Компьютерный анализ информационного содержания геномов

Юрий Львович Орлов

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

Структура лекции

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

Задачи исследования информационных процессов в молекулярной генетике:

Обзор подходов

Поиск и предсказание генов, функциональная аннотация

- ✓Генетическое рестрикционное картирование
- ✓Физическое картирование
- ✓Секвенирование
- ✓Поиск сходства
- ✓Предсказание генов
- ✓Анализ мутаций
- ✓Множественное выравнивание
- ✓Поиск сигналов в ДНК

Литература:

Pevzner, Pavel. Computational Molecular Biology: an algorithmic approach. MIT Press, 2000, 311 pages.

Основные задачи компьютерного анализа генетических текстов:

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;

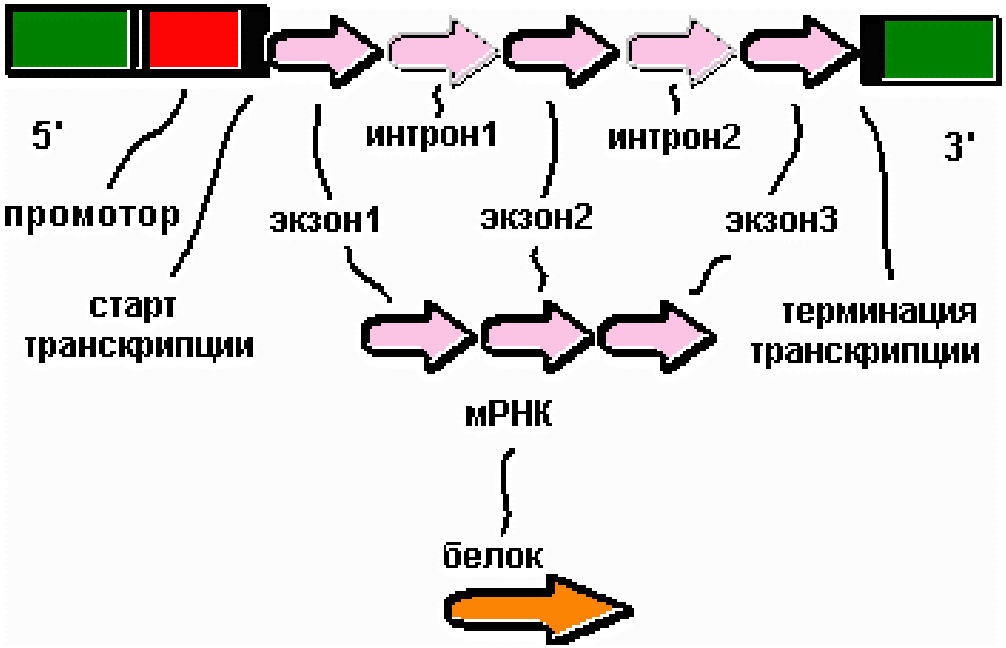
4) анализ вторичной структуры РНК и сигналов трансляции;

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.

Схема структуры гена



Пример – линейное расположение функциональных участков гена

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: сайт рестрикции *Hind*II **CTGCAC, GTTAAC**

Участки, которые узнают на ДНК белки-регуляторы работы генов, не столь однозначны. 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.

Понятие генетического текста

Генетическим текстом называется представление реальной макромолекулы (ДНК, РНК и белков) в виде конечной последовательности символов соответствующего алфавита.

Такое представление отражает наиболее существенные особенности генетических макромолекул, а именно – наличие в них генетической информации и линейного способа ее записи и хранения.

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**.

Свойства молекул ДНК как генетических текстов

Рассматривая последовательность ДНК как генетический текст, предполагается, что это последовательность ДНК от 5' к 3'-концу, соответствующая ей комплементарная последовательность не записывается.

Особенности организации геномов и генов эукариот и прокариот были изложены в предыдущих лекциях.

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.

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).

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 относиться как к нуклеотидным сайтам, так и к активным сайтам белков.

Представление в форме консенсуса.

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 NNNCNGSSSSCCCTTTWWWWAAAGSSSSSSSSCNNG.

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.

actgtgtatataagagctct
ttctcgtataaatactctac
acttcatataagtatcctac
aggctatttatttattcagc
accggcctataaattgcccg

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-ДельгарноВ ДНК прокариот к числу таких регуляторных сайтов относятся, например, операторы, которые расположены, рядом с промоторами или перекрываются с ними. [Зенгбуш П., 1982; Стент Г., Кэлиндар Р., 1981]. Для эукариот структура регуляторных сайтов описывается более сложным образом.

Сложность внутренней структуры реальных генетических текстов

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.

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.

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.

Повторы в нуклеотидных последовательностях подразделяются на прямые, инвертированные, симметрические, а также палиндромы и комплементарные палиндромы.

Прямой

ATGC...ATGC
TACG...TACG

Симметричный:

ATGC...CGTA
TACG...GCAT

Инвертированный:

ATGC...GCAT
TACG...CGTA

Прямой комплементарный:

ATGC...TACG
TACG...ATGC

TAAT
ATTA

Палиндром

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.

>d3sdha_ 1.1.1.1.1 Hemoglobin I [ark clam (Scapharca inaequivalvis)]

SVYDAAAQLTADVKKDLRDSWKVIGSDKKGNGVALMTTLFADNQETIGYF
KRLGNVSQGMANDKLRGHSITLMYALQNFIDQLDNPDDLVCVVEKFAVNH
ITRKISAAEFGKINGPIKKVLASKNFGDKYANAWAKLVAVVQAAL

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:

7. **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].

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].

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)

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

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].

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).

Не существует рациональных алгоритмов для определения такой сложности – есть лишь различные реализации в рамках заданных ограничений.

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

Наибольшее распространение получила мера сложности, введенная Лемпелем и Зивом (мера LZ). 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.

$$S = S_1 S_2 \ldots S_m$$

$$H(S) = S[1:i_1]\, S[i_1+1:i_2] \ldots S[i_{k-1}+1:i_k] \ldots S[i_{m-1}+1:N]$$

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.

GTAGTCTGATGCA

Сложность текста и сложностные разложения

Терминология:

Декомпозиция = разложение = факторизация

Факторизация – представление последовательности ДНК S в форме набора неперекрывающихся фрагментов

$$S = S_1 S_2 \ldots S_m$$

GTAGTCTGATGCA

Сложностные разложения связаны с теорией сжатия данных

Схема Лемпеля и Зива: Каждый фрагмент является копией другого участка последовательности

GTAGTCTGATGCA

$$H(S) = S[1:i_1]\,S[i_1+1:i_2]\ldots S[i_{k-1}+1:i_k]\ldots S[i_{m-1}+1:N]$$

Число m минимально, когда длины S_i максимальны

Расширение схемы Лемпеля и Зива для последовательностей ДНК:

Каждый фрагмент является копией или комплементарной копией 5’-района

Возможные операции: прямое и обратное копирование и прямое и обратное комплементарное копирование

Классификация повторов :

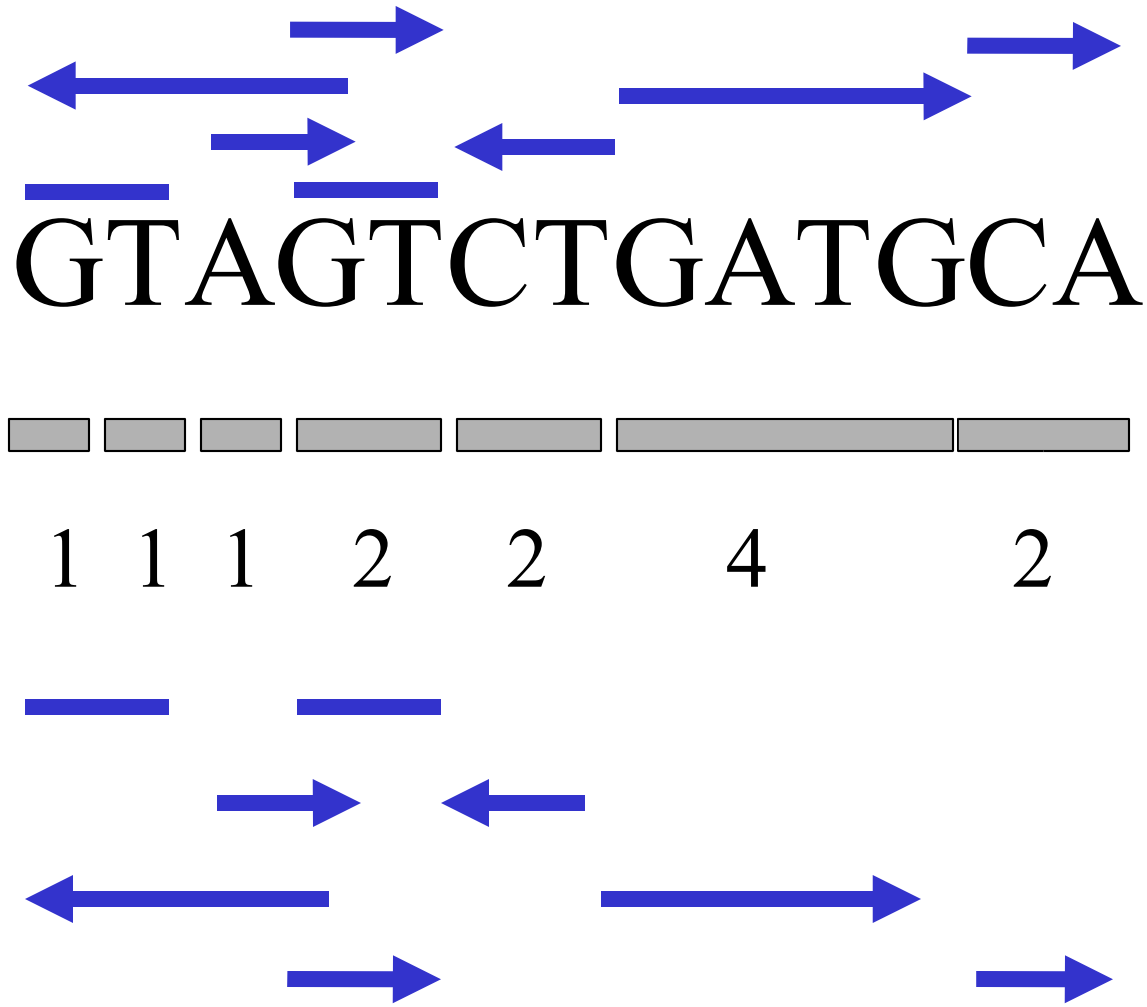
Прямой
ATGC...ATGC
TACG...TACG

Симметричный:
ATGC...CGTA
TACG...GCAT

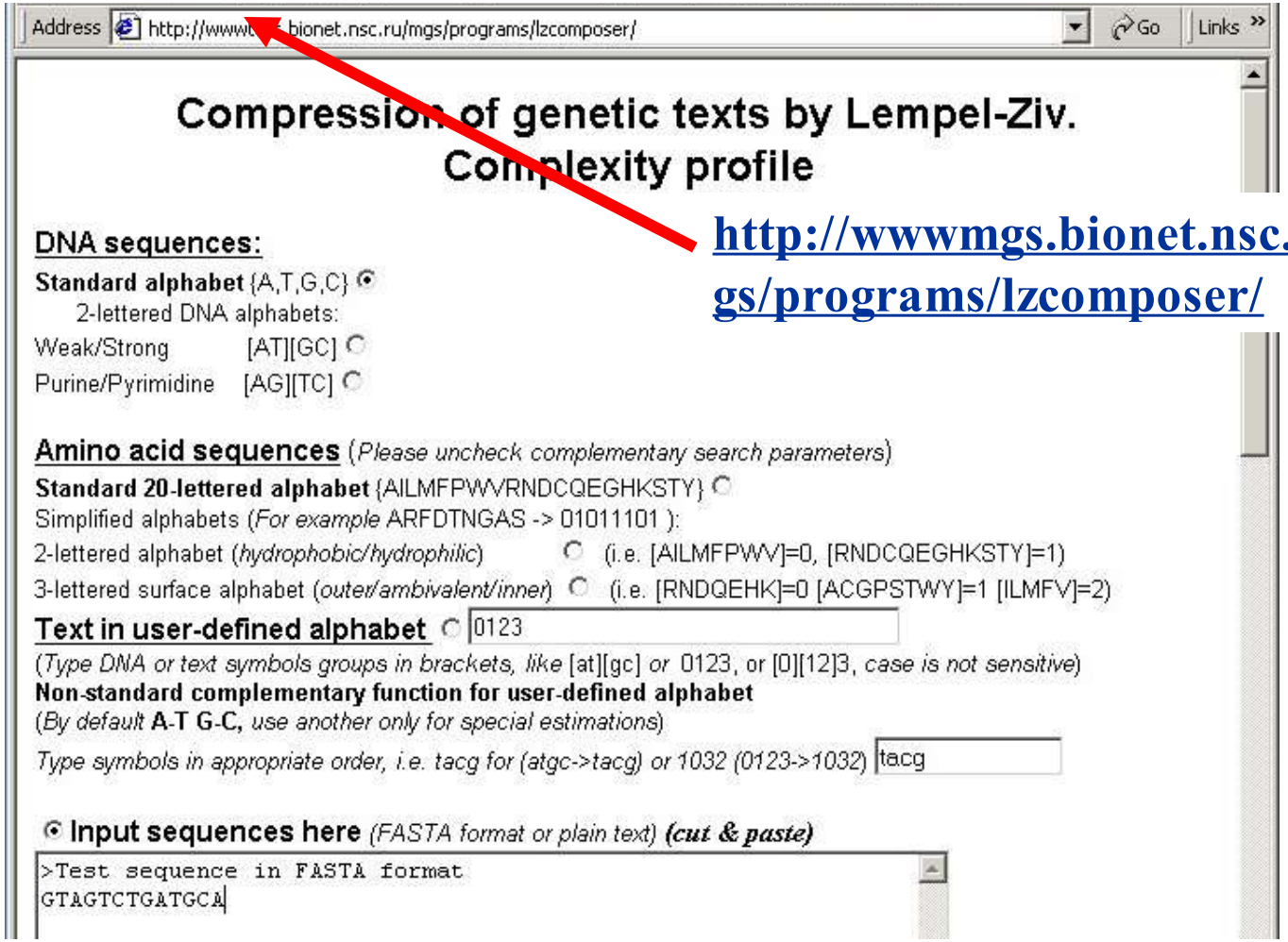
Инвертированный:
ATGC...GCAT
TACG...CGTA

Прямой комплементарный:
ATGC...TACG
TACG...ATGC

Пример сложностного разложения последовательности:



Интернет-интерфейс программы сложностных разложений



http://wwwmgs.bionet.nsc.ru/mgs/programs/lzcomposer/

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

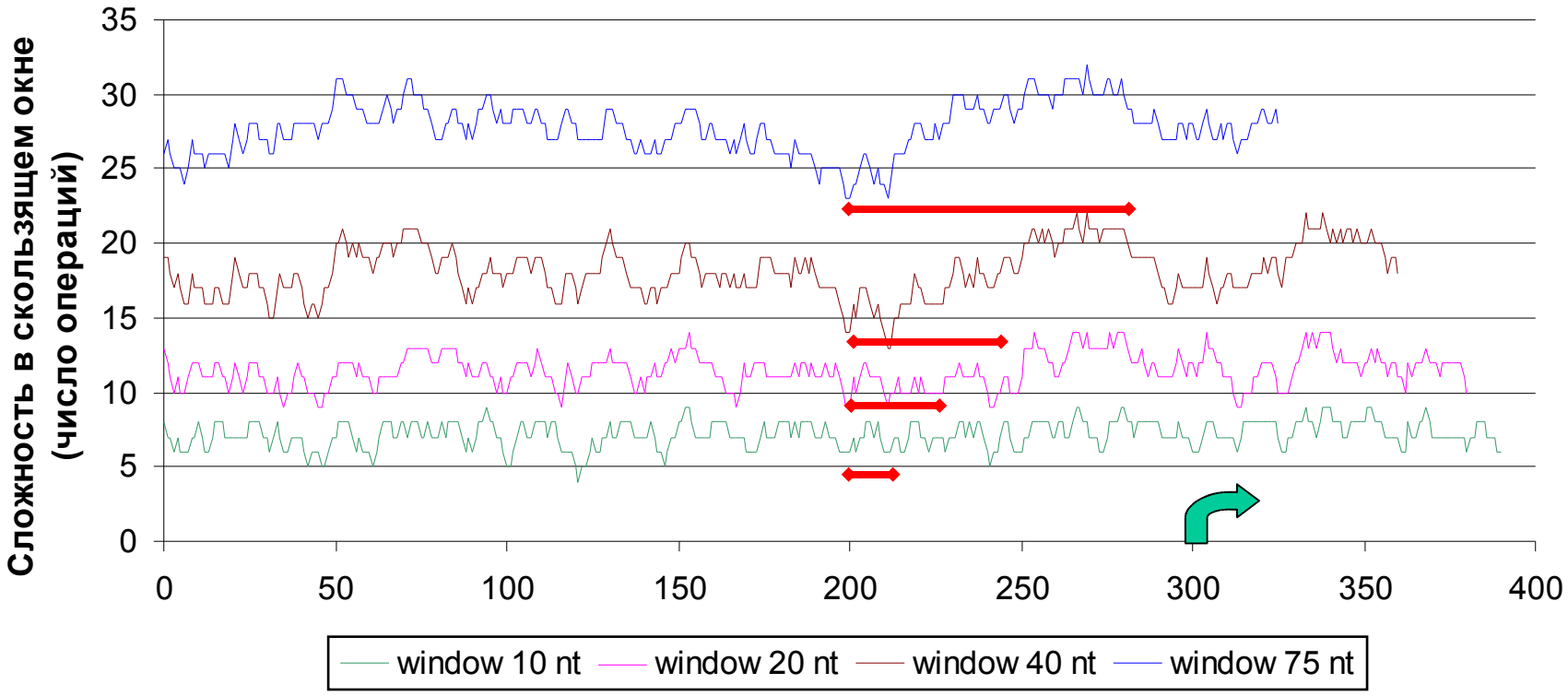
Цели использования сложностных разложений

- 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

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)

Профиль сложности в скользящем окне позволяет найти районы низкой сложности в последовательности



Промоторная последователь [-300;+100] гена интерферон-бета мыши EMBL ID:MMIFNBG

```
>A00039_EMBL; MMIFNBG; X14029; ST:1193_Region 893-1292
ctgtttgagagttcttttatcttcagggctgtctcctttctgttcttctc
tcctggatatttctcttcctttgctccagcaattggtgaaactgtacaag
attttataaatccttagtttgtatatattttaacccagtacatagcatat
aaaatagccaggagcttgaataaaatgaatattagaagctgttagaataa
gagaaaatgacagaggaaaactgaaagggagaactgaaagtgggaaattc
ctctgaggcagaaaggaccatccctTATAAATagcacaggccatgaagga
agatcattctcactgcagcctttgacagcctttgcctcatcttgcaggta
gcagccgacaccagcctggcttccatcatgaacaacaggtggatcctcca
```

Старт транскрипции

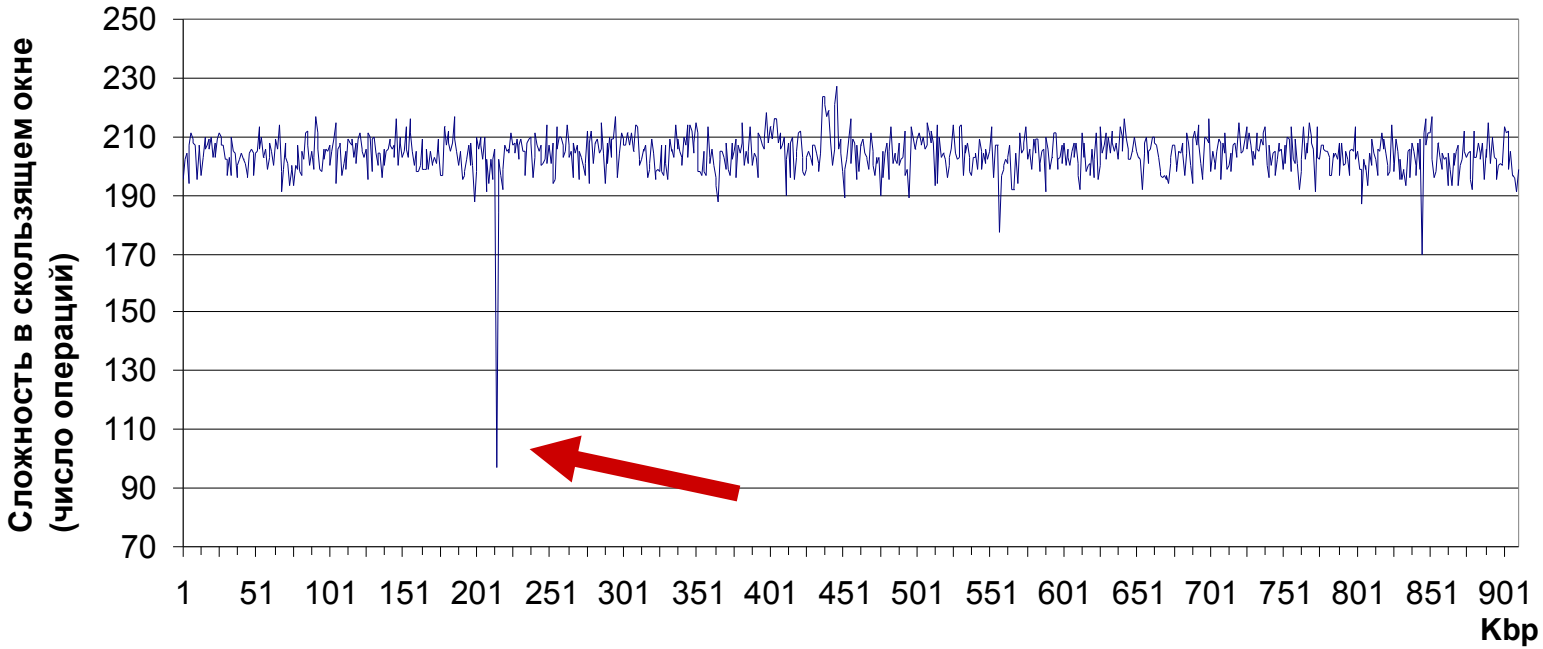
[–300;+100] промоторный район гена MMIFNBG.

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.

Профили сложности для протяженных последовательностей

Профиль сложности для генома *Borrelia burgdorferi*
в скользящем окне 1000 нт

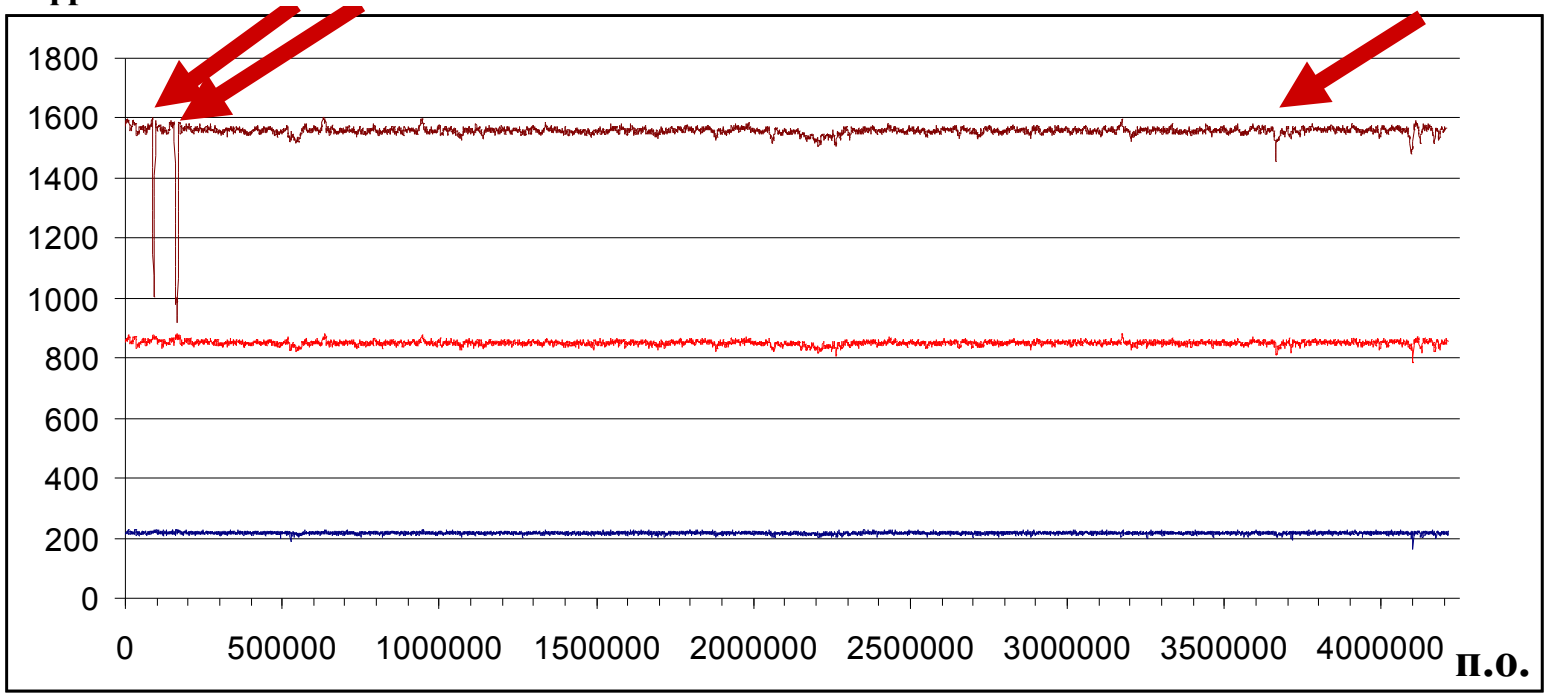


Отмеченный сегмент соответствует гену BB0210, кодирующему поверхностный трансмембранный белок 1 (surface-located membrane protein 1, lmp1). Он содержит два прямых повтора, 234 нт и 315 нт длиной, расположенные в кодирующем районе.

Профиль сложности для генома *Bacillus subtilis*

(EMBL:AL009126|BSUB) Скользящее окно 1000, 5000 и 10000 нт

Число фрагментов в окне

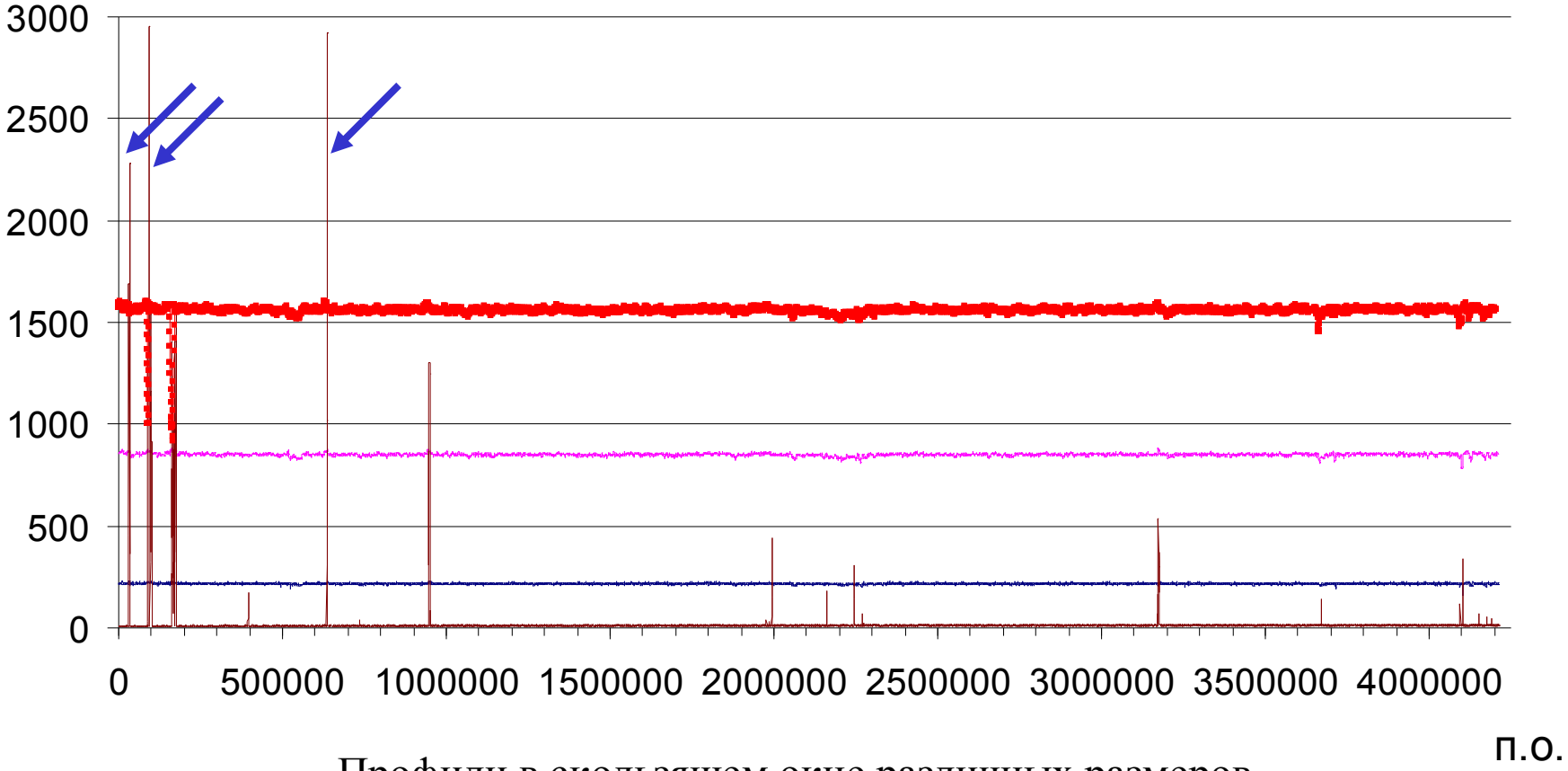


Район 90,000-100,000 п.о. Содержит кластер генов рибосомальных РНК (16S,23S, и 5S) и транспортных РНК для Val, Thr, Lys, Leu, Gly, Arg, Pro, и Ala.

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.

Профиль сложности генома *Bacillus subtilis* совмещенный с длинами компонент разложения



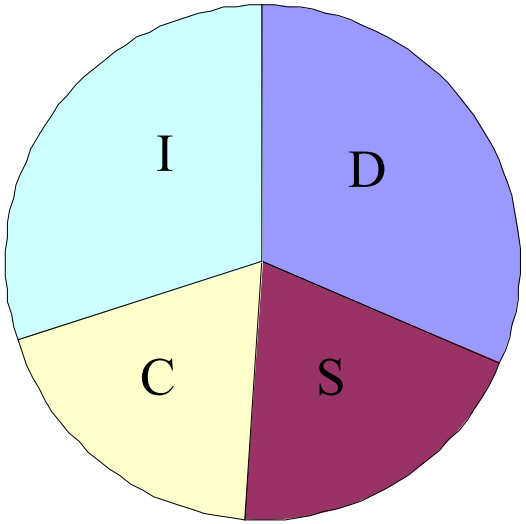
Профили в скользящем окне различных размеров

Коричневая линия (пики) соответствует длинам компонент в полном сложностном разложении.

Статистика частот повторов в сложностных разложениях генома *Bacillus subtilis*

Длина повтора	Общее число	Тип повтора: Dir.	Sym.	Compl.	Inverted
2:	8	4	1	2	1
3:	11	2	2	4	3
4:	49	11	10	13	15
5:	146	32	48	27	39
6:	463	126	120	94	123
7:	1546	421	377	367	381
8:	5546	1518	1389	1189	1450
9:	19498	5654	4457	4114	5273
10:	61474	18387	13283	12563	17241
11:	110536	34756	21890	21032	32858
12:	94291	30445	17642	16864	29340
13:	46299	15076	8367	8055	14801
14:	16875	5580	2823	2814	5658
15:	5412	1851	889	870	1802
16:	1628	592	217	240	579
17:	506	183	59	69	195
18:	168	68	15	10	75
19:	72	37	6	2	27
20:	27	17	0	1	9
21:	18	7	0	0	11
22:	23	19	0	0	4
23:	8	6	0	0	2
24:	10	7	0	0	3
25:	14	12	0	0	2
…	…	…	…	…	…

Доля каждого типа повторов в разложении :

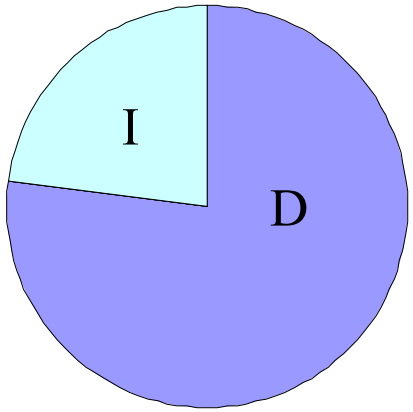


Статистика сложностных разложений для генома *Bacillus subtilis*

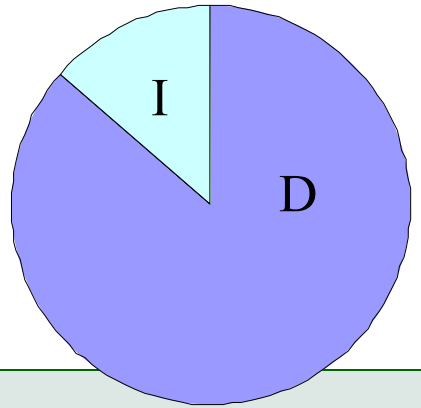
Распределение хин компонент, больших 20 нт

Длина повтора	Общее число	Тип повтора: D.	S.	C.	I.
19:	72	37	6	2	27
20:	27	17	0	1	9
21:	18	7	0	0	11
22:	23	19	0	0	4
23:	8	6	0	0	2
24:	10	7	0	0	3
25:	14	12	0	0	2
26:	8	4	0	0	4
27:	9	7	0	0	2
28:	5	5	0	0	0
29:	4	3	0	0	1
30:	9	6	0	0	3
31:	6	5	0	0	1
32:	4	4	0	0	0
33:	8	6	0	0	2
34:	5	4	0	0	1
35:	5	3	0	0	2
...					
1691:	1	1	0	0	0
2281:	1	1	0	0	0
2918:	1	1	0	0	0
2952:	1	1	0	0	0

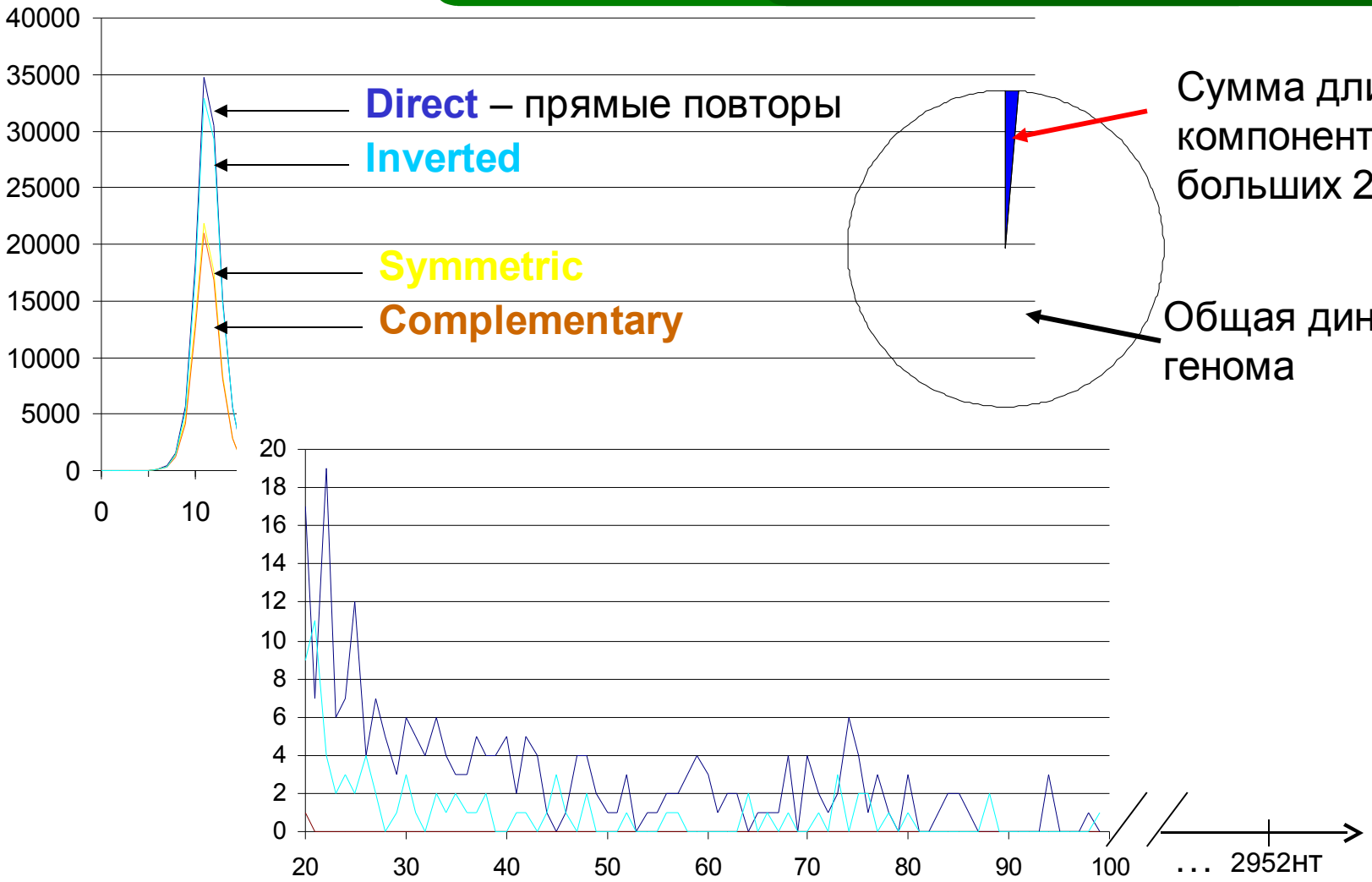
Доля каждого типа повторов (>20 nt)



Доля суммарных длин компонент (>20 nt)



Распределение длин компонент в разложении генома *B. subtilis*



```
FT   tRNA            32019..32093
FT                   /gene="trnA-Ala"
FT                   /product="transfer RNA-Ala"
FT   rRNA            32175..35102
FT                   /gene="rrnA-23S"
FT                   /product="ribosomal RNA-23S"
FT   rRNA            35235..35346
FT                   /gene="rrnA-5S"
FT                   /product="ribosomal RNA-5S"
FT   terminator      35360..35378
FT                   /gene="rrnA-5S"
```

Максимальный повтор в геноме *B. subtilis* :

Ген 23S-RNA,

Фрагмент 2952 по

Копия: 32150-35102

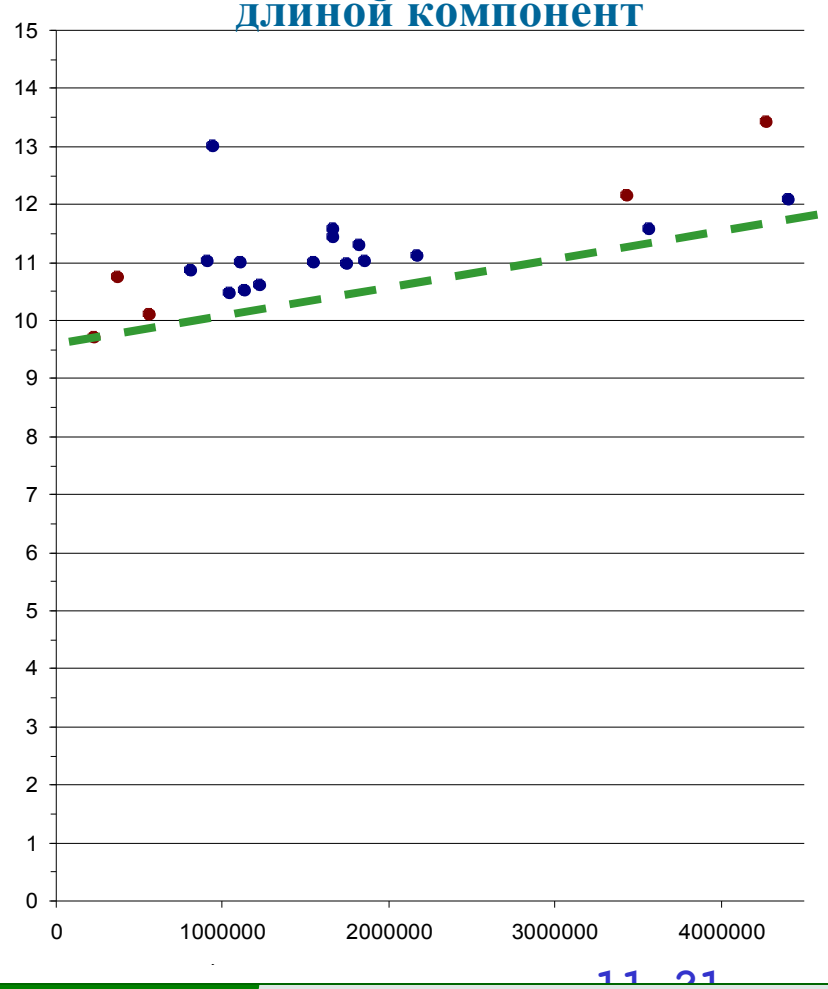
Другая копия: 92226-95178

Этот повтор не выявляется с помощью техники скользящего окна

```
FT   rRNA            90533..92085
FT                   /gene="rrnJ-16S"
FT                   /product="ribosomal RNA-16S"
FT   rRNA            92251..95178
FT                   /gene="rrnJ-23S"
FT                   /product="ribosomal RNA-23S"
FT   rRNA            95234..95349
FT                   /gene="rrnJ-5S"
FT                   /product="ribosomal RNA-5S"
```

Организм	Размер генома (по)	Средняя длина компонент (нт)
Mycoplasma pneumoniae	816394	**10.86**
Borrelia burgdorferi	910724	**11.02**
Plasmodium falciparum chr.2	947089	**12.99**
Chlamydia trachomatis	1042519	:
Rickettsia prowazekii		
Treponema pallidum	1111523	:
Chlamydia pneumoniae		
Aquifex aeolicus	1137944	:
Methanococcus jannaschii		
Helicobacter pylori 26695	1230230	:
M. thermoautotrophicum		
Haemophilus influenzae Rd	1551335	:
Thermotoga maritima		
Archaeoglobus fulgidus	1664957	:
Synechocystis PCC6803		
Mycobacterium tuberculosis	1667825	:
Среднее		
Mouse, chromosome 1	1751377	:
Mouse, chromosome 2		
Mouse, chromosome Y	1830023	:
***S. cerevisiae* Chromosome I**		
***S. cerevisiae* Chrom. VIII**	1860725	:
Случайная последовательность	2178400	:

Корреляция между длиной последовательности и средней длиной компонент



Организм	Максимальная длина компоненты (нт) и ее тип	Средняя длина компонент для повторов типов: D	S	C	I
Mycoplasma pneumoniae	459,D	12.01	10.23	10.23	10.56
Borrelia burgdorferi	1430,D	11.13	10.90	10.95	11.05
Plasmodium falciparum chr.2	475,I	13.83	12.23	12.22	13.33
Chlamydia trachomatis	4904,D	10.64	10.33	10.37	10.44
Rickettsia prowazekii	487,D	11.04	10.93	10.95	11.01
Treponema pallidum	3278,D	10.84	10.28	10.31	10.47
Chlamydia pneumoniae	1371,D	10.77	10.47	10.52	10.59
Aquifex aeolicus	5270,I	11.10	10.80	10.84	11.16
Methanococcus jannaschii	1018,I	11.66	11.33	11.34	11.73
Helicobacter pylori 26695	4847,I	11.67	10.93	10.93	11.70
M. thermoautotrophicum	1856,D	11.28	10.61	10.64	10.99
Haemophilus influenzae Rd	5791,I	11.50	10.89	10.91	11.61
Thermotoga maritima	917,D	11.17	10.80	10.79	11.12
Archaeoglobus fulgidus	1209,D	11.37	10.83	10.85	11.21
Synechocystis PCC6803	5353,I	11.82	11.18	11.20	11.82
Mycobacterium tuberculosis	1697,D	12.41	11.72	11.72	12.20
Среднее по бактер. геномам :		11.51	10.90	10.92	11.31
Mouse, chromosome 1	10187,I	14.32	11.51	11.48	14.63
Mouse, chromosome 2	43688,I	13.11	11.40	11.36	12.04
Mouse, chromosome Y	10095,I	11.02	9.85	9.71	11.70
***S. cerevisiae* Chromosome I**	810,I	9.75	9.15	9.19	10.62
***S. cerevisiae* Chrom. VIII**	1988,D	10.29	9.90	9.92	10.26
Случайная последовательность	21,D,S,C,I	9.09	9.09	9.08	9.10

Компьютерная симуляция повторов в случайной последовательности

Компьютерная симуляция для случайной последовательности 1 Мб с равными частотами нуклеотидов:

- **Максимальный размер повтора - 21 нт для всех типов повторов.**
- **Средние длины не различаются:**

  9.09, 9.09, 9.08, *и* 9.10

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.

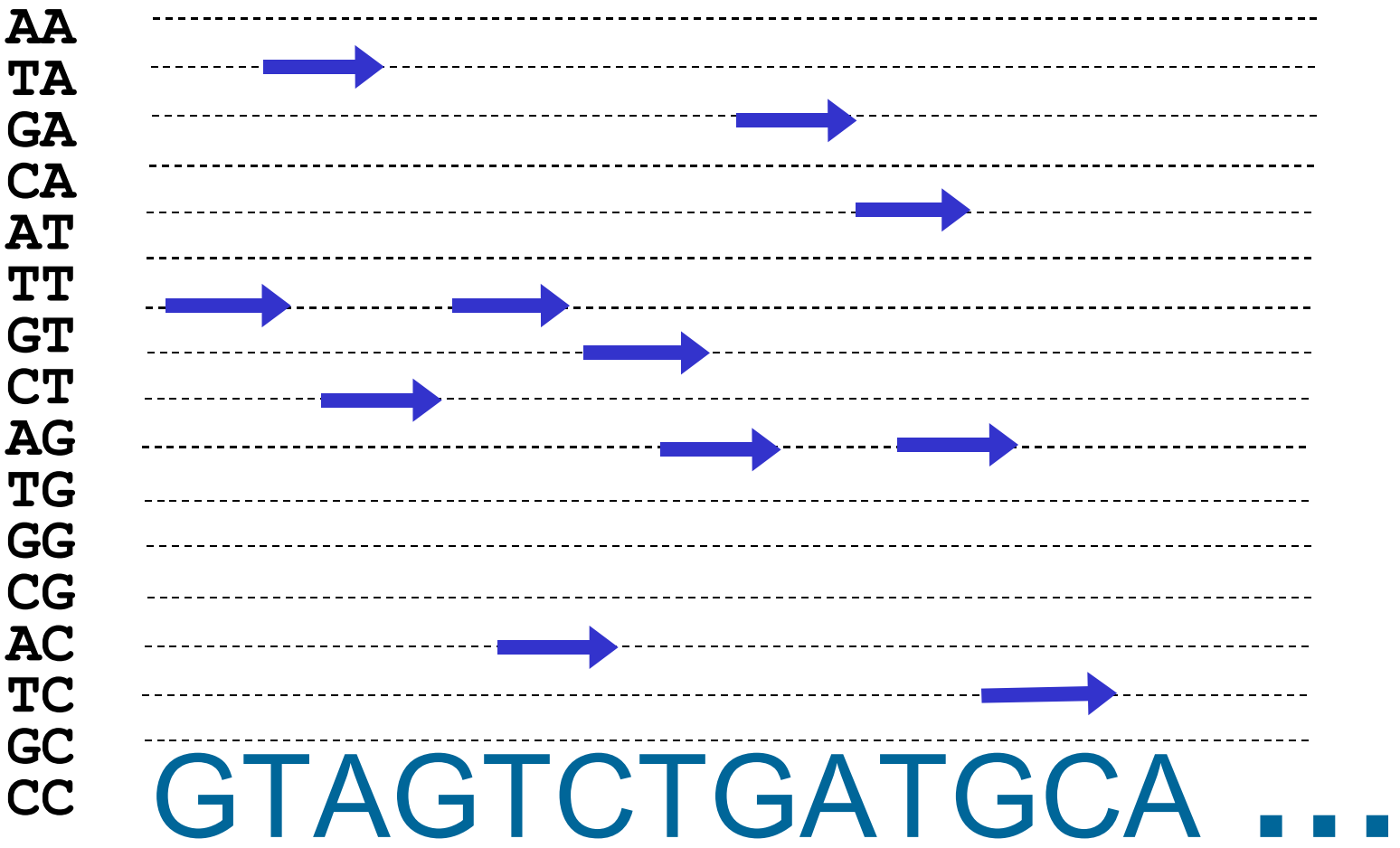
Чуриков Н.А. , Чернов В.Б., Голова Ю.Б., Нечипуренко Ю.Д. Параллельная ДНК – возможность существования // ДАН. 1988. Т. 303, № 5. С. 1254–1258.

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

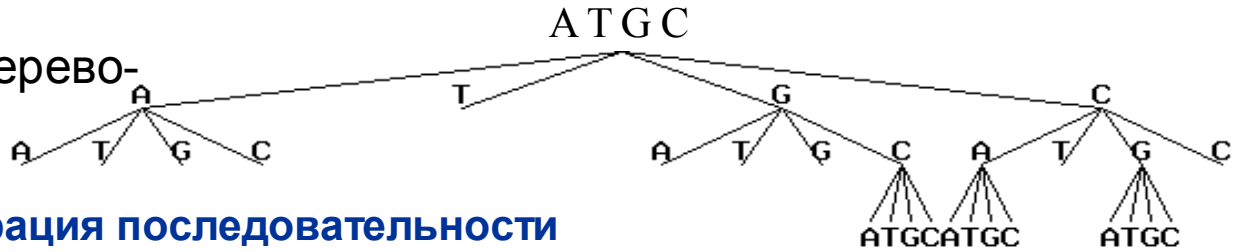
Рассмотрим разложение нуклеотидной последовательности на перекрывающиеся фрагменты

Пример разложение на перекрывающиеся динуклеотиды

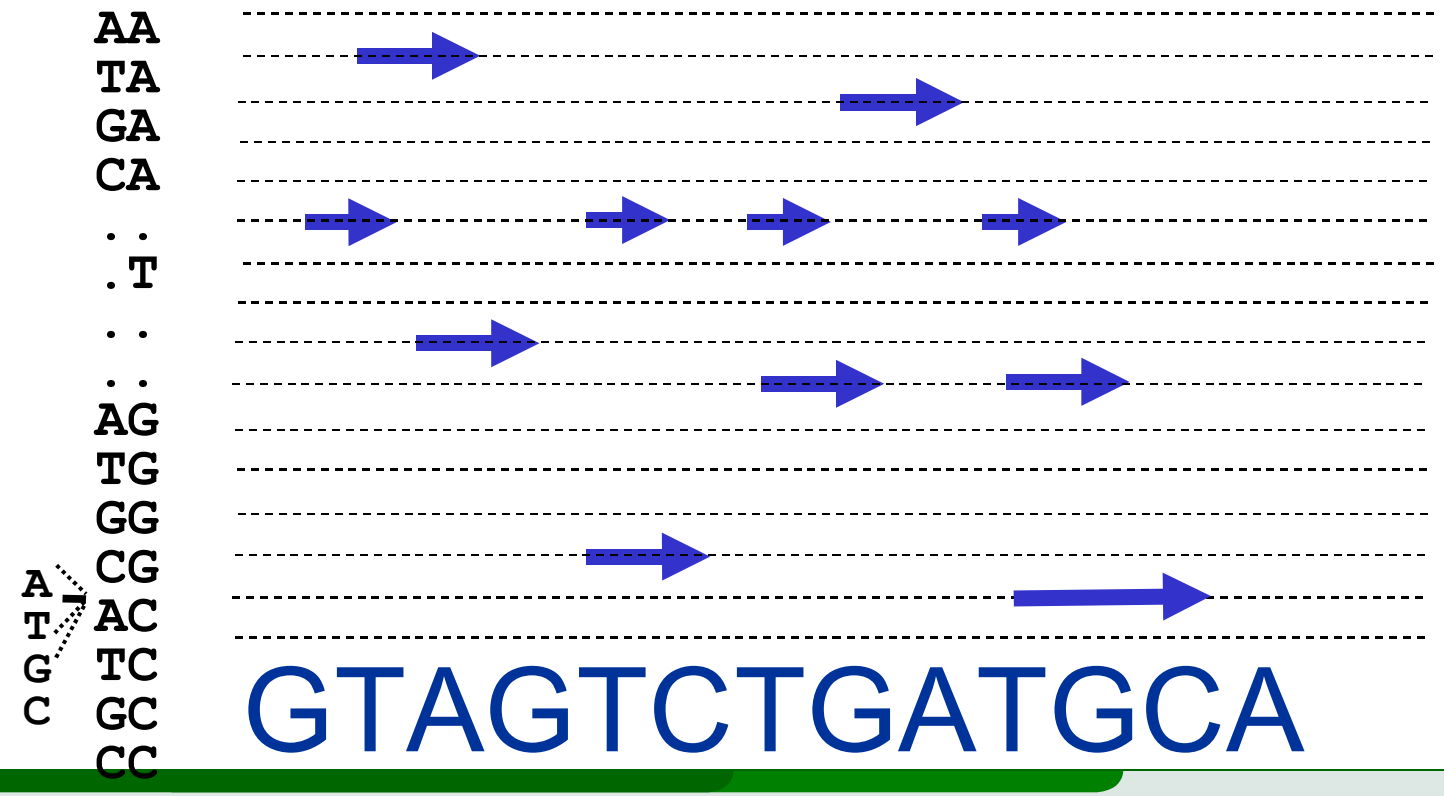
Пошаговая генерация последовательности
ДНК на основе коротких предшествующих контекстов:

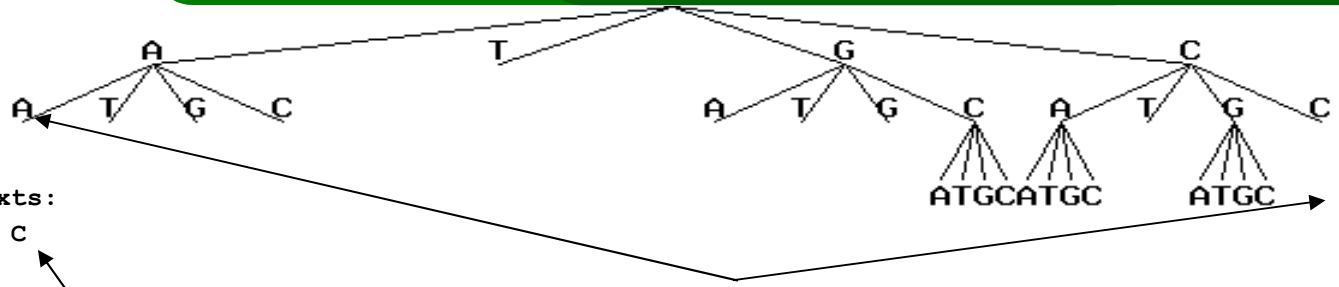


Контекстное дерево-источник:



Пошаговая генерация последовательности ДНК на основе дерева предшествующих контекстов:



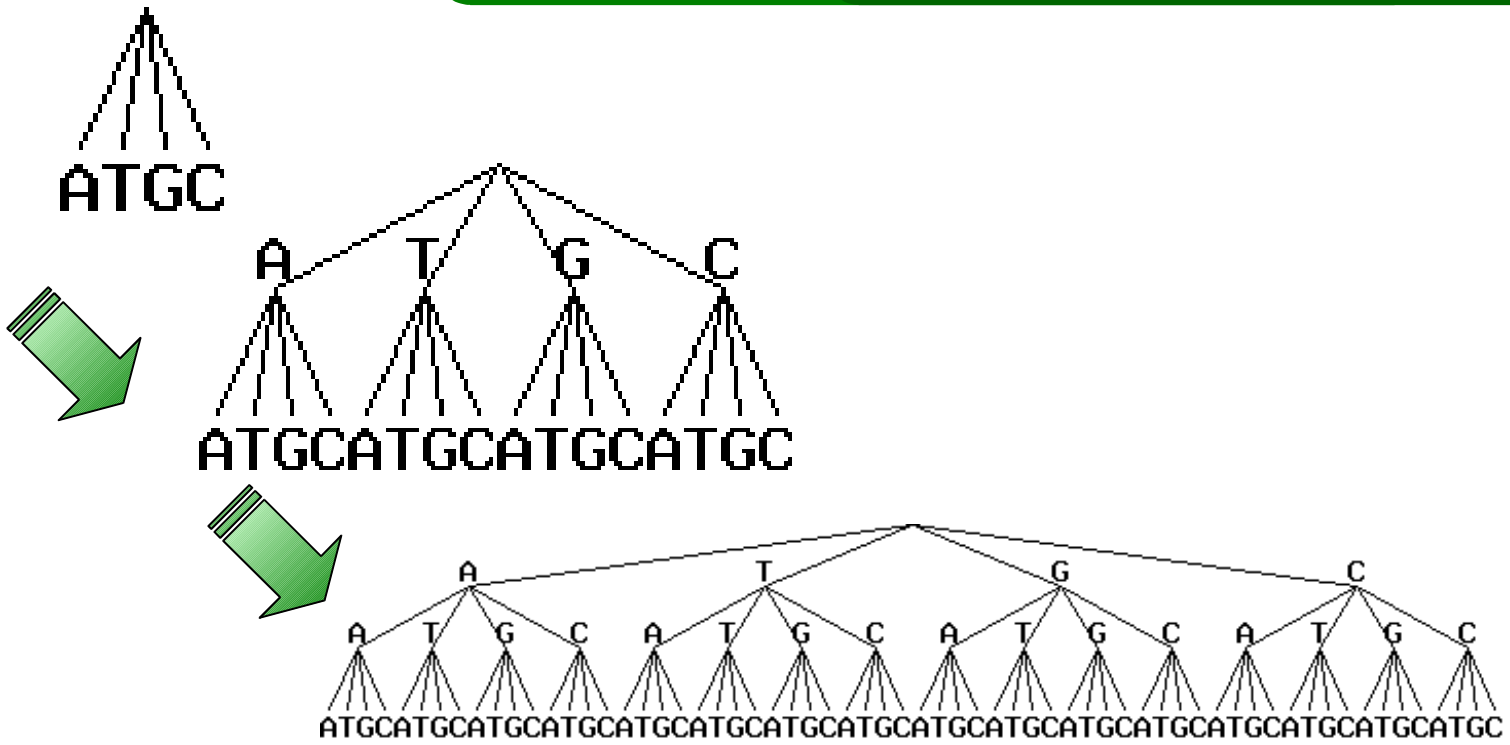



```
Frequencies after
corresponding contexts:
/contextbegin A T G C
ACG= 40 9 9 9 13
TCG= 45 6 6 15 18
GCG=160 22 20 59 59
CCG=121 16 9 34 62
AAC= 92 26 31 6 29
TAC= 46 20 14 3 9
GAC=111 23 51 23 14
CAC=120 35 43 8 34
AGC=144 38 43 23 40
TGC=177 30 51 36 60
GGC=180 35 42 56 47
CGC=148 22 25 45 56
Length is 3.
AA=472 161 104 116 91
TA=284 98 64 78 44
GA=501 111 80 197 113
CA=548 104 127 199 118
AG=584 110 144 187 143
TG=621 164 115 163 179
GG=739 165 127 266 181
TC=482 174 121 45 142
CC=638 149 161 122 206
Length is 2.
T=1868 287 474 616 491
Length is 1.
Total number of leaves 22
```

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 $4 \times 22 = 88$ значений.

Процедура построения порождающих деревьев

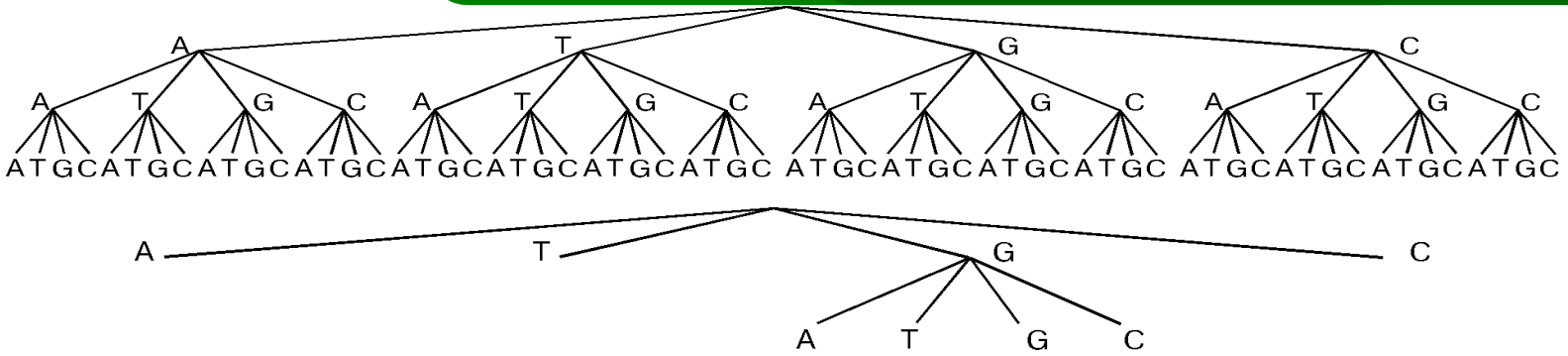


Примеры полных суффиксных деревьев, соответствующих полным наборам

4 нуклеотидов,

16 динуклеотидов,

64 тринуклеотидов



Марковский источник 3-го порядка.

Контекстное дерево-источник, полученное из полного дерева.

Вероятность порождения последовательности $X^n=X_1X_2\ldots X_n$ определяется вероятностями появления символов X_i, составляющих X^n в соответствующих контекстах S_i.

$P(X^n) = P(X_1|S_1)\,P(X_2|S_2)\ldots P(X_n|S_n)$,

где $S_i \in T'$ – контекст, в котором буква X_i содержится в слове X^n, T' - порождающее дерево.

$P(X_i|S_j)=\theta^{i}_{j}$ и $\theta^{i}_{j} = 1$ где $X_i \in \{A,T,G,C\}$,. $S_j \in D^k$, $j=1,\ldots 4^k$

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

Алгоритм выбора модели – дерева-источника

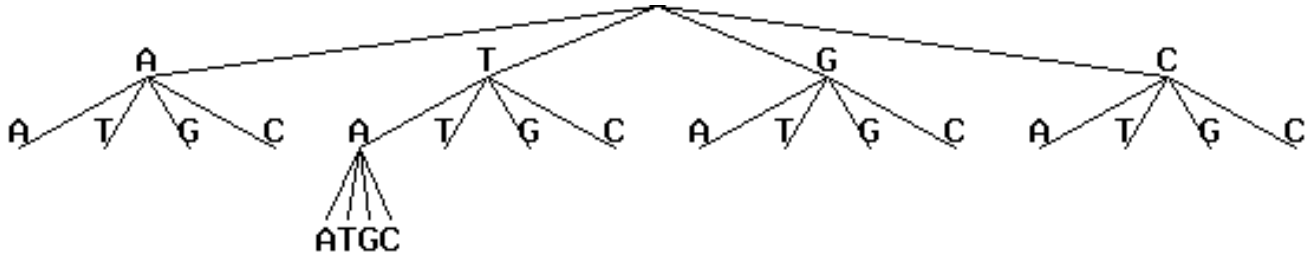
Стохастической сложностью L^0 данных относительно модели M называется минимальная сумма L сложностей сообщения X^n, при некотором заданном параметрами модели распределении θ, и сложности описания этих параметров H_n (сложности модели M).

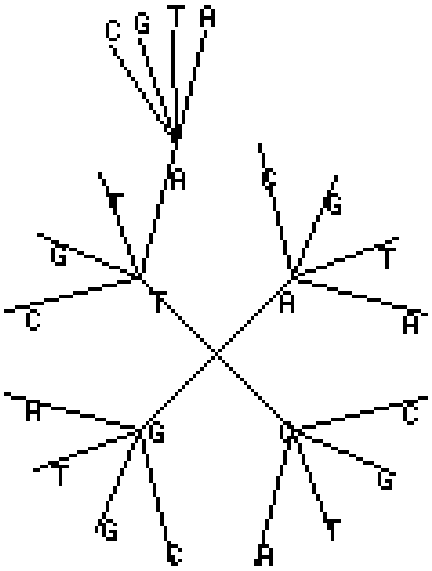
$$L^0(X^n,M) = L(X^n,\theta(X^n),M)+H_n(M)$$

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

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

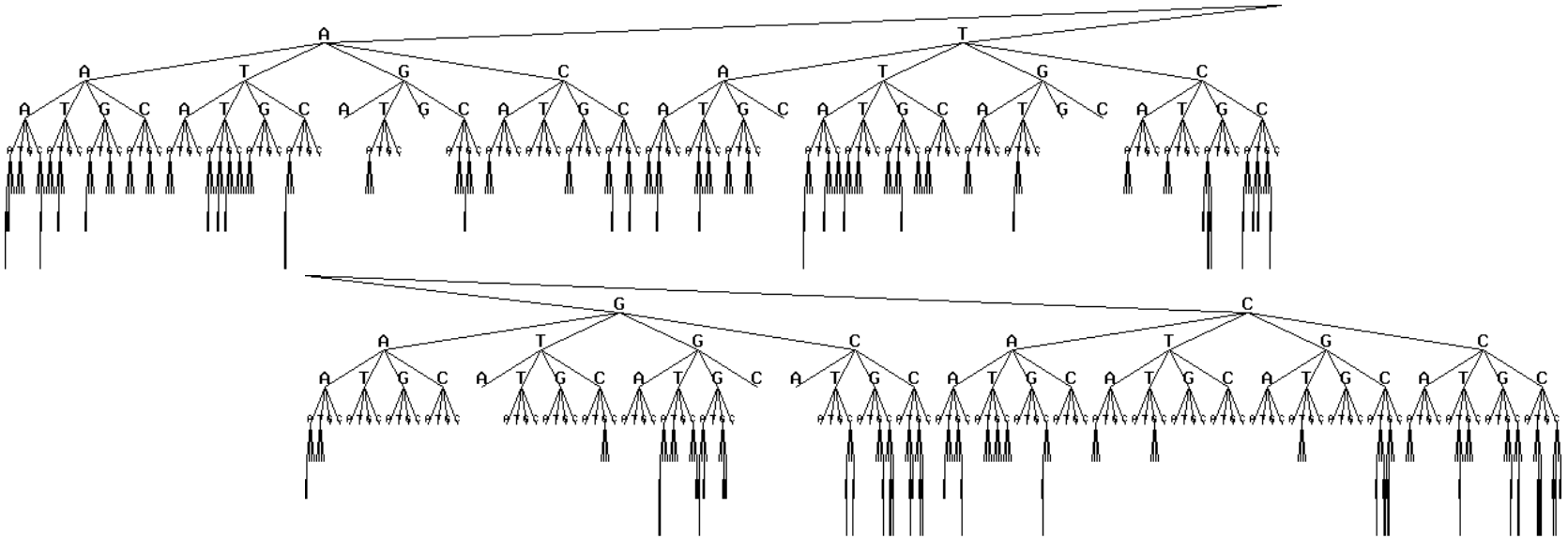
Стандартный термин древовидной модели по отношению к Марковским моделям - Variable Memory Markov Model





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 форме окружности. Рисунки автоматически генерируются программой оценки сложности и построения контекстных деревьев, доступной по адресу

http://wwwmgs.bionet.nsc.ru/mgs/programs/complexity/

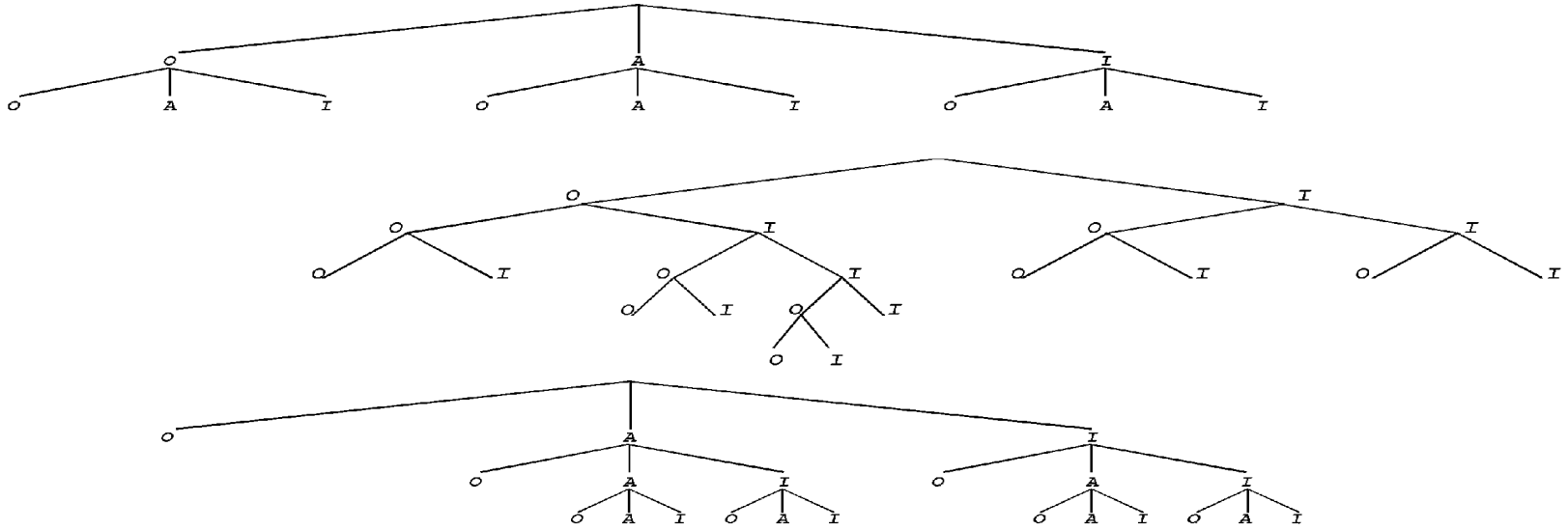


Визуальное представление полного набора статистически значимых контекстов

Mycoplasma pneumoniae полный геном. Контекстное дерево разделено на две части.

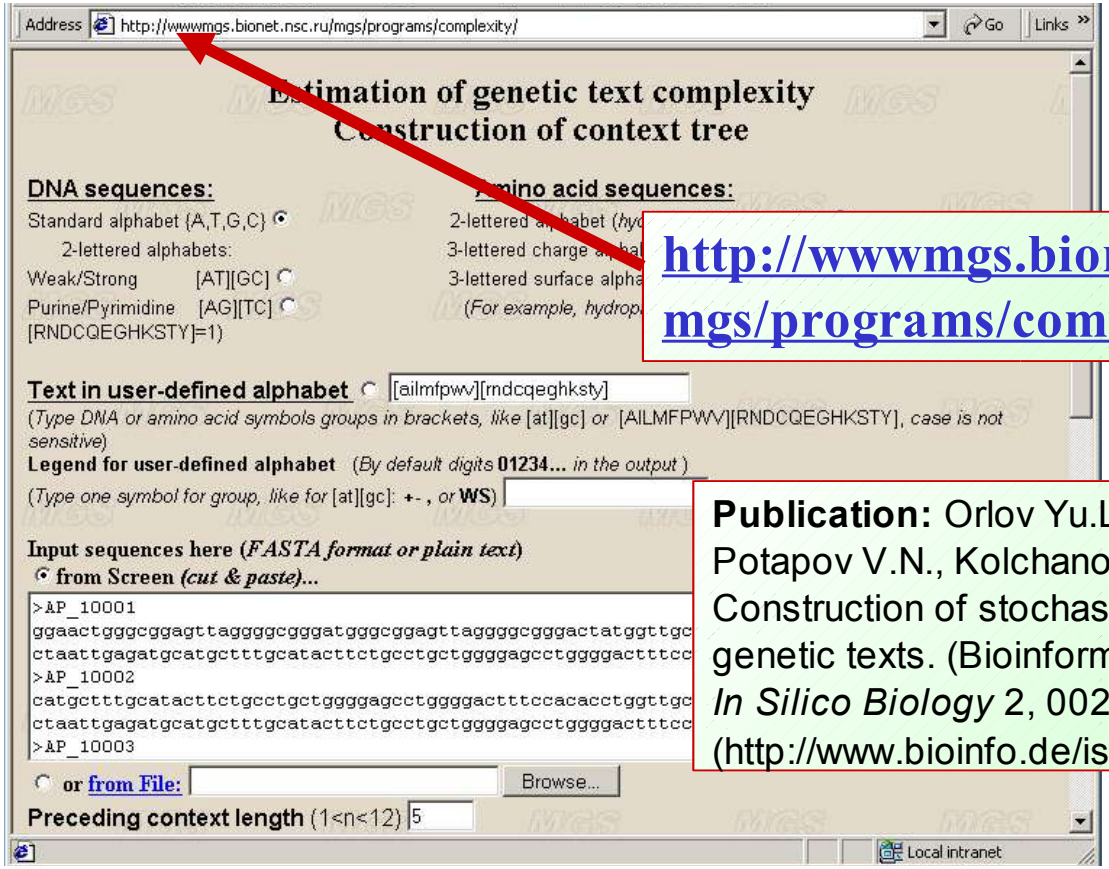
Анализ аминокислотных последовательностей с помощью деревьев-источников

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.

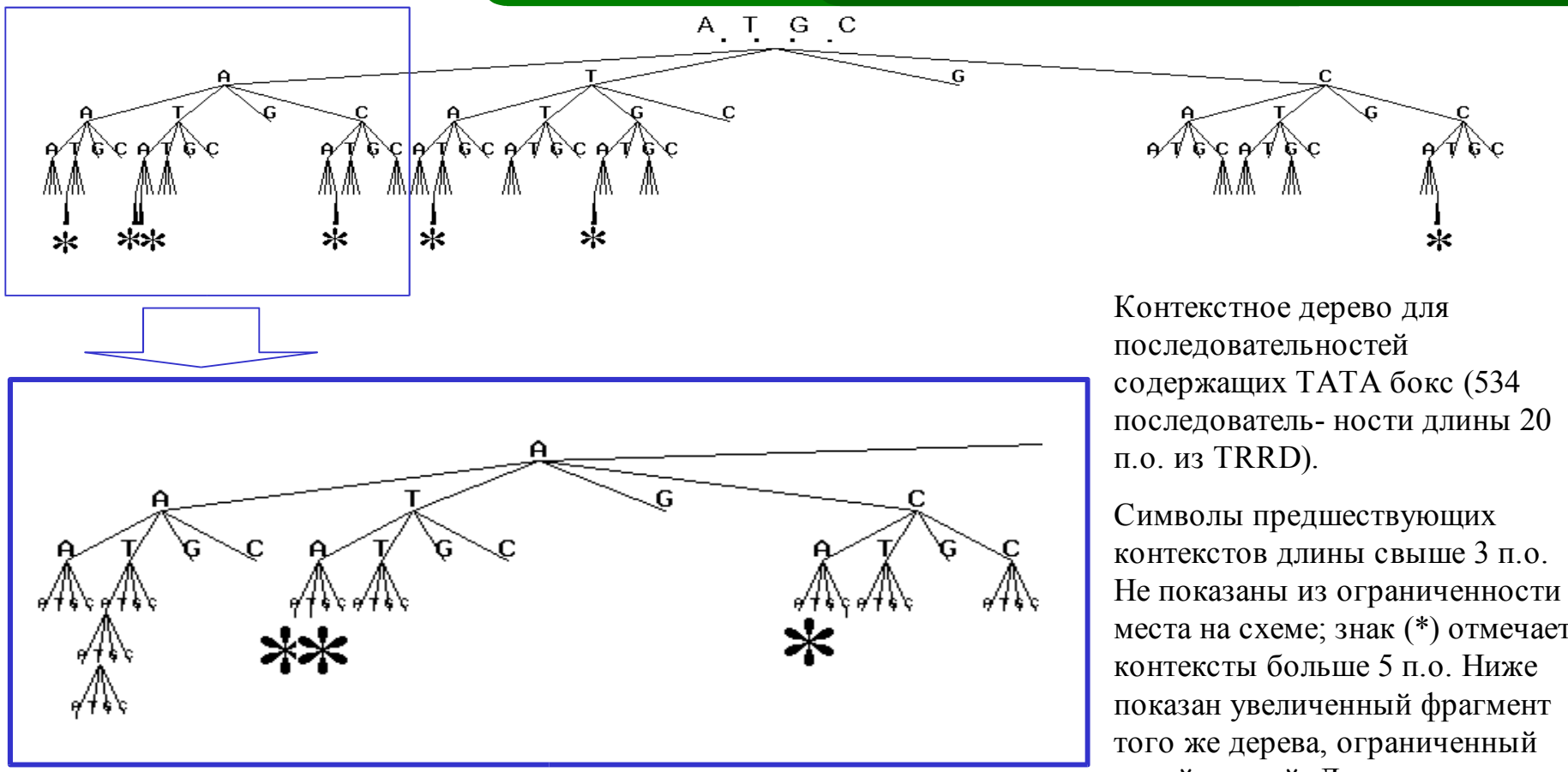
Интернет-интерфейс программы Complexity для анализа сложности генетических текстов с помощью моделей деревьев-источников



http://wwwmgs.bionet.nsc.ru/mgs/programs/complexity/

Publication: Orlov Yu.L., Filippov V.P., Potapov V.N., Kolchanov N.A. (2002) Construction of stochastic context trees for genetic texts. (Bioinformation Systems e.V.) *In Silico Biology* 2, 0022 (http://www.bioinfo.de/isb/2002/02/0022/)

Структура дерева может быть достаточно сложной для визуализации



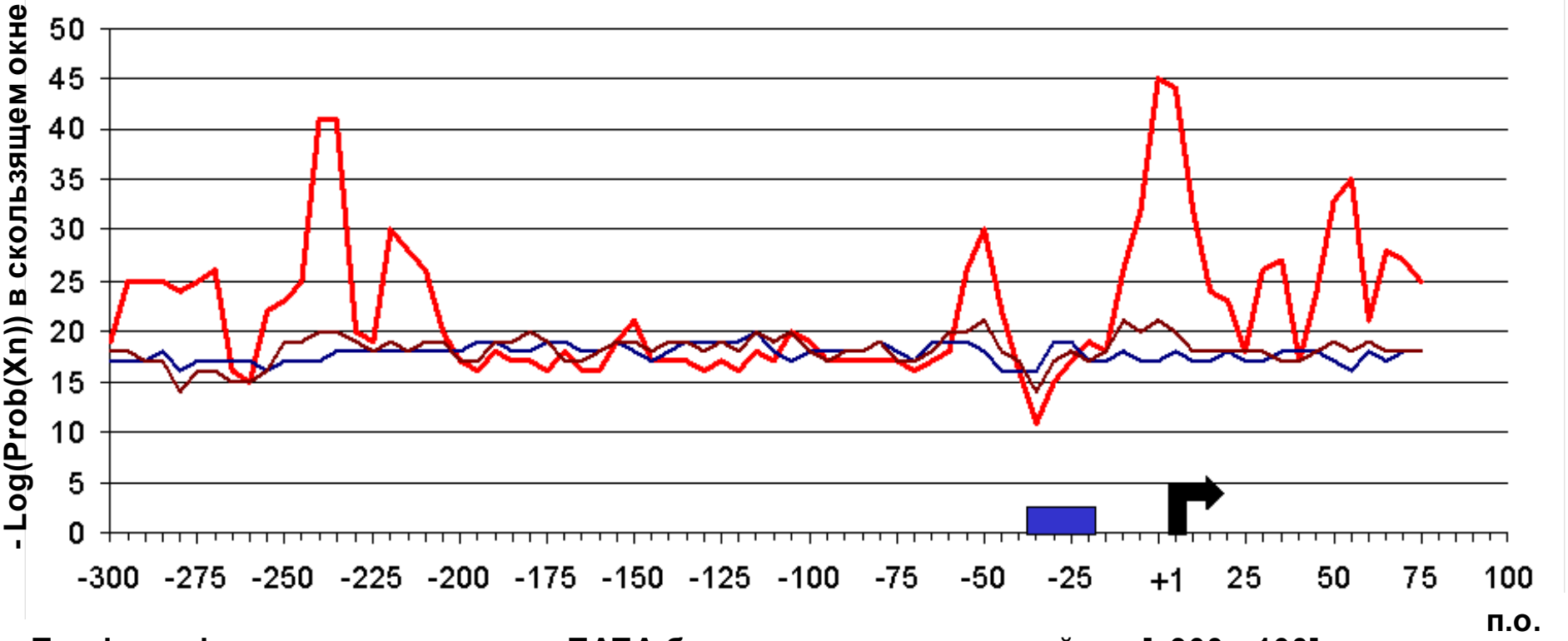
Контекстное дерево для последовательностей содержащих TATA бокс (534 последователь- ности длины 20 п.о. из TRRD).

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.

```
>S300;
cccagctctaatttcccaag
>S2777;
Ttttctataaaaagacaaa
. . .
```

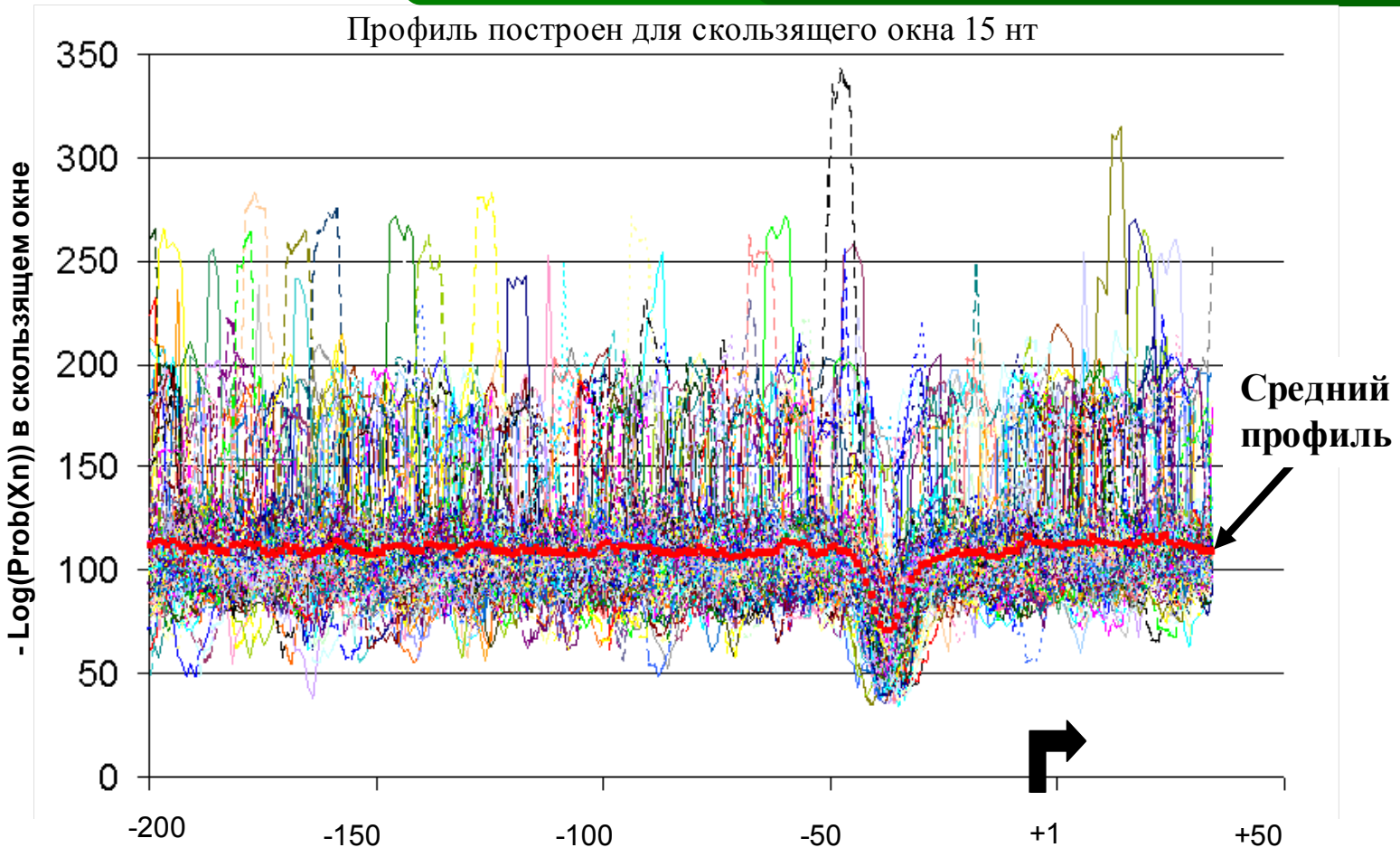
Примеры последовательностей из выборки

Предсказание TATA box-бокса с помощью марковской модели с переменной памятью



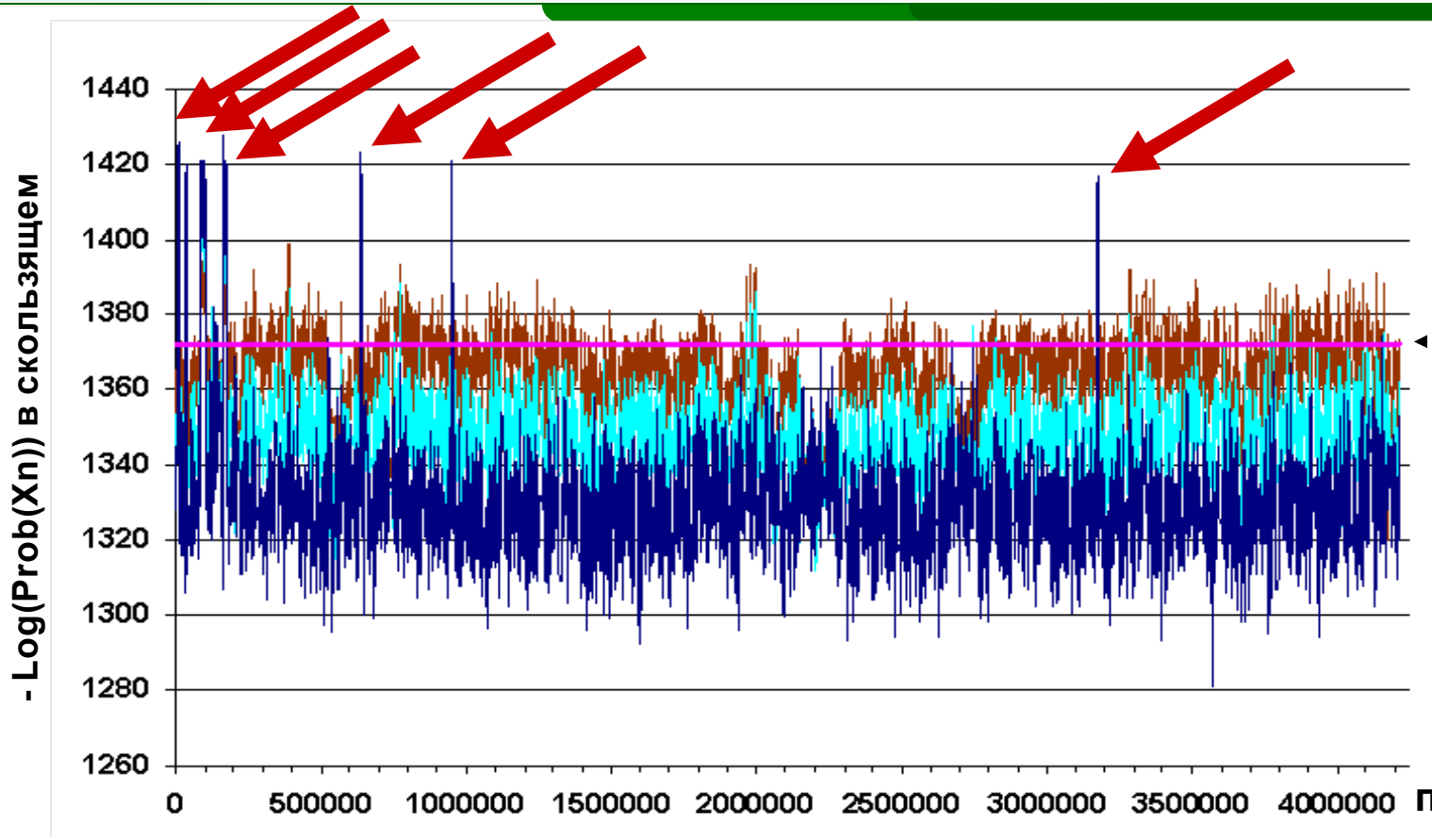
Профиль функции предсказания TATA-бокса в промоторном районе [–300;+100] гена металлотинотеин-I (metallothionein-I).
Красная линия показывает профиль предсказания с помощью модели дерева-источника; синяя и коричневая кривые соответствуют предсказанию с помощью только частот нуклеотидов и динуклеотидов в том же скользящем окне (марковские модели нулевого и первого порядка). Отмечено положение TATA бокса аннотированное в TRRD и старт транскрипции.

Предсказание TATA box-бокса с помощью марковской модели с переменной памятью



124 последовательности, содержащие TATA бокс сфазированы относительно старта транскрипции

Профиль сегментации для генома *B. subtilis*



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)

Профиль соответствия локальной модели порождения символов в скользящем окне 1000 п.о.

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)

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.