностями появления символов Xi, составляющих Xn в соответствующих контекстах Si.
P(Xn) = P(X1|S1) P(X2|S2)…P(Xn|Sn),
где Si принадлежит T' – контекст, в котором буква Xi содержится в слове Xn, T' - порождающее дерево.

Задача – построить оптимальный источник для генерации текстов и исследовать его свойства для разных классов генетических текстов

Рис.48.
Вопрос оптимального выбора порождающего дерева-источника.
Алгоритм выбора модели – дерева-источника.
Стохастической сложностью L^0 данных относительно модели M называется минимальная сумма L сложностей сообщения X^n, при некотором заданном параметрами модели распределении teta, и сложности описания этих параметров H_n (сложности модели M).

$$L^0(X^n,M) = L(X^n,teta(X^n),M)+H_n(M)$$

Если оценить вероятности через частоты, сложность с точностью до нормировки на длину последовательности n будет равна энтропии Шеннона - $\Sigma\, p \log p$, сумме произведений вероятностей символов на логарифм этих вероятностей. Оценка сложности $H_n(M_0)$ описания параметров модели для четырехбуквенного алфавита равна

$$H_n(M_0) = (3)/2 \log n + 1/2 \log (pi) - 3/2$$

Стандартный термин древовидной модели по отношению к Марковским цепям - Variable Memory Markov Model

Рис.49.
Пример порождающего дерева-источника. Дерево построено для последовательности ДНК кластера бета глобинов человека, хромосома 11, 73308 п.о. (EMBL ID: HSHBB). Оба рисунка, сверху и слева, представляют один и тот же граф, соответственно, в стандартной форме и в радиальной форме (по окружности). Рисунки автоматически генерируются программой оценки сложности и построения контекстных деревьев, доступной по адресу http://wwwmgs.bionet.nsc.ru/mgs/programs/complexity/

Рис.50.
Дерево может быть построено по обучающей последовательности или выборке последовательностей.
Деревья могут быть достаточно громоздкими и не вмещаться на экран компьютера.
Пример
Mycoplasma pneumoniae полный геном. Контекстное дерево разделено на две части.

Рис.51.
Анализ аминокислотных последовательностей с помощью деревьев-источников
Чтобы получить двухбуквенный алфавит 20 аминокислотных остатков были разбиты на две группы - поверхностные, гидрофильные внешние (Outer) O={ R N D C Q E G H K S T Y }и гидрофобные, внутренние (Inner) I={ A I L M F P W V }. Разбиение на три группы по степени представленности на поверхности белка - внешние, O (outer) {R N D Q E H K}, амбивалентные , A(ambiv.) { A C G P S T W Y } и внутренние, I(inner) { I L M F V }.

Полученные контекстные деревья для различных типов доменов проинтерпретированы в терминах вторичной структуры глобулярных белков.
Результаты анализа выборок аминокислотных последовательностей альфа-спиральных и бета-структурных белков из базы данных SCOP в двухбуквенном алфавите и трехбуквенном алфавитах.

Рис.52.
Компьютерная реализация изложенного выше алгоритма.
Интернет-интерфейс программы Complexity для анализа сложности генетических текстов с помощью моделей деревьев-источников

Рис.53.
Структура дерева может быть достаточно сложной для визуализации.
Последовательность может анализироваться не одна, а несколько одновременно.
Контекстное дерево для последовательностей содержащих TATA бокс (534 последовательности длины 20 п.о. из TRRD).
Символы предшествующих контекстов длины свыше 3 п.о. Не показаны из ограниченности места на схеме; знак (*) отмечает контексты больше 5 п.о. Ниже показан увеличенный фрагмент того же дерева, ограниченный синей линией. Длинные контексты, маркированные звездочкой, показаны ниже полностью: ATATAA, TTATAA, GTATAA, и CTATAA.

Рис.54.
Предсказание TATA box-бокса с помощью марковской модели с переменной памятью.
Поскольку у нас есть модель (дерево-источник) мы можем исследовать профиль - вероятность наблюдения символа вдоль последовательности.
Такой профиль лучше вычислять в логарифмической шкале, что и представлено на рисунке.

Профиль функции предсказания TATA-бокса в промоторном районе [300;+100] гена металлотинотеин-I (metallothionein-I).
Наибольшие значения профилей показывают районы, наиболее НЕ-характерные для последовательности в целом, наименьшие - наиболее характерные.
Красная линия показывает профиль предсказания с помощью модели дерева-источника; синяя и коричневая кривые соответствуют предсказанию с помощью только частот нуклеотидов и динуклеотидов в том же скользящем окне (марковские модели нулевого и первого порядка). Отмечено положение TATA бокса аннотированное в базе данных TRRD и старт транскрипции.
Пик вниз показывает правильное предсказание.

Рис.55.
Предсказание TATA box-бокса с помощью марковской модели с переменной памятью
Профиль построен для скользящего окна 15 нт
124 последовательности, содержащие TATA бокс
сфазированы относительно старта транскрипции.
Усредненный профиль распознавания показан жирной красной линией.
Пик вниз соответствует характерному положению TATA-бокса в районе -30 относительно страта транскрипции.

Рис.56.
На данном рисунке приведен профиль локальной близости последовательности в скользящем окне размера 1000 нт для разных статистических моделей.
Обучение, подсчет частот и построение модели шло по всей последовательности.

Коричневый, профиль построен только по частотам нуклеотидов (оцененных по всему геному); голубой, профиль по частотам динуклеотидов; и синий, профиль построен по марковской модели с переменной памятью (порождающее дерево источник построено по всей геномной последовательности)

Наибольшие значения профилей показывают районы, наиболее НЕ-характерные для последовательности в целом, наименьшие - наиболее характерные.
Сейчас нас интересуют наиболее нехарактерные.

Это районы генов рибосомальных белков (14 kbp), рибосомальных и транспортных РНК (98–100 kbp, 165–170 kbp, 635 kbp, 946–950 kbp, и 3,172 kbp, соответственно)

Таким образом гены РНК имеют контекстную структуру, отличную от основного состава (белок-кодирующих генов).

Рис.57.
Программы и материалы доступны в Интернете на сайте ИЦиГ СО РАН. Древовидные деревья-источники:
http://wwwmgs.bionet.nsc.ru/mgs/programs/complexity/,
Сложностные разложения по Лемпелю и Зиву:
http://wwwmgs.bionet.nsc.ru/mgs/programs/lzcomposer/,
http://wwwmgs.bionet.nsc.ru/mgs/programs/gc_net/
Сайт Интеграционного проекта СО РАН по моделированию фундаментальных генетичечских процессов и систем содержит ряд последних работ по анализу текстов на русском языке, полезные Интернет-ссылки и литературу.
http://www.bionet.nsc.ru/ICIG/report/2001/icg_im/.
Литература:
Франк-Каменецкого М.Д. под ред. (1990) Компьютерный анализ генетических текстов // Москва, Наука, 1990, 267 с.