Курс «Компьютерная геномика». Лекция 2

Обзор задач биоинформатики, связанных с анализом и обработкой текстовых последовательностей

Орлов Ю.Л.

Кафедраинформационной биологии ФЕННГУ

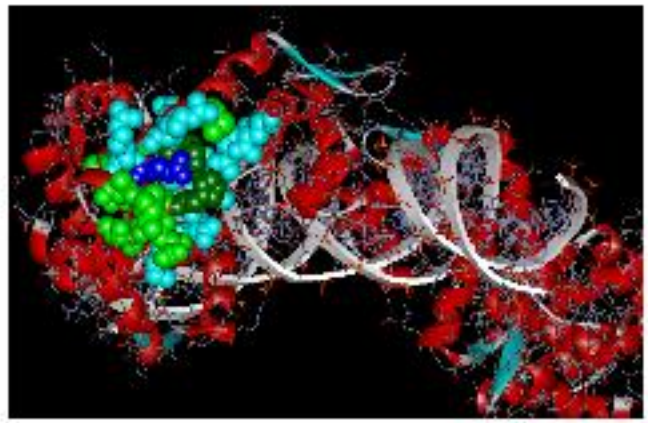
Обзор задач биоинформатики как задач компьютерного анализа генетических текстов.

Проблемы компьютерного анализа генетических текстов

Распознавание и функциональная аннотация регуляторных последовательностей генов эукариот имеет большое значение для молекулярной биологии в наступившую "пост-геномную" эру.

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.

Пример – сайты рестриктаз и сайты связывания ТФ.



Рекомендуемая литература:

David W. Mount «Bioinformatics. Sequence and genome analysis»

«Компьютерный анализ генетических текстов» (Ред.Франк-Каменецкий), 1990

Основные задачи компьютерного анализа генетических текстов:

1) поиск гомологии и выравнивание генетических текстов, множественное выравнивание

2) статистический анализ генетических текстов, исследование структуры повторов и модели порождения символьных последовательностей, сегментация геномов;

3) предсказание кодирующих участков генов и открытых рамок считывания;

4) предсказание функциональных сигналов (функциональных сайтов и регуляторных районов);

5) **анализ вторичной структуры РНК** и сигналов трансляции;

6) **анализ аминокислотных последовательностей** белков, предсказание вторичной структуры, функциональных сайтов и доменов глобулярных белков по их аминокислотным последовательностям;

7) **филогенетические сравнения**;

8) ДНК-чипы – экспрессионные кривые

9) задачи оперирования с большими массивами информации и управления (Интернет-навигации) разрозненными специализированными базами данных

Основные задачи компьютерного анализа генетических текстов:

1) поиск гомологии и выравнивание генетических текстов, множественное выравнивание

Рассматриваемые вопросы

- Дот-матрица или метод диаграмм для сравнения последовательностей
- Выравнивание последовательностей с помощью динамического программирования
- Поиск локального выравнивания последовательностей.
- Множественное выравнивание последовательностей.
- Поиск гомологии в базах данных. Методы FASTA и BLAST для поиска в базах данных.

Дот-матрица или метод диаграмм для сравнения последовательностей

1970 - A.J.Gibbs and G.A.McIntyre

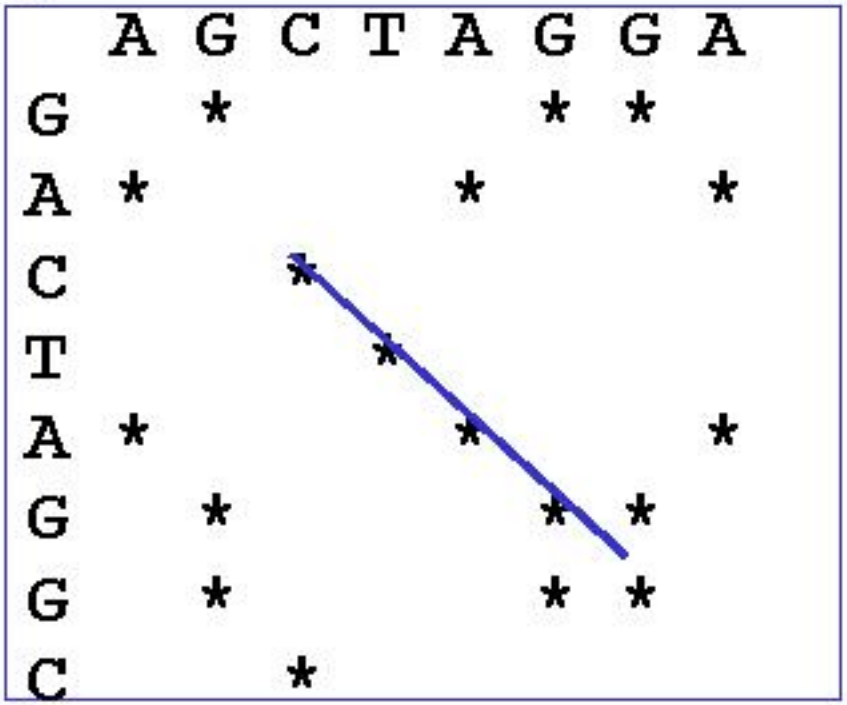
Сравнение последовательностей.

Точечная матрица

совпадений –

dot-matrix

(Дот-матрица).

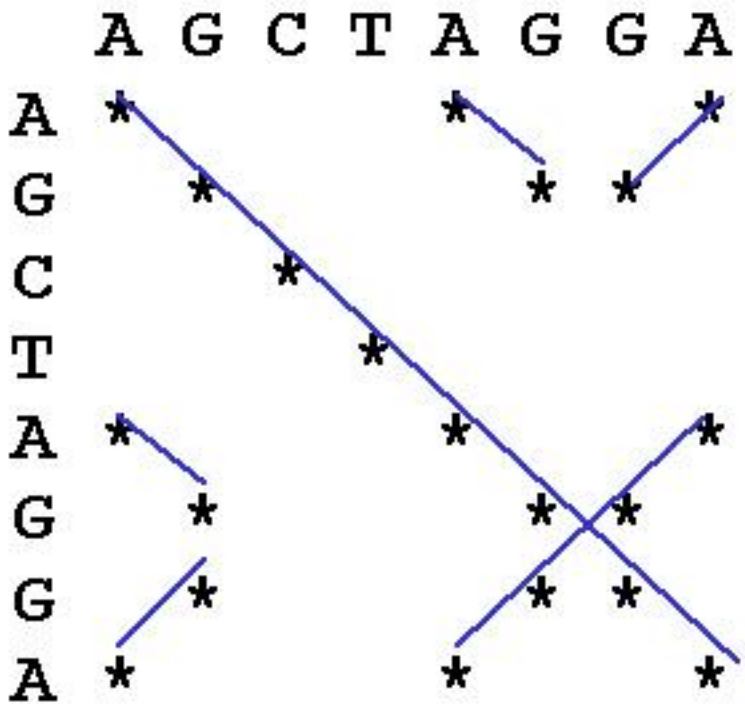


Рассмотрим две последовательности: AGCTAGGA и GACTAGGC. Диагональ точек выявляет общую подпоследовательность CTAGG

Рассмотрим одну последовательность AGCTAGGA.

В этом случае матрица всегда симметрична.

Диагонали точек выявляют прямые и симметричные повторы



Короткий повтор: AG

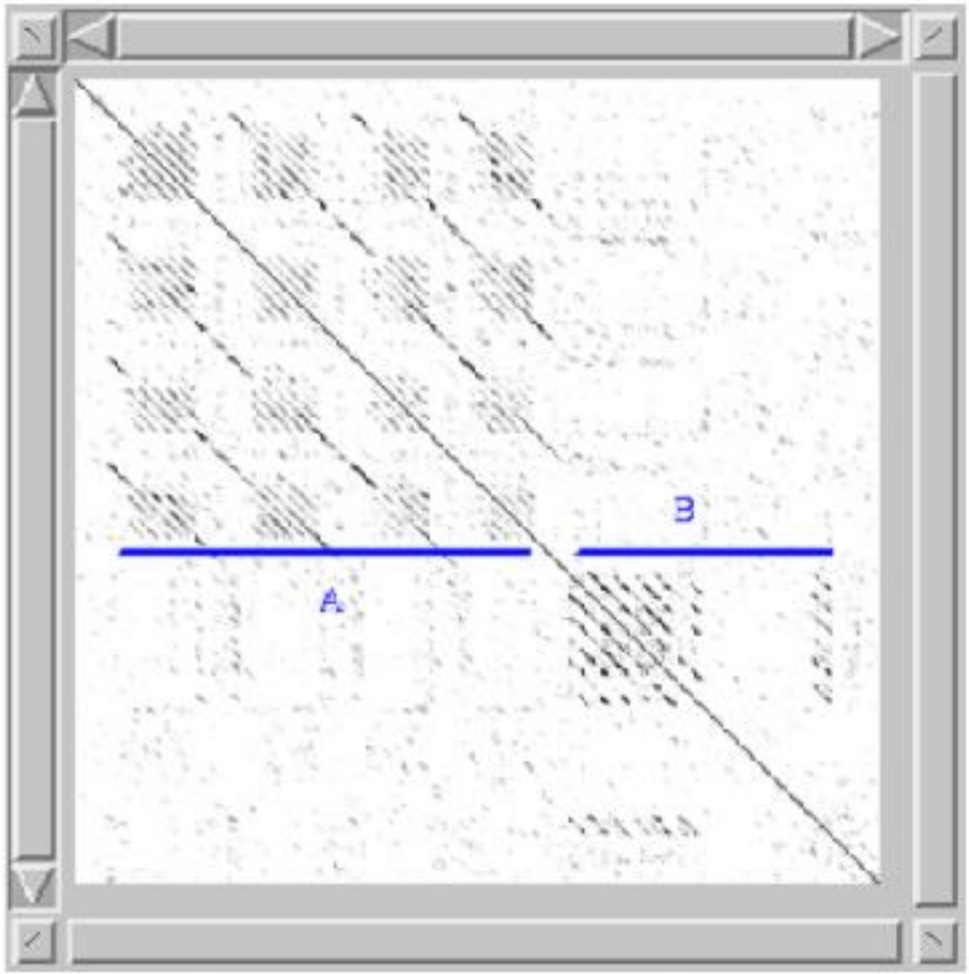
Симметричный повтор: AGGA

Пример произвольной символьной последовательности.

Матрица позволяет находить общие подслова в последовательностях символов любого алфавита – нуклеотидном, аминокислотном, на естественных языках.

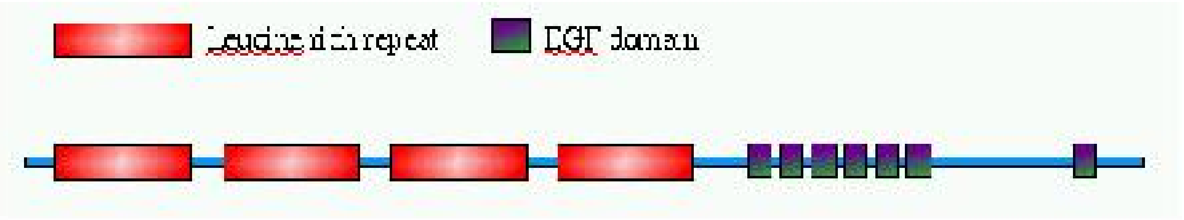
http://www.isrec.isb-sib.ch/java/dotlet/repeats.html

Усовершенствование сравнения.



На данном примере представлен график (точечная матрица) для белка SPLIT *D.melanogaster*. Видны 4 повтора в N-концевом домене и (A) и 6 в C-концевом (B)

Результат поиска повторов с помощью точечной матрицы гомологии. Белок SPLIT *D.melanogaster.* Аминокислотная последовательность и структура повторов.



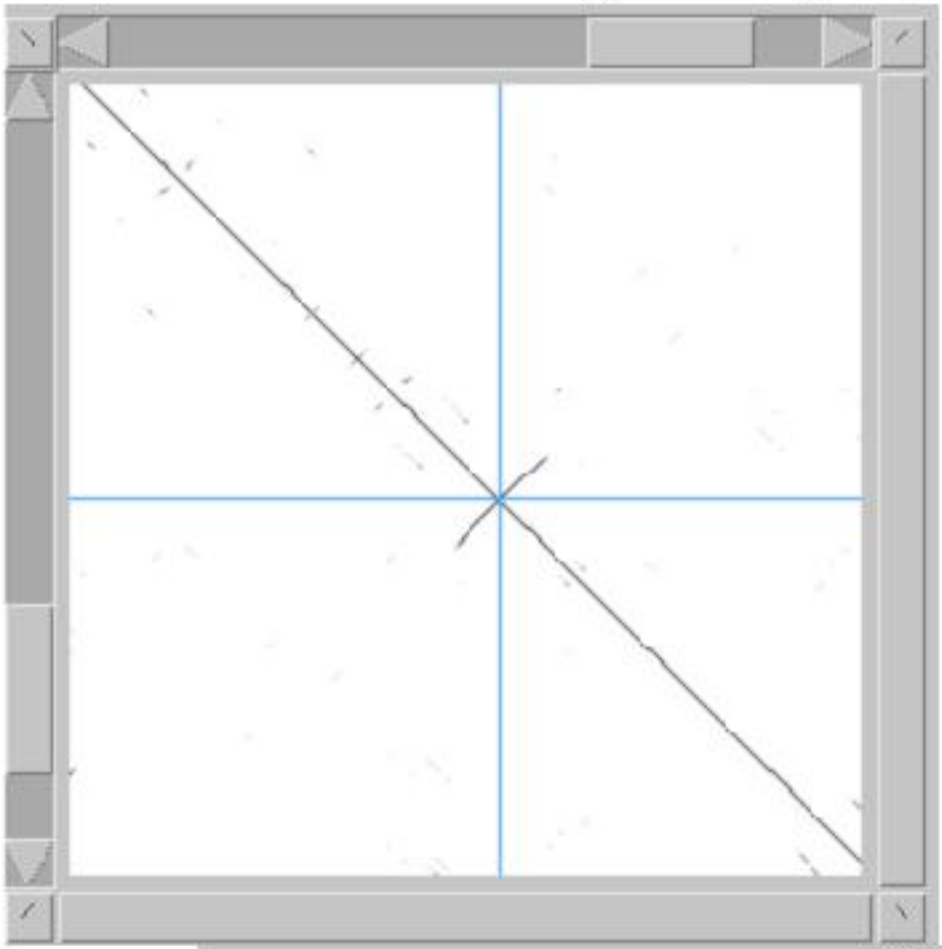
- SLIT_DROME (P24014):

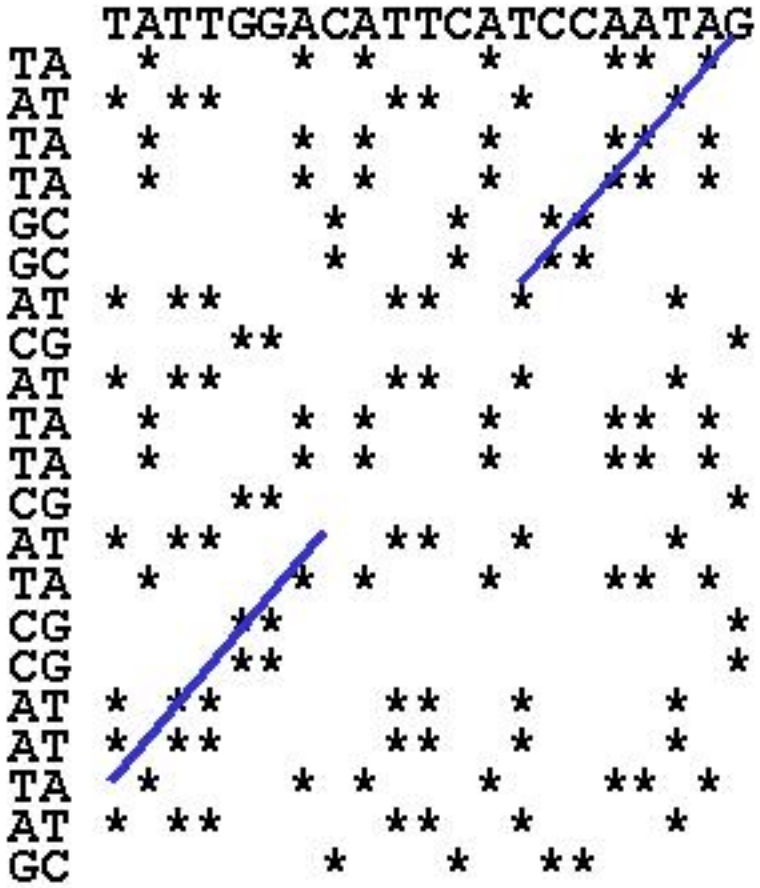
MAAPSRTTLMPPPFRLQLRLLILPILLLLRHDAVHAEPYSGGFGSSAVSSGGLGSVGIHIPGGGVGVITEARCPRVCSCT
GLNVDCSHRGLTSVPRKISADVERLELQGNNLTVIYETDFQRLTKLRMLQLTDNQIHTIERNSFQDLVSLERLDISNNVI
TTVGRRVFKGAQSLRSLQLDNNQITCLDEHAFKGLVELEILTLNNNNLTSLPHNIFGGLGRLRALRLSDNPFACDCHLSW
LSRFLRSATRLAPYTRCQSPSQLKGQNVADLHDQEFKCSGLTEHAPMECGAENSCPHPCRCADGIVDCREKSLTSVPVTL
PDDTTDVRLEQNFITELPPKSFSSFRRLRRIDLSNNNISRIAHDALSGLKQLTTLVLYGNKIKDLPSGVFKGLGSLRLLL
LNANEISCIRKDAFRDLHSLSLLSLYDNNIQSLANGTFDAMKSMKTVHLAKNPFICDCNLRWLADYLHKNPIETSGARCE
SPKRMHRRRIESLREEKFKCSWGELRMKLSGECRMDSDCPAMCHCEGTTVDCTGRRLKEIPRDIPLHTTELLLNDNELGR
ISSDGLFGRLPHLVKLELKRNQLTGIEPNAFEGASHIQELQLGENKIKEISNKMFLGLHQLKTLNLYDNQISCVMPGSFE
HLNSLTSLNLASNPFNCNCHLAWFAECVRKKSLNGGAARCGAPSKVRDVQIKDLPHSEFKCSSENSEGCLGDGYCPPSCT
CTGTVVACSRNQLKEIPRGIPAETSELYLESNEIEQIHYERIRHLRSLTRLDLSNNQITILSNYTFANLTKLSTLIISYN
KLQCLQRHALSGLNNLRVVSLHGNRISMLPEGSFEDLKSLTHIALGSNPLYCDCGLKWFSDWIKLDYVEPGIARCAEPEQ
MKDKLILSTPSSSFVCRGRVRNDILAKCNACFEQPCQNQAQCVALPQREYQCLCQPGYHGKHCEFMIDACYGNPCRNNAT
CTVLEEGRFSCQCAPGYTGARCETNIDDCLGEIKCQNNATCIDGVESYKCECQPGFSGEFCDTKIQFCSPEFNPCANGAK
CMDHFTHYSCDCQAGFHGTNCTDNIDDCQNHMCQNGGTCVDGINDYQCRCPDDYTGKYCEGHNMISMMYPQTSPCQNHEC
KHGVCFQPNAQGSDYLCRCHPGYTGKWCEYLTSISFVHNNSFVELEPLRTRPEANVTIVFSSAEQNGILMYDGQDAHLAV
ELFNGRIRVSYDVGNHPVSTMYSFEMVADGKYHAVELLAIKKNFTLRVDRGLARSIINEGSNDYLKLTTPMFLGGLPVDP
AQQAYKNWQIRNLTSFKGCMKEVWINHKLVDFGNAQRQQKITPGCALLEGEQQEEEDDEQDFMDETPHIKEEPVDPCLEN
KCRRGSRCVPNSNARDGYQCKCKHGQRGRYCDQGEGSTEPPTVTAASTCRKEQVREYYTENDCRSRQPLKYAKCVGGCGN
QCCAAKIVRRRKVRMVCSNNRKYIKNLDIVRKCGCTKKCY

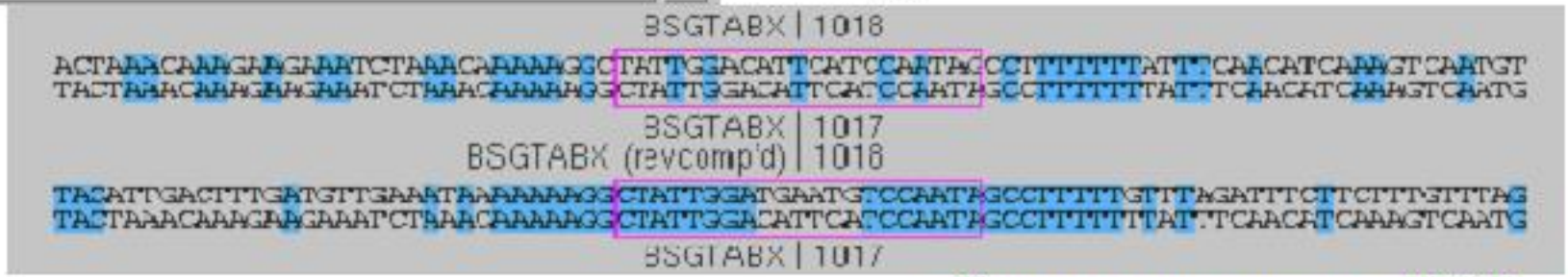
Пример взят: http://www.isrec.isb-sib.ch/java/dotlet/repeats.html

Точечная матрица гомологии. Представлен инвертированный повтор (X-форма)

Bacillus subtilis UTP-glucose-1-phosphate uridylyltransferase gene







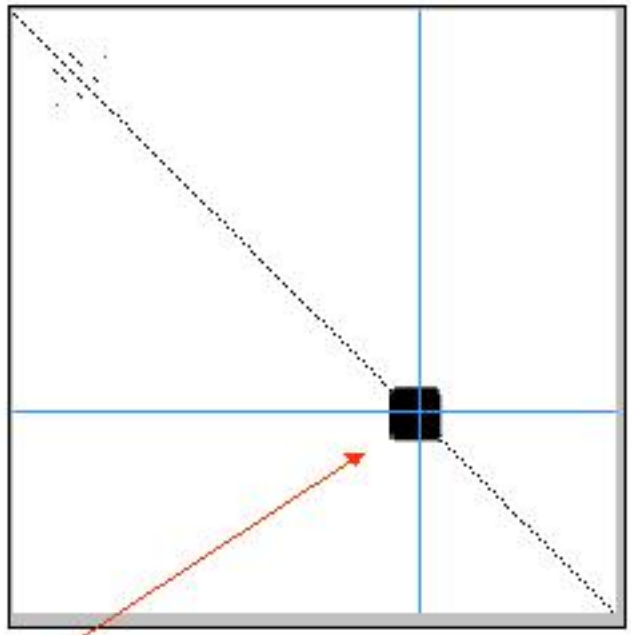
Точечная матрица гомологии.

Участки низкой сложности

Белок SERA_PLAFG (P13823) малярийного плазмодия

содержит серин-богатый участок

Расшифровать структуру квадрата



- SERA_PLAFG (P13823):
  MKSYISLFFILCVIFNKNVIKCTGESQTGNTGGGQAGNTVGDQAGSTGGSPQGSTGASQPGSSEPSNPVSSGHSVSTVSV
  SQTSTSSEKQDTIQVKSALLKDYMGLKVTGPCNENFIMFLVPHIYIDVDTEDTNIELRTTLKETNNAISFESNSGSLEKK
  KYVKLPSNGTTGEQGSSTGTVRGDTEPISDSSSSSSSSSSSSSSSSSSSSSSSSSSSSSESLPANGPDSPTVKP
  PRNLQNICETGKNFKLVVYIKENTLIIKWKVYGETKDTTENNKVDVRKYLINEKETPFTSILIHAYKEHNGTNLIESKNY
  ALGSDIPEKCDTLASNCFLSGNFNIEKCFQCALLVEKENKNDVCYKYLSEDIVSNFKEIKAETEDDDEDDYTEYKLTESI
  DNILVKMFKTNENNDKSELIKLEEVDDSLKLELMNYCSLLKDVDTTGTLDNYGMGNEMDIFNNLKRLLIYHSEENINTLK
  NKFRNAAVCLKNVDDWIVNKRGLVLPELNYDLEYFNEHLYNDKNSPEDKDNKGKGVVHVDTTLEKEDTLSYDNSDNMFCN
  KEYCNRLKDENNCISNLQVEDQGNCDTSWIFASKYHLETIRCMKGYEPTKISALYVANCYKGEHKDRCDEGSSPMEFLQI
  IEDYGFLPAESNYPYNYVKVGEQCPKVEDHWMNLWDNGKILHNKNEPNSLDGKGYTAYESERFHDNMDAFVKIIKTEVMN
  KGSVIAYIKAENVMGYEFSGKKVQNLCGDDTADHAVNIVGYGNYVNSEGEKKSYWIVRNSWGPYWGDEGYFKVDMYGPTH
  CHFNFIHSVVIFNVDLPMNNKTTKKESKIYDYYLKASPEFYHNLYFKNFNVGKKNLFSEKEDNENNKKLGNNYIIFGQDT
  AGSGQSGKESNTALESAGTSNEVSERVHVYHILKHIKDGKIRMGMRKYIDTQDVNKKHSCTRSYAFNPENYEKCVNLCNV

Выравнивание последовательностей с помощью динамического программирования

Выравнивание последовательностей с помощью динамического программирования. По этапам

Рассмотрим две последовательности: GATCTA и GATCA

Needleman-Wunsch

Алгоритм Нидльмана-Вунша

		G	A	T	C	T	A
	G	1	0	0	0	0	0
	A	0	1	0	0	0	1
	T	0	0	1	0	1	0
	C	0	0	0	1	0	0
	A	0	1	0	0	0	1

		G	A	T	C	T	A
	G	1	0	0	0	0	0
	A	0	2	0	0	0	1
	T	0	0	3	0	1	0
	C	0	0	0	4	0	0
	A	0	1	0	0	0	5

Итоговое выравнивание двух последовательностей:

GATCTA

GATC-A

Пример – выравнивание слов COELANCANTH и PELICAN с использованием простой схемы подсчета: +1 для совпадающих букв (match), -1 для несовпадений (mismatches), и -1 для пробелов (gaps).

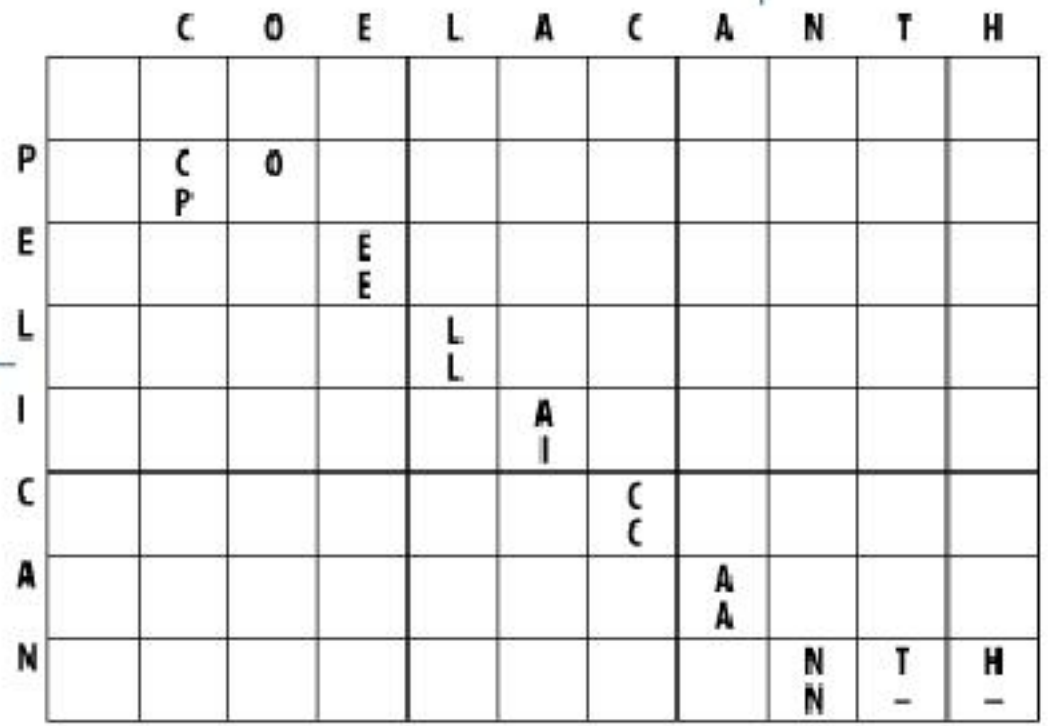
COELACANTH

P-ELICAN--

COELACANTH

-PELICAN--

Пример матрицы выравнивания



Инициализация матрицы сравнения

		C	O	E	L	A	C	A	N	T	H
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
P	↑ -1										
E	↑ -2										
L	↑ -3										
I	↑ -4										
C	↑ -5										
A	↑ -6										
N	↑ -7										

Начало заполнения матрицы сравнения

```
CO      CO
-P      P-
```

		C	O	E	L	A	C	A	N	T	H
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
P	↑ -1	↖ -1									

Заполнение второй строки матрицы сравнения

		C	O	E	L	A	C	A	N	T	H
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
P	↑ -1	↖ -1	↖ -2								

Заполненная матрица сравнения для тестового примера

		C	O	E	L	A	C	A	N	T	H
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
P	↑ -1	↖ -1	↖ -2	↖ -3	↖ -4	↖ -5	↖ -6	↖ -7	↖ -8	↖ -9	↖ -10
E	↑ -2	↖ -2	↖ -2	↖ -1	← -0	← -3	← -4	← -5	← -6	← -7	← -8
L	↑ -3	↖ -3	↖ -3	← -2	↖ -2	← -1	← -2	← -3	← -4	← -5	← -6
I	↑ -4	↖ -4	↑ -4	↑ -3	↑ -1	↖ -1	↖ -2	↖ -1	↖ -4	↖ -5	↖ -6
C	↑ -5	↖ -3	← -4	↑ -4	↑ -2	↖ -2	↖ -0	← -1	← -2	← -3	← -4
A	↑ -6	↑ -4	↖ -4	↖ -5	↑ -3	↖ -1	↑ -1	↖ -1	← -0	← -1	← -2
N	↑ -7	↑ -5	↖ -5	↖ -5	↑ -4	↑ -2	↖ -2	← -0	↖ -2	← -1	← -0

COELACANTH

-PELICAN-

Local Alignment: Smith-Waterman

Алгоритм Смита-Уотермана является модификацией алгоритма Нидльмана-Вунша

Отличия

•Углы матрицы инициализируются 0 вместо увеличения штрафа за пробел.

•Максимальный счет никогда не бывает меньше нуля, и указатель не записывается до тех пор пока счет больше нуля.

•Обратный проход начинается с наибольшего счета в матрице (а не с конца) и заканчивается при счете 0 (а не в начале матрицы).

ELACAN

ELICAN

		C	O	E	L	A	C	A	N	T	H
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	1	0	0	0	0	0	0	0
L	0	0	0	0	↖2	←1	0	0	0	0	0
I	0	0	0	0	↑1	↖1	0	0	0	0	0
C	0	1	0	0	0	0	↖2	0	0	0	0
A	0	0	0	0	0	1	0	↖3	←2	←1	0
N	0	0	0	0	0	0	0	↑1	↖4	←3	←2

Поиск локального выравнивания последовательностей

1981 - Mike Waterman, Temple Smith

Наиболее биологически значимые районы в ДНК и белках – локальные районы, которые выравниваются хорошо, в то время как остающиеся участки менее значимы.

Две меры оценки выравнивания (подсчета, скора – score) – счет сходства (близости) последовательностей и счет расстояния между ними.

Smith T.F. and Waterman M.S. (1981) Identification of common molecular subsequences. *J.Mol.Biol.* **147**: 195-197.

Значимость счета выравнивания

Karlin S. and Altschul S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA 87:2264-2268.

Множественное выравнивание последовательностей

Johnson and Doolitle, 1986

GCG PILEUP

CLUSTALW (Thompson et al., 1994)

(Baylor College of Medicine,

http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html)

Gribskov et al., 1987

```
A00825_hsp17_L.Merr.            TTACAATCTCCCTAGTTTC---TAATCTCAGCTAAGAAAA-
ACCAAAAGA
A00826_hsp17.5-M_L.Merr.        TTTTGATCTCCCAAGTTTC---
AAATCTCGCGAATAAATATATCAAAAGA
A00662_mult.r.g.1b_Rat          GCCGCTGCTCCCATCTTC----
GAGGCTCAGCTCAACTCAGAGCTACTT-
A00172_mult.r.g.1b_Mus          GCCGCTGCTTCCATCTTCT---
GAGGTTCCGCTCAACTCAGAGCTACT--
A00597_M35021_70hsp_Mus         TCCAGAG---ACAAGCGAA---GACAAGAGAAGCAGAGC-
GAGCGGCGC-
A00597_M76613_70hsp_Mus         TCCAGAG---ACAAGCGAA---
GACAAGAGAAGCAGAGCAGAGCGGCGC-
A00551_heme_ox.1_Homo           GCACGAA----CGAGCCC-----
AGCACCGGCCGGATGGAGCGTCCGCAA
A00569_heme_ox.1_Rat            GCCGGAG----CAGAGCC-----
ATCTCGAGCGGAGCCCGGAGCCTGAAG
A00552_hsp70_Xenopus            TACTTAC----TCCAAA
```

Поиск гомологии в базах данных. Методы FASTA и BLAST для поиска в базах данных

FASTA (Pearson and Lipman, 1988)

BLAST (Altschul et al., 1990)

www.ncbi.mlm.nih.gov/BLAST

Gapped-BLAST (в три раза быстрее)

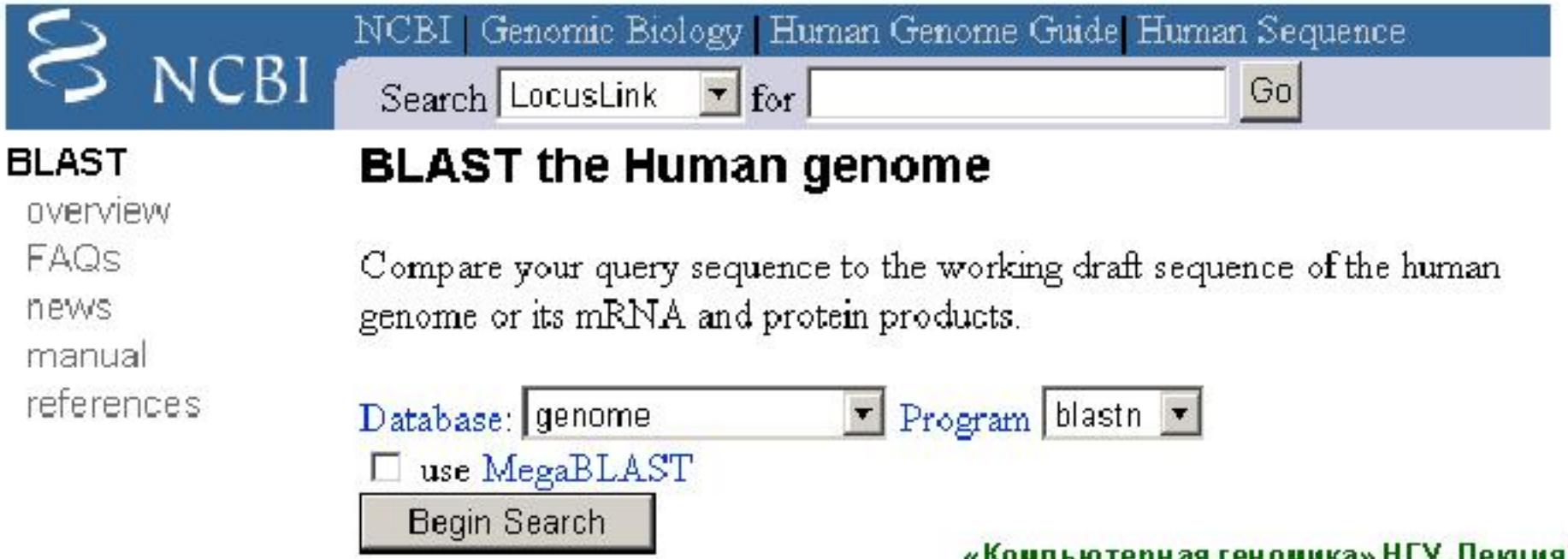
PSI-BLAST (большая чувствительность)

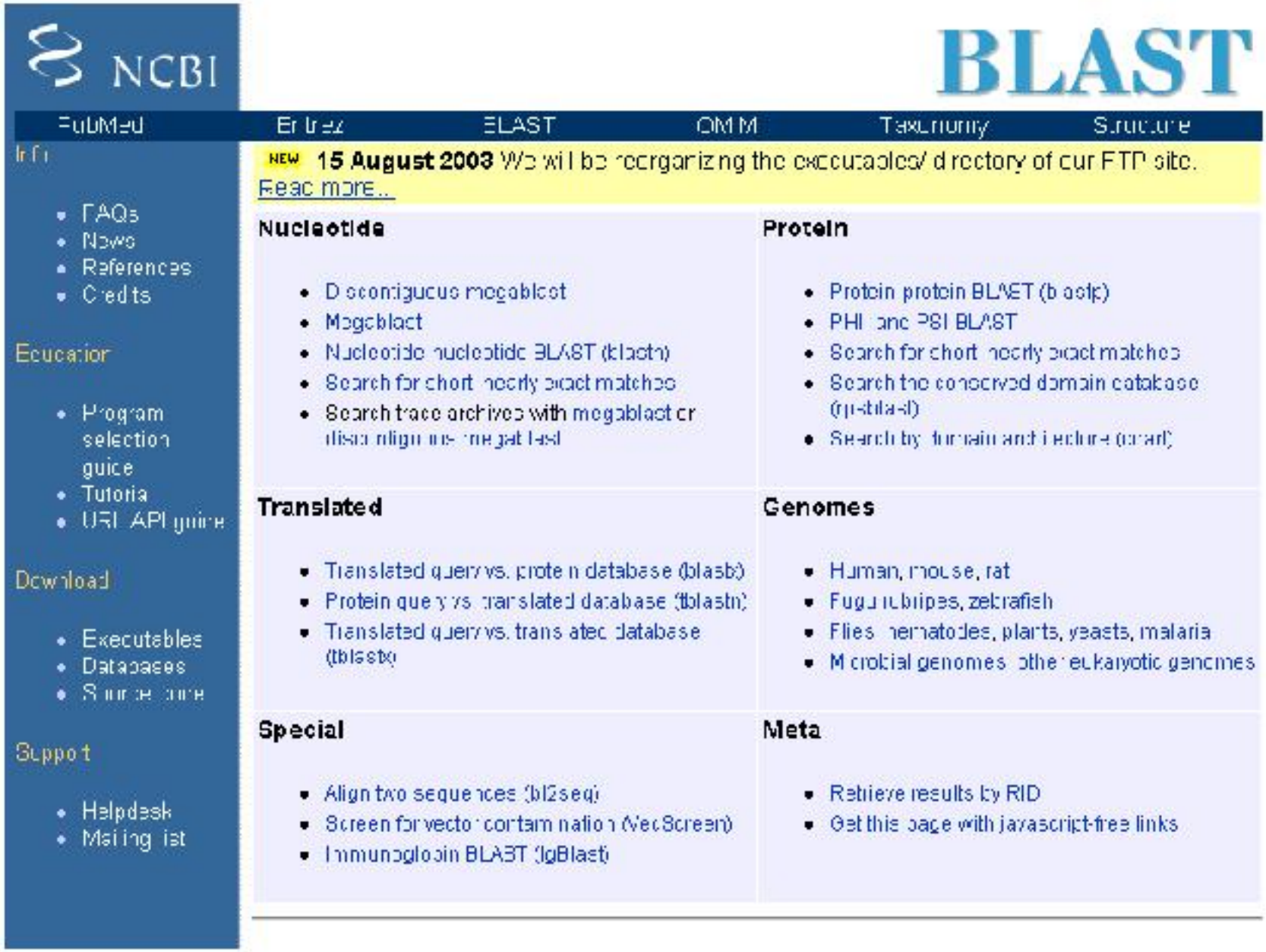
Основная идея – поиск коротких полностью совпадающих фрагментов и расширение выравнивания

Поиск по базам данных нуклеотидных и белковых последовательностей с помощью программы BLAST

Основная задача, которая решается с помощью программы **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool) - быстрый поиск сходных (гомологичных) участков последовательностей в банках данных нуклеотидных или аминокислотных последовательностей. **BLAST** не является программой выравнивания.

Входными данными для программы **BLAST** является нуклеотидная или аминокислотная последовательность, для которой производится поиск сходных последовательностей среди всех последовательностей из банков данных. Выходными данными являются последовательности из банков данных, которые имеют участки статистически значимого сходства с тестируемой последовательностью.




NCBI
BLAST
PubMed
Entrez
BLAST
OMIM
Taxonomy
Structure
Info
FAQs
News
References
Credits
Education
Program selection guide
Tutorials
URL API guide
Download
Executables
Databases
Source code
Support
Helpdesk
Mailing list
NEW 15 August 2003 We will be reorganizing the executables/ directory of our FTP site.
Read more...
Nucleotide
Discontiguous megablast
Megablast
Nucleotide-nucleotide BLAST (blastn)
Search for short, nearly exact matches
Search trace archives with megablast or
discontiguous megablast
Protein
Protein-protein BLAST (blastp)
PHI and PSI BLAST
Search for short, nearly exact matches
Search the conserved domain database
(rpsblast)
Search by domain architecture (cobalt)
Translated
Translated query vs. protein database (blastx)
Protein query vs. translated database (tblastn)
Translated query vs. translated database
(tblastx)
Genomes
Human, mouse, rat
Fugu, ruberipes, zebrafish
Flies, nematodes, plants, yeasts, malaria
Microbial genomes, other eukaryotic genomes
Special
Align two sequences (bl2seq)
Screen for vector contamination (VecScreen)
Immunoglobin BLAST (IgBlast)
Meta
Retrieve results by RID
Get this page with javascript-free links

Основные задачи компьютерного анализа генетических текстов:

2) статистический анализ генетических текстов, исследование структуры повторов и модели порождения символьных последовательностей, сегментация геномов

Статистическая обработка последовательностей

(частоты нуклеотидов и аминокислот, олигонуклеотидов, *l*-граммы)

Поиск паттернов

Поиск консервативных компонент в наборе последовательностей.

PROFILE

PSSM (Position Specific Score Matrix)

Stormo, Staden

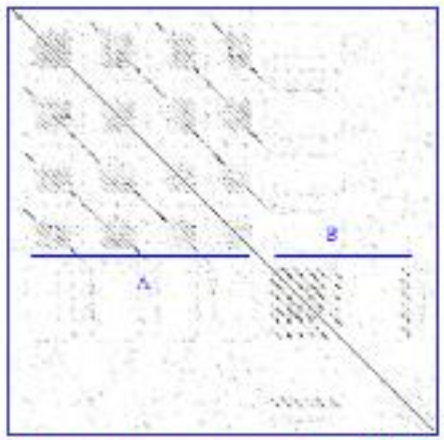
Нейронные сети , скрытые марковские модели, максимизация ожидания, выравнивание по Гиббсу.

Статистический анализ...

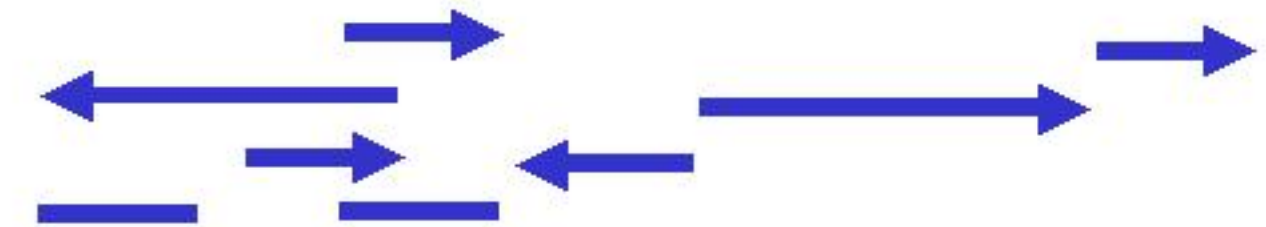
Классификация повторов

Название	Пример	Направление	Комплементарность
Прямой	AGCTTT … AGCTTT TCGAAA … TCGAAA	Вперед	Нет
Инвертированный	AGCTTT … AAAGCT TCGAAA … TTTCGA	Назад	Есть
Симметричный	AGCTTT … TTTCGA TCGAAA … AAAGCT	Назад	Нет
Прямой комплементарный	AGCTTT … TCGAAA TCGAAA … AGCTTT	Вперед	Есть
Палиндром	AAGCCGAA TTCGGCTT	Назад	Нет
Комплементарный палиндром	AAGCGCTT TTCGCGAA	Назад	Есть

Повторы могут пересекаться и накладываться друг на друга в последовательности.



Тандемные и диспергированные повторы



Последовательность ДНК: GTAGTCTGATGCA

Прямой

Инвертированный

Симметричный

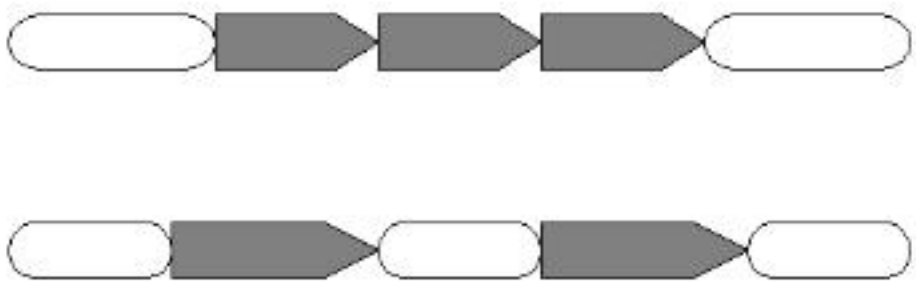
Комплементарный

Повторы могут пересекаться и накладываться друг на друга в последовательности.

Повторы могут иметь длину от нескольких пар до нескольких тысяч пар оснований.

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.

Тандемные и диспергированные повторы



Вопросы из комбинаторики:

Определить вероятность получить последовательность длины l, вероятность получить повтор длины l в последовательности длины n

Статистический анализ.
Модели порождения последовательностей

Анализ закодированных в последовательностях ДНК функциональных сигналов требует применения современных методов распознавания образов, статистических подходов и специальных оптимизированных алгоритмов для преодоления вычислительных трудностей, связанных с обработкой огромных массивов информации.

Вероятностные предсказания функции ДНК возможны лишь при достаточном объеме известных экспериментальных данных, накопленных в специализированных базах данных и при адекватном выборе моделей, описывающих выполнение функции. Надежность предсказания зависит от степени полноты априорной информации (наличия и полноты обучающей выборки, эволюционной близости анализируемых последовательностей).

P(ATGCGG)=?

P или *Pr* – probability, вероятность

Модели порождения последовательности:

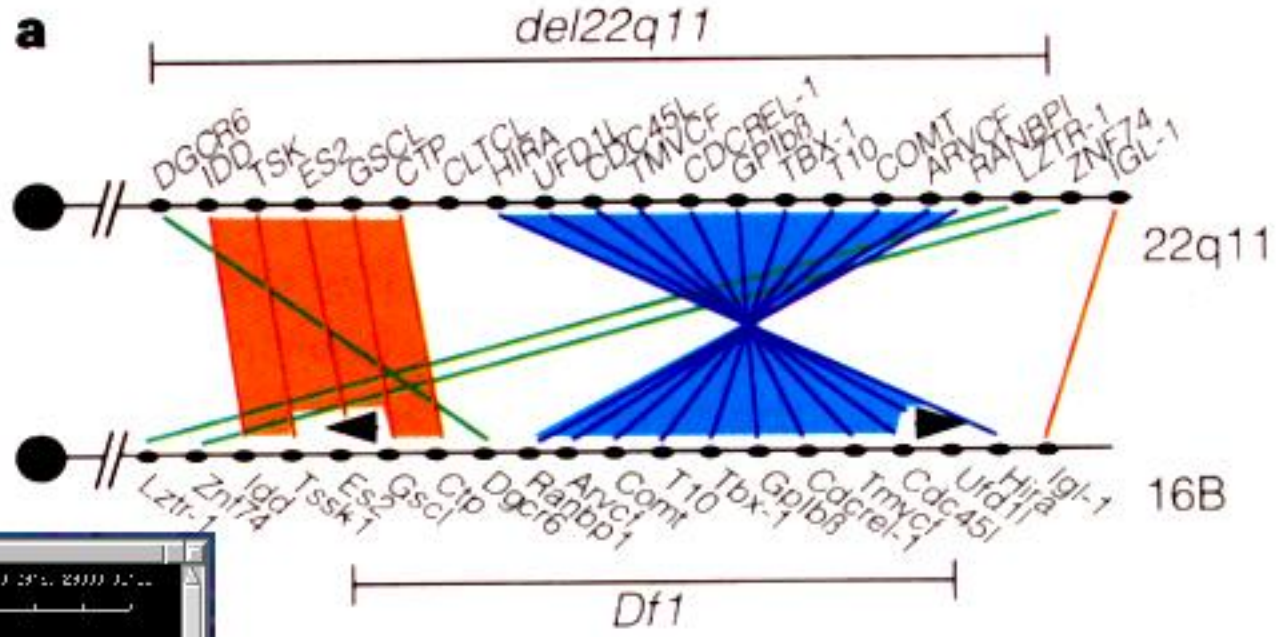
Равновероятная модель: $P(ATGG)=(0.25)^4$

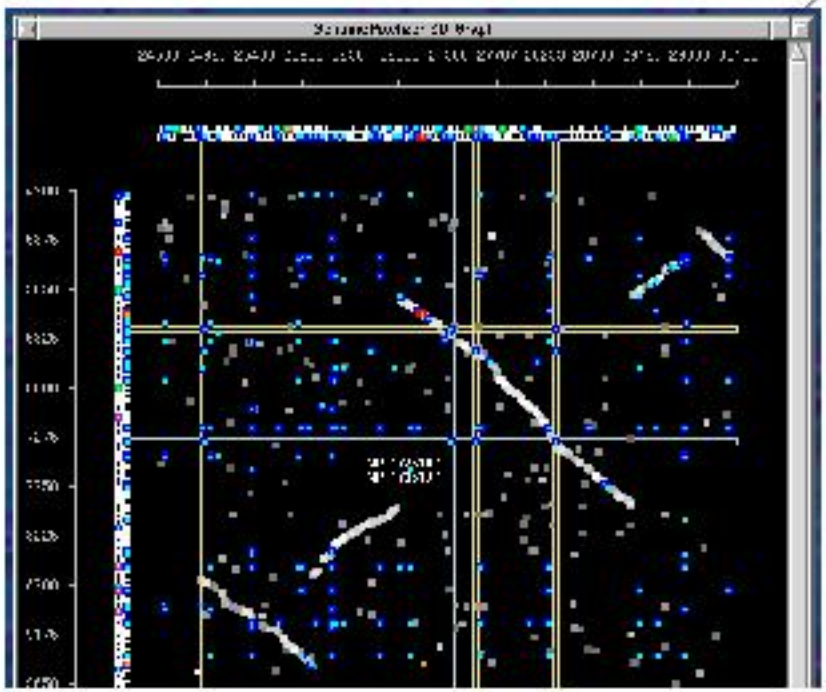
Модель Бернулли: P(ATGG)=P(A)*P(T)*P(G) *P(G)

Марковская модель 2-го порядка:

P(ATGG)=P(A)*P(T|A)*P(G|T) *P(G|G)

Сравнительный анализ района 22 хромосомы человека и 16 хромосомы мыши

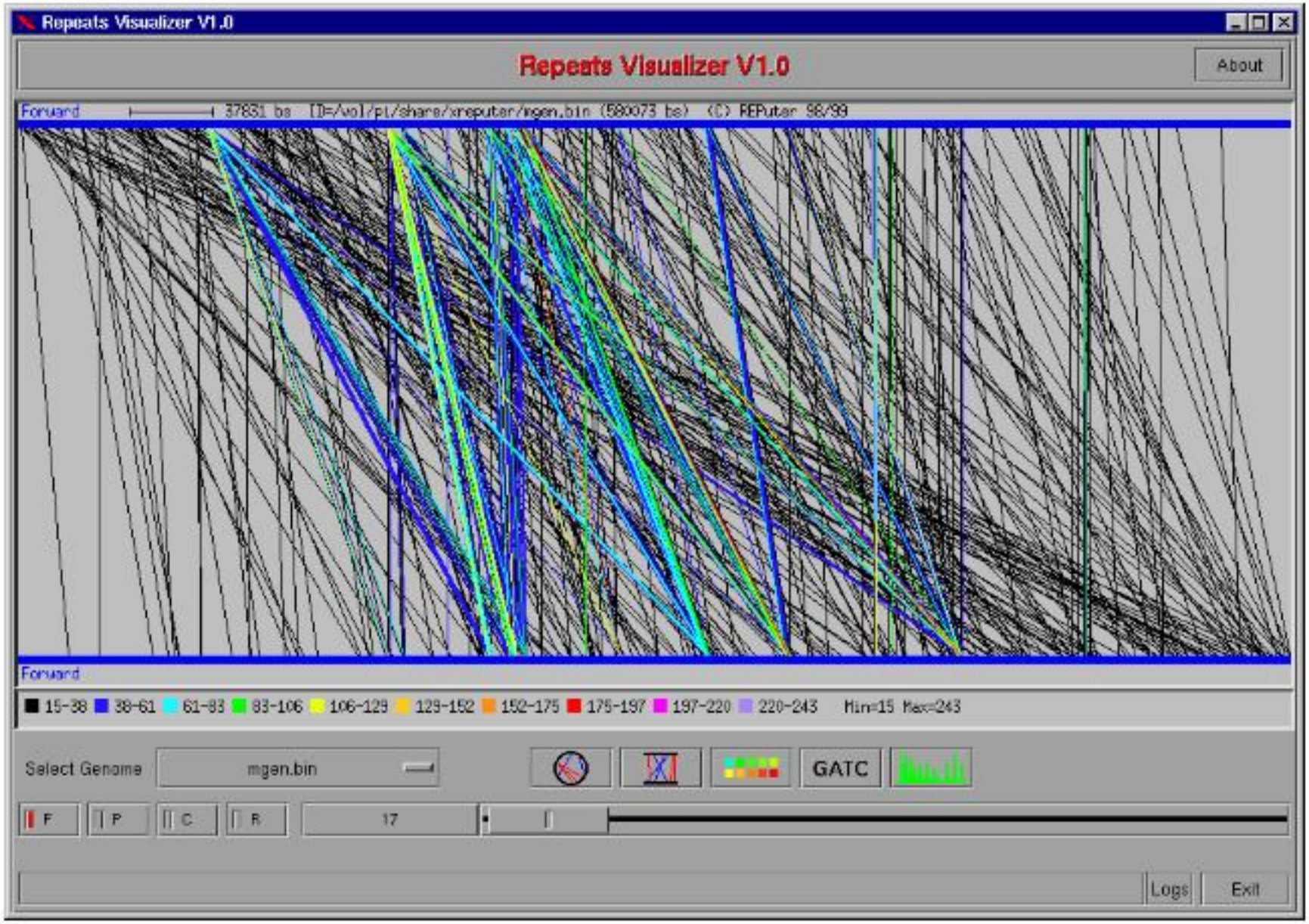




Сравнение порядка расположения генов может представлено в виде дот-матрицы

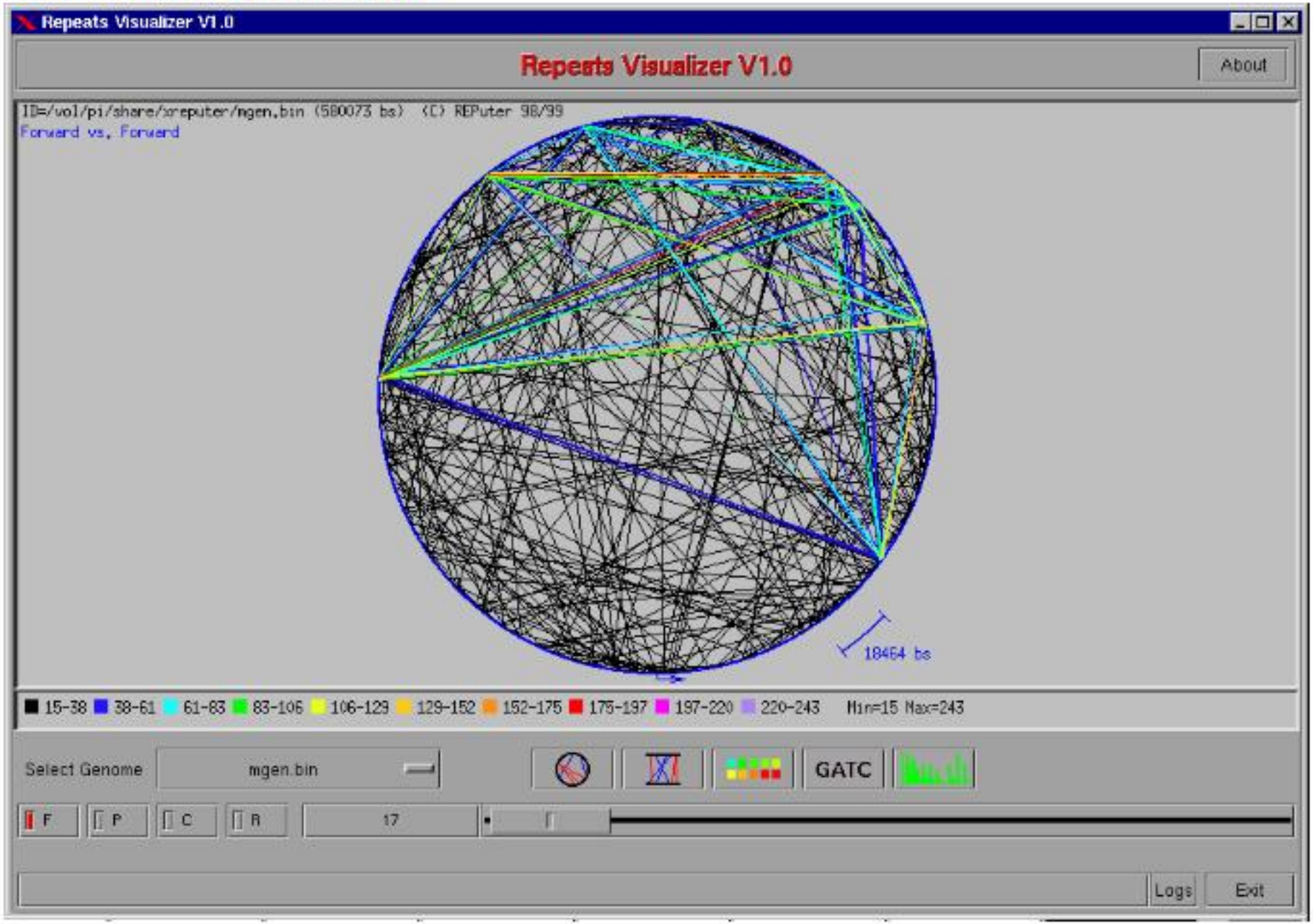
http://atgc.org/GenoPix_2D_Plotter/

Статистический анализ.
Сравнение и сегментация геномов



Визуализация сравнения полных геномов с помощью программы REPuter (http://bibiserv.techfak.uni-bielefeld.de/reputer/)

Статистический анализ.
Сравнение и сегментация геномов



Визуализация сравнения полных геномов с помощью программы REPuter (http://bibiserv.techfak.uni-bielefeld.de/reputer/)

Основные задачи компьютерного анализа генетических текстов:

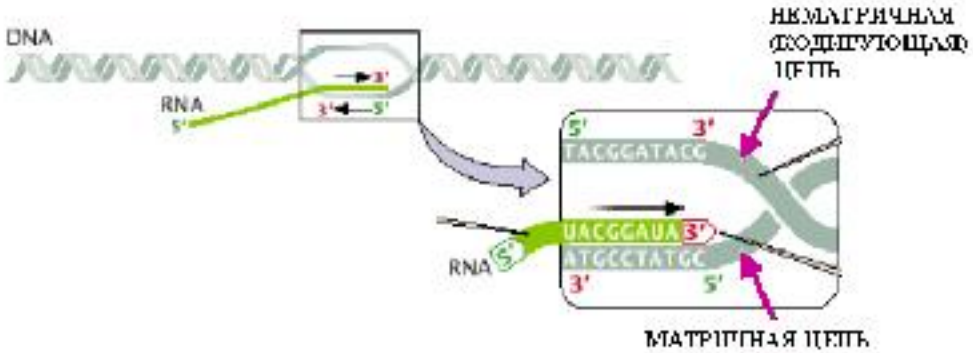
3) предсказание кодирующих участков генов и открытых рамок считывания;

Бактериальный ген (прокариоты)

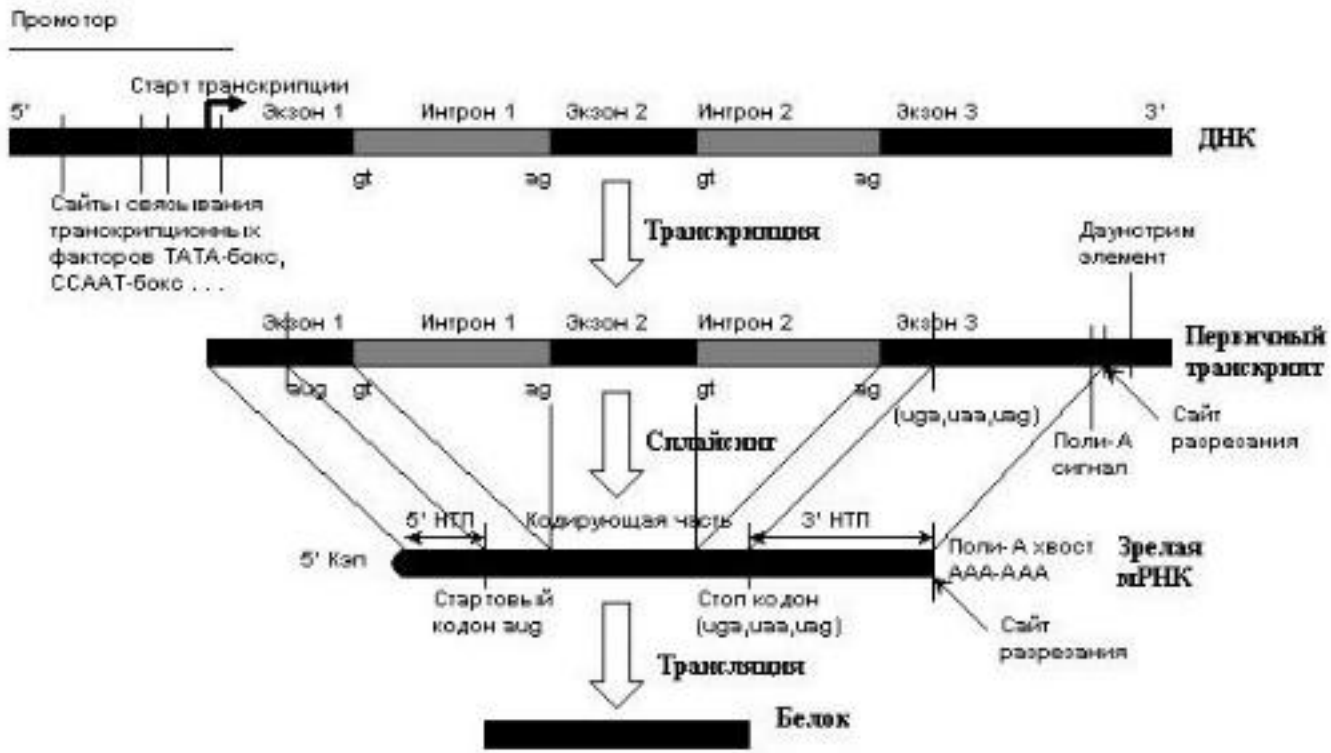
Задача предсказания структуры гена остается одной из важнейших задач биоинформатики, несмотря на полное секвенирование геномов и новые экспериментальные методики

Перекодирование последовательности в аминокислотную

Кодирующий потенциал – частоты гексануклеотидов



Предсказание кодирующих участков. Гены эукариот



ПРИНЦИПЫ работы программ распознавания структуры гена:

Кодирующий потенциал

Информация о сайтах сплайсинга

Скрытые марковские модели

Сравнение с известными генами по базам данных

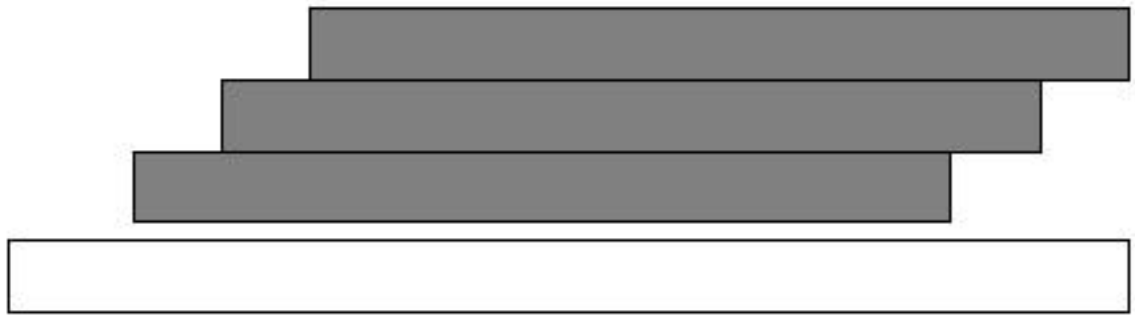
Экспертные системы

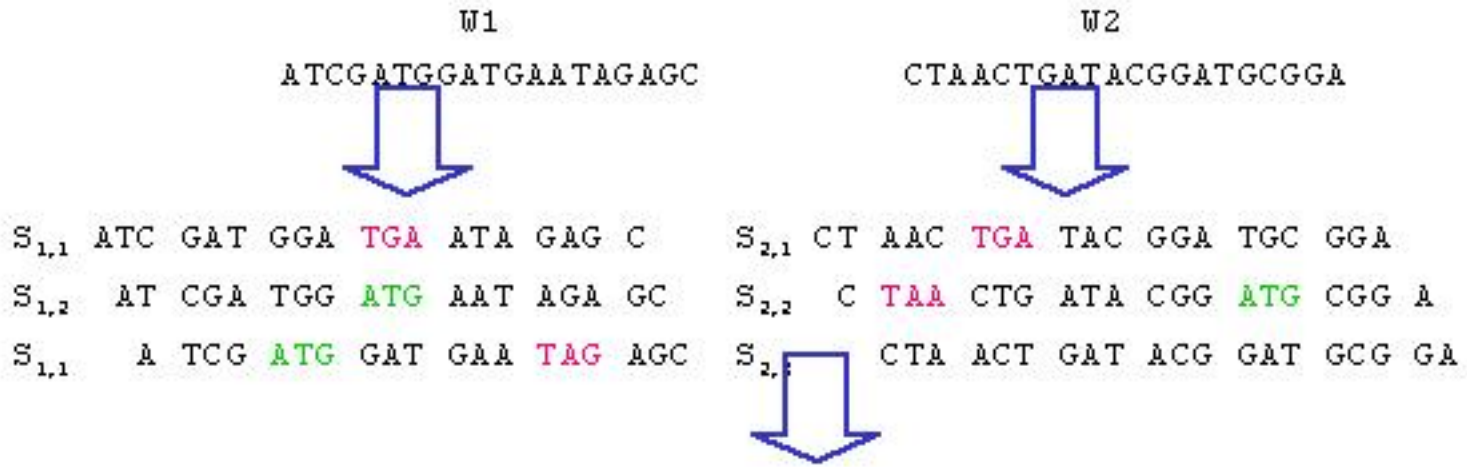
Предсказание кодирующих участков. Гены эукариот

Понятие рамки считывания ORF.

Сдвиг рамки считывания

Статистическая задача предсказания экзонов





Кодирующий потенциал

(частоты использования кодонов)

Определение маршрута, начинающегося с AUG и заканчивающегося стоп-кодоном или концом последовательности

Предсказание кодирующих частей генов с помощью программ ORFScan, ESTScan, GENSCAN, GRAIL, GENFIND на примере DT_101886.

Ген human ferrochelatase. Ошибочно определенные позиции маркированы цветом

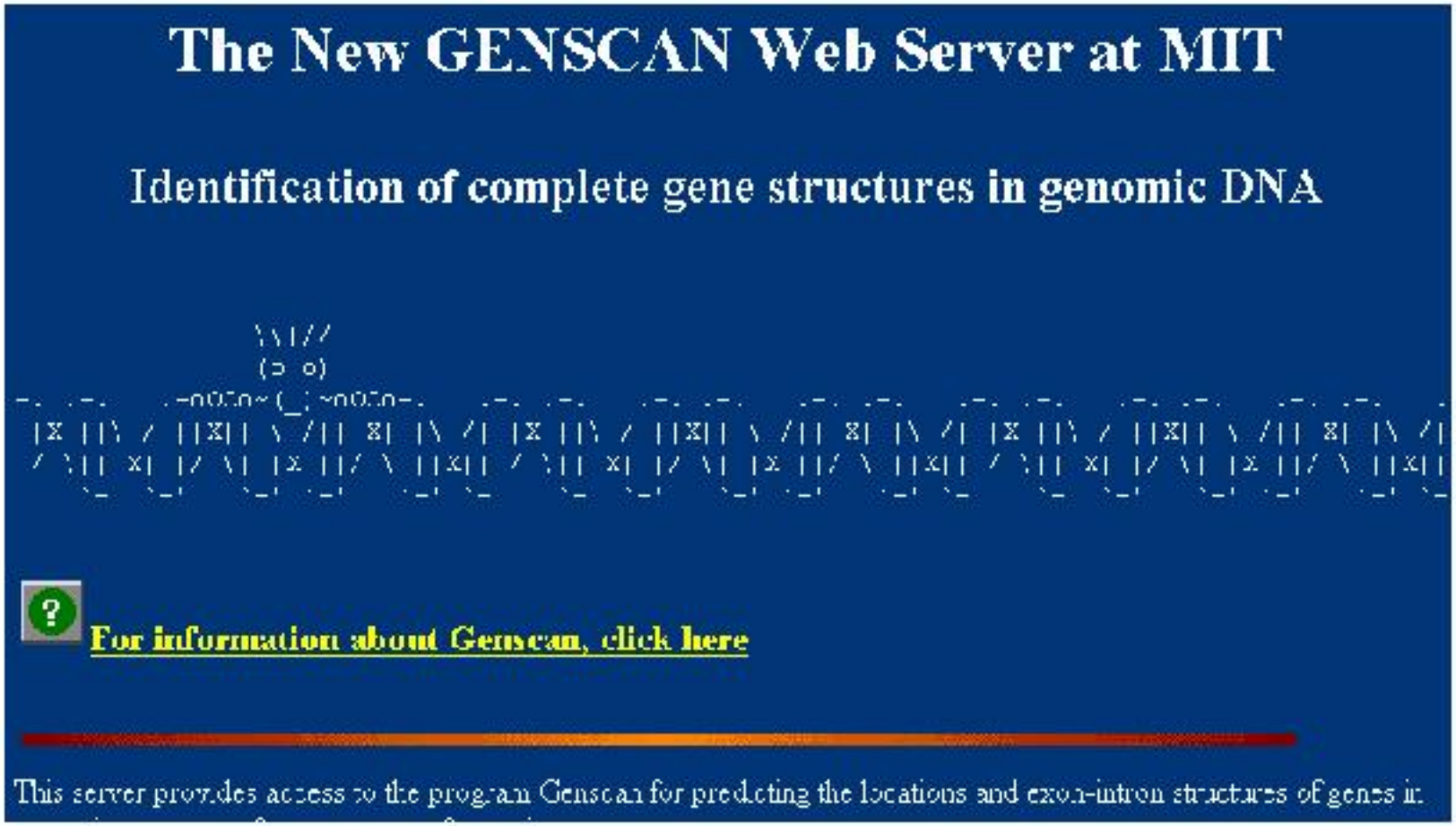
```
true     : mrslgarmaaalraagvllrdplassswrvcqpwrwksgaaaaavttetaqhaqgakpqvqpqkrkpktgilmlrmggpetlgdvhdfllrlfldqdlmtlpiqrklapf
ORFScan  : MRSLGAMMAAALRAAGVLLRDPLASSSWRVCQPWRWKSGAAAAAVTTETAQHAQGAKPQVQPQKRKPKTGILMLRMGGPETLGDVHDFLLRLFLDQDLMTLPIQRKLAPF
ESTScan  : NAFIGAMMAAALRAAGVLLRXSAGIQQLEGLSAMEVEVGAAAAAVTTETAQHAQGAKPQVQPQKRKPKTGILMLRMGGPETLGDVHDFLLRLFLDQDLMTLPIQRKLAPF
GENSCAN  : _______MAAALRAAGVLLRDPLASSSWRVCQPWRWKSGAAAAAVTTETAQHAQGAKPQVQPQKRKPKTGILMLRMGGPETLGDVHDFLLRLFLDQDLMTLPIQRKLAPF
GRAIL    : MRSLGAMMAAALRAAGVLLRDPLASSSWRVCQPWRWKSGAAAAAVTTETAQHAQGAKPQVQPQKRKPKTGILMLRMGGPETLGDVHDFLLRLFLDQDLMTLPIQRKLAPF
GenFind  : _______MAAALRAAGVLLRDPLASSSWRVCQPWRWKSGAAAAAVTTETAQHAQGAKPQVQPQKRKPKTGILMLRMGGPETLGDVHDFLLRLFLDQDLMTLPIQRKLAPF

true     : iakrrtpkiqeqyrrigggspikiwtskqgegmvklldelspntaphkyyigfryvhplteeaieemerdgleraiaftqypqyscsttgsslnaiyryynqvgrkptmk
ORFScan  : IAKRRTPKIQEQYRRLEADPPSRY??SKQGEGMVKLLDELSPNTAPHKYYIGFRYVHPLTEEAIEEMERDGLERAIAFTQYPQYSCSTTGSSLNAIYRYYNQVGRKPTMK
ESTScan  : IAKRRTPKIQEQYRRXWRRIPHQIMDFQQGEGMVKLLDELSPNTAPHKYYIGFRYVHPLTEEAIEEMERDGLERAIAFTQYPQYSCSTTGSSLNAIYRYYNQVGRKPTMK
GENSCAN  : IAKRRTPKIQEQYRRXWRRIPHQIMDFQAGRGH__________APHKYYIGFRYVHPLTEEAIEEMERDGLERAIAFTQYPQYSCSTTGSSLNAIYRYYNQVGRKPTMK
GRAIL    : IAKRRTPKIQEQYRRTWRRIPHQIMDFQAGRGH_____________________________________________________________________
GenFind  : IAKRRTPKIQEQYRRTWRRIPHQIMDFQAGRGH_____________________EEAIEEMERDGLERAIAFTQYPQYSCSTTGSSLNAIYRYYNQVGRKPTMK

true     : wstidrwpthhlliqcfadhilkeldhfplekrsevvilfsahslpmsvvnrgdpypqevsatvqkvmerleycnpyrlvwqskvgpmpwlgpqtdesikglcergrkni
ORFScan  : WSTIDRWPTHHLLIQCFADHILKELDHFPLEKRSEVVILFSAHSLPMSVVNRGDPYPQEVSATVQKVMERLEYCNPYRLVWQSKVGPMPWLGPQTDESIKGLCERGRKNI
ESTScan  : WSTIDRWPTHHLLIQCFADHILKELDHFPLEKRSEVVILFSAHSLPMSVVNRGDPYPQEVSATVQKVMERLEYCNPYRLVWQSKVGPMPWLGPQTDESIKGLCERGRKNI
GENSCAN  : WSTIDRWPTHHLLIQCFADHILKELDHFPLEKRSEVVILFSAHSLPMSVVNRGDPYPQEVSATVQKVMERLEYCNPYRLVWQSKVGPMPWLGPQTDESIKGLCERGRKNI
GRAIL    : _____________________________________________GEAAGQEVSATVQKVMERLEYCNPYRLVWQSKVGPMPWLGP_____________________
GenFind  : WSTIDRWPTHHLLIQCFADHILKELDHFPLEKRSEVVILFSAHSLPMSVVNRGDPYPQEVSATVQKVMERLEYCNPYRLVWQSKVGPMPWLGPQTDESIKGLCERGRKNI

true     : llvpiaftsdhietlyeldieysqvlakecgvenirraeslngnplfskaladlvhshiqsnelcskqltlscplcvnpvcretksfftsqql
ORFScan  : LLVPIAFTSDHIETLYELDIEYSQVLAKECGVENIRRAESLNGIHCS?KALADLV?FTHPVKRAVFQAADPDCPLCVNPVCRETKSFFT?PAAVTPAGGPRGVSKCPTSRYLRCGEGVI
ESTScan  : LLVPIAFTSDHIETLYELDIEYSQVLAKDVGVENIXKELSLFNGKFHCSLKAPGRLGAFTHPSQTSCVPKQLTPELVNPVCRETKSFXHPASSCEPPPVDPVALANAQPPDTSDVER
GENSCAN  : LLVPIAFTSDHIETLYELDIEYSQVLAKECGS_____________________________________________________________________
GRAIL    :                                                                                  CPLCVNPVCRETKSFFTQPAAVNPRRVTPVR
```

Пример программы предсказания структуры гена GenScan



http://genes.mit.edu/GENSCAN.html

Предсказание кодирующих участков. Гены эукариот

1	+	3434	3681	0	Initial	248		
2	+	4405	4462	0	Initial	58		
3	+	7478	7517	2	Terminal	40		
4	-	13450	13528	2	Terminal	79		
5	-	13709	13807	0	Terminal	99		
6	-	13876	13951	0	Internal	76		
7	-	16097	16308	0	Initial	212	77	Marginal
8	+	19541	19632	0	Initial	92	95	Excellent
9	+	19755	19977	2	Internal	223	91	Excellent
10	+	20833	20951	0	Internal	119	93	Excellent
11	+	22124	22173	1	Terminal	50	61	Marginal
12	+	25777	25798	2	Terminal	22	93	Excellent
13	+	34531	34622	0	Initial	92	96	Excellent
14	+	34745	34982	2	Internal	238	100	Excellent
15	+	35854	35982	0	Terminal	129	89	Good
16	+	39467	39558	0	Initial	92	96	Excellent
17	+	39681	39918	2	Internal	238	100	Excellent
18	+	40770	40856	0	Internal	87	93	Excellent
19	+	41914	42021	0	Internal	108	76	Good
20	+	45653	45753	0	Initial	101	63	Marginal
21	+	45995	46210	0	Single	216	90	Excellent
22	+	47254	47417	0	Initial	164	64	Marginal
23	+	54790	54881	0	Initial	92	97	Excellent
24	+	55010	55232	2	Internal	223	100	Excellent
25	+	56131	56259	0	Terminal	129	93	Excellent
26	+	62187	62278	0	Initial	92	96	Excellent
27	+	62409	62631	2	Internal	223	100	Excellent
28	+	63482	63610	0	Terminal	129	94	Excellent

```
FT   CDS  join
(19541..19632,19755..19977,20833..20961)
FT                    /note="epsilon-globin"
FT   CDS  join
(34531..34622,34745..34967,35854..35982)
FT                    /note="G-gamma globin"
FT   CDS join(39467..39558,39681..39903,40770..40898)
FT                    /note="A-gamma globin"
FT   CDS join(54790..54881,55010..55232,56131..56259)
FT                    /note="delta-globin"
FT   CDS join(62187..62278,62409..62631,63482..63610)
FT                    /note="beta-globin"
FT   CDS join(62187..62278,62390..62408)
FT                    /note="beta-globin thalassemia"
```

Результаты предсказания кластера глобиновых генов человека (73308 по) с помощью программы GrailEXP v3.31 [March, 2002]

http://grail.lsd.ornl.gov/grailexp/

Предсказание кодирующих участков. Гены эукариот

Название программы, ссылкаИнтернет-адрес, краткое описание

GENEID (Wiehe et al., 2001) http://www1.imim.es/geneid.html (R. Guigo, Spain Institut Municipal de Investigacio Medica, Испания)

SLAM(Pachter et al., 2002) http://bio.math.berkeley.edu/slam/ (Марковские модели и парное выравнивание)

GENIE (Reese et al., 2000) http://www.cse.ucsc.edu/~dkulp/cgi-bin/genie (Скрытые марковские модели)

SELFID (Audic and Claverie, 1998) http://igs-server.cnrs-mrs.fr/~audic/selfid.html (Франция, поиск генов в микробной ДНК)

MZEF(Zhang, 1997) http://argon.cshl.edu/genefinder/ (Cold Spring Harbor Lab, США, Квадратичный дискриминантный анализ)

WEBGENE (Milanesi et al., 1999) http://www.itba.mi.cnr.it/webgene/ (ITBA, CNR, Milan, Italy)

GeneMark(Lukashin and Borodovsky, 1998) http://opal.biology.gatech.edu/GeneMark/ (GIT, Borodovsky's lab, School of Biology, США, Скрытые марковские модели)

FrameD(Schiex et al., 2000) http://www.toulouse.inra.fr/FrameD/cgi-bin/FD (INRA, Toulouse, Франция) поиск генов и рамок считывания в G+C богатых прокариотических последовательностях

EuGene(Schiex et al., 2001) http://www-bia.inra.fr/T/EuGene/ поиск генов Arabidopsis thaliana

GLIMMER(Delcher et al., 1999) http://www.tigr.org/~salzberg/glimmer.html (TIGR, Salzberg's lab) поиск генов в ДНК микробов

VEIL(Henderson et al., 1997) http://www.tigr.org/~salzberg/veil.html (VEIL - the Viterbi Exon-Intron Locator. Скрытая марковская модель)

MORGAN(Salzberg et al., 1998) http://www.tigr.org/~salzberg/morgan.html (Решающие деревья для поиска генов в ДНК позвоночных)

GENESCAN(Tiwari et al., 1997) http://202.41.10.146/ (Jawaharlal Nehru Univ., Индия. Поиск генов с использованием преобразований Фурье)

GENSCAN http://genes.mit.edu/GENSCAN.html (C. Burge, Massachusetts Institute of Technology, США)

Diogenes http://www.cbc.umn.edu/diogenes/index.html (США. Предсказание для коротких последовательностей)

GRAIL(Uberbacher et al., 1996) http://compbio.ornl.gov/Grail-bin/EmptyGrailForm (Oak Ridge National Lab, США)

FGENES (Solovyev and Salamov, 1997) http://genomic.sanger.ac.uk/gf/gf.html (Sanger Centre, UK) http://www.softberry.com/berry.phtml (новая версия FGENES)

HMMGENE(Krogh, 1997) http://www.cbs.dtu.dk/services/HMMgene/ (Technical Univ. of Denmark, Дания. Скрытые марковские модели)

YEASTGENE(Zhang and Wang, 2000) http://tubic.tju.edu.cn/cgi-bin/Yeastgene.cgi (TianJin University, Китай, техника распознавания Z-curve)

GENEPARSER(Snyder and Stormo, 1995) http://beagle.colorado.edu/~eesnyder/GeneParser.html (Динамическое программирование и нейронные сети)

Вопрос компьютерного предсказания генов, экзон-интронной структуры по нуклеотидной последовательности остается открытым.

В настоящее время даже число генов в полностью секвенированном геноме человека остается неизвестным – разные компьютерные методики дают оценку от минимум **24500** до **45000** генов. Минимальную оценку дают сравнения с кодирующими частями родственных геномов (например мыши), максимальную – предсказание по свойствам самой последовательности ДНК (GENSCAN).

Эти цифры сильно уменьшены по сравнению с первоначальными оценками в 100 тысяч генов, затем 50 тысяч генов.

Есть проблемы в определении понятия гена – как считать варианты альтернативного сплайсинга, низкоэкспрессирующиеся гены, короткие гены (один экзон) и гены, кодирующие только РНК, но не белок.

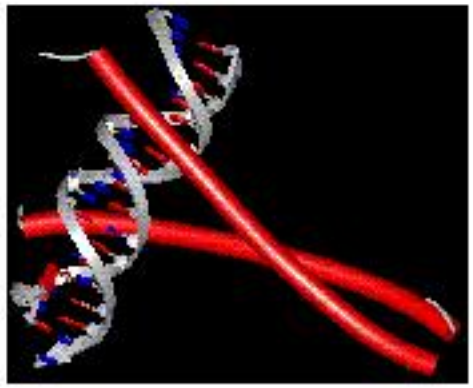
Проблемой остается предсказание генов на так называемой «темной стороне» (dark side) генома человека – районов с низкой плотностью генов. Существующие программы предсказания генов не могут адекватно находить гены в этих частях генома.

Elizabeth Pennisi. (2003) Gene Counters Struggle to Get the Right Answer. *Science* Aug 22 : 1040-1041.

Основные задачи компьютерного анализа генетических текстов:

4) предсказание функциональных сигналов (функциональных сайтов и регуляторных районов)

Понятие консенсуса, весовой матрицы, паттерна



консенсус ТАТА-бокс связывающего белка (TBP) имеет вид TATAAAAA, расширенный аналог в вырожденном 15-буквенном коде имеет вид STWTAWADRSSSSSS

Соответствие представлений функц.сайта в 4-буквенном и 15-буквенном алфавитах

```
---TATAAAAA----

STWTAWADRSSSSSS
```

Более точным способом представления и анализа выборок выровненных последовательностей длины L являются весовые матрицы размерности $L \times 4$. Элемент f(i,j) весовой матрицы $F = |f(i, j)|$ определяет частоту встречаемости нуклеотида i (i =1,2,3,4 соответствует символам A, T, G и C) в позиции j ($j = 1,.., L$), подсчитанную по выборке выровненных нуклеотидных последовательностей. Оптимизированная весовая матрица $W = |w(i,j)|$ может быть вычислена в логарифмической форме с учетом ожидаемых частот

$$w(i, j) = \ln\left(\frac{f(i, j)}{e(i, j)} + \frac{s}{100}\right) + c(i)$$

Понятие весовой матрицы

A	5	0	0	3	18	0	3	4	2
T	4	1	1	0	9	2	1	1	5
G	7	3	2	1	13	22	19	15	17
C	27	39	40	39	3	19	20	23	19
	C	**C**	**C**	**C**	**A/G**	**G/C**	**G/C**	**G/C**	**G/C**
	C	**C**	**C**	**C**	**R**	**S**	**S**	**S**	**S**

R	**G/A**
Y	**T/C**
M	**A/C**
K	**G/T**
W	**A/T**
S	**G/C**
B	**–A**
V	**–T**
H	**–G**
D	**–C**
N	**A/T/G/C**

Пример весовой матрицы и консенсуса TATA-бокса в промоторах эукариот

- TATA-box HMM trained from 900 unrelated general promoter sequences:

Position	1	2	3	4	5	6	7	8	9	10	11	12
% A	21.4	15.9	3.7	91.1	0.0	94.5	67.3	97.3	52.1	40.7	16.5	23.6
% C	22.7	39.3	9.8	0.0	0.0	0.0	0.0	0.0	0.0	9.1	34.8	37.1
% G	28.2	35.2	2.9	0.0	0.0	0.0	0.0	2.7	12.0	40.2	38.0	30.4
% T	27.7	9.6	83.6	8.9	100.0	5.5	32.7	0.0	35.9	10.0	10.7	8.9
Consensus			T	A	T	A	W	A	W	R		

TATA-бокс в промоторах растений:

- TATA-box HMM trained from 134 unrelated plant promoter sequences:

Position	1	2	3	4	5	6	7	8	9	10	11	12
% A	31.6	16.3	2.0	90.8	0.0	94.9	57.1	100.0	27.6	69.4	11.2	24.5
% C	24.5	60.2	3.0	2.1	0.0	0.0	0.0	0.0	0.0	3.1	39.8	52.0
% G	15.3	10.2	0.0	2.0	1.0	0.0	0.0	0.0	2.0	13.3	37.8	21.4
% T	28.6	13.3	94.9	5.1	99.0	5.1	42.9	0.0	70.4	14.3	11.2	2.1
Consensus			T	A	T	A	W	A	W	A		

http://www.epd.isb-sib.ch/promoter_elements/

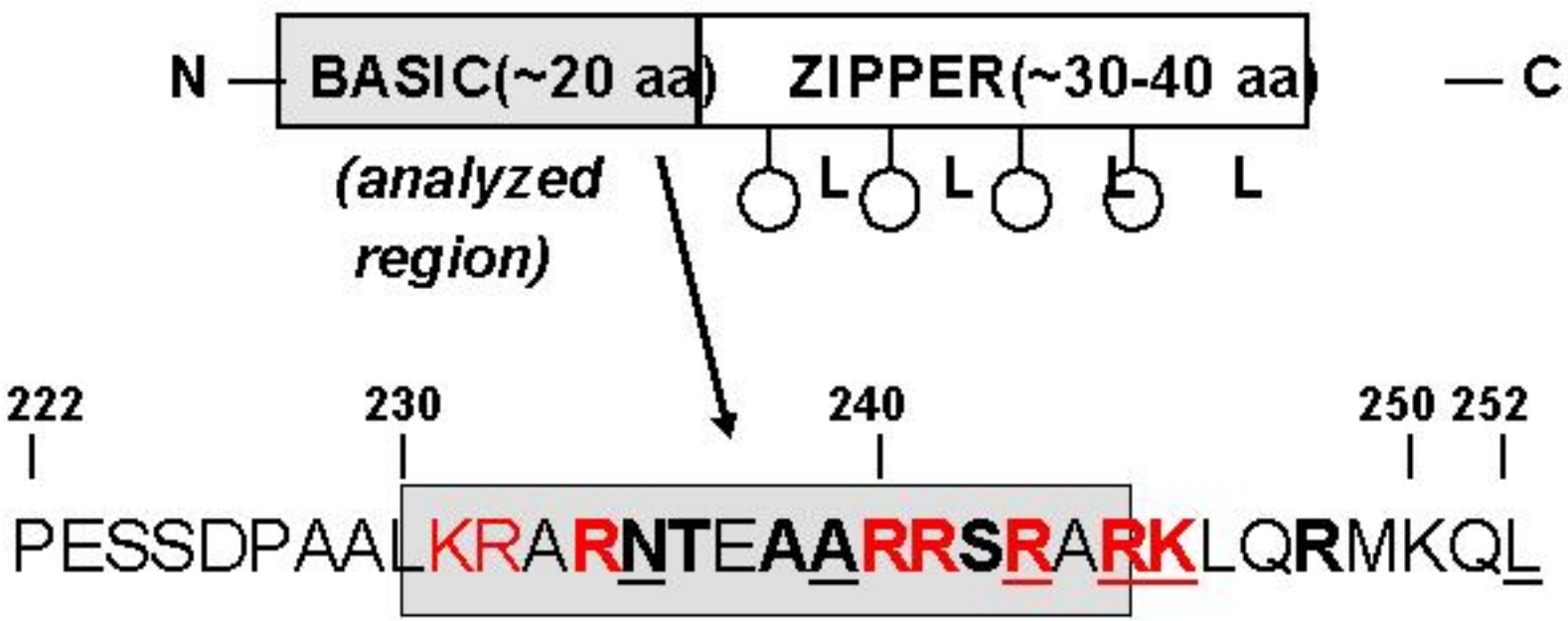
Понятие паттерна как вида записи функционального сайта:

$TAT(A)_{2-5}$

$R[Y]SA[G][T]G_{0-3}C_{0-6}$

структура домена bZIP

Функция: *связывание с ДНК* *формирование димера*



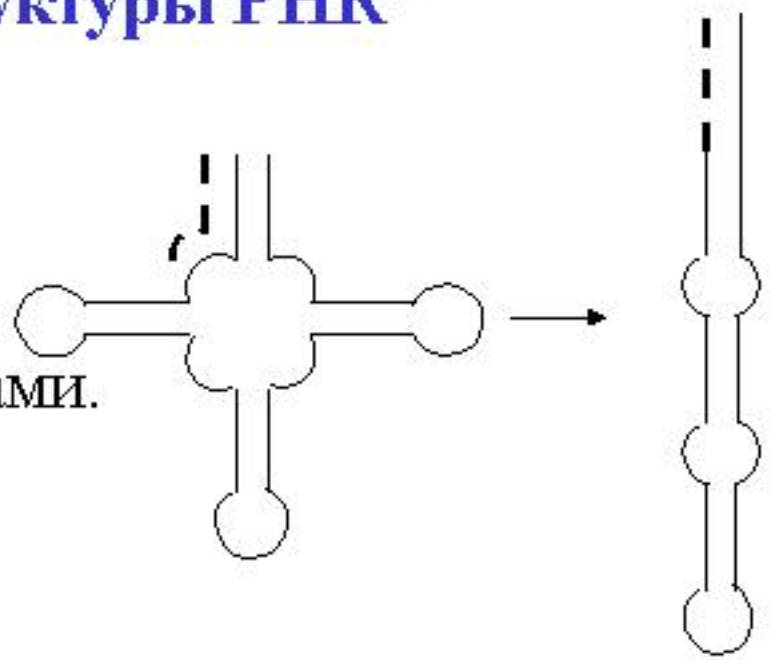
5) анализ вторичной структуры РНК и сигналов трансляции;

Предсказание вторичной структуры РНК

Шпилька РНК –

Оценка энергии по числу

водородных связей между нуклеотидами.



G-C (3H)

A-U (2H)

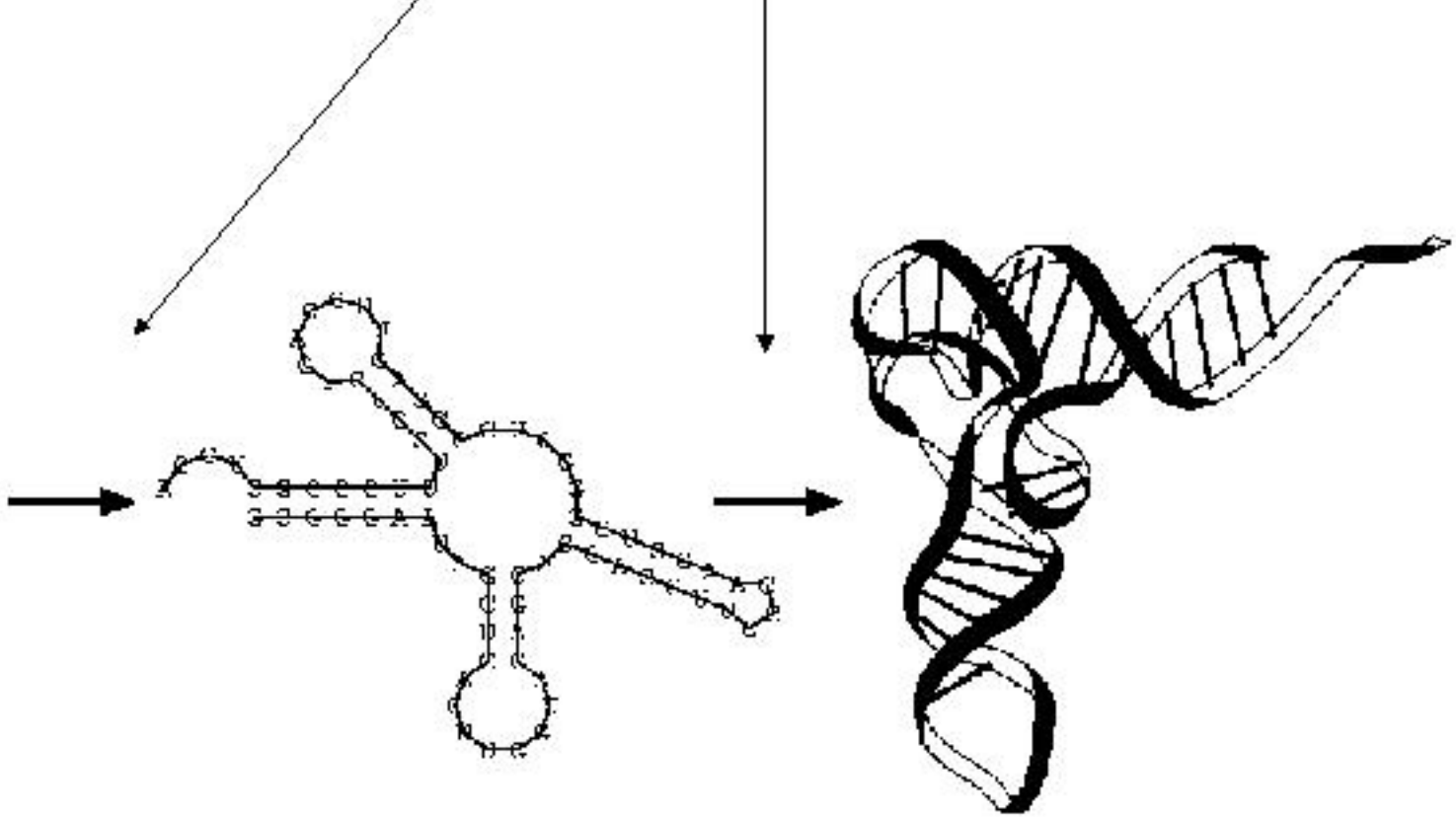
G-U (1H)

Задача – выбор оптимальной, наиболее энергетической вторичной структуры по анализу нуклеотидной последовательности

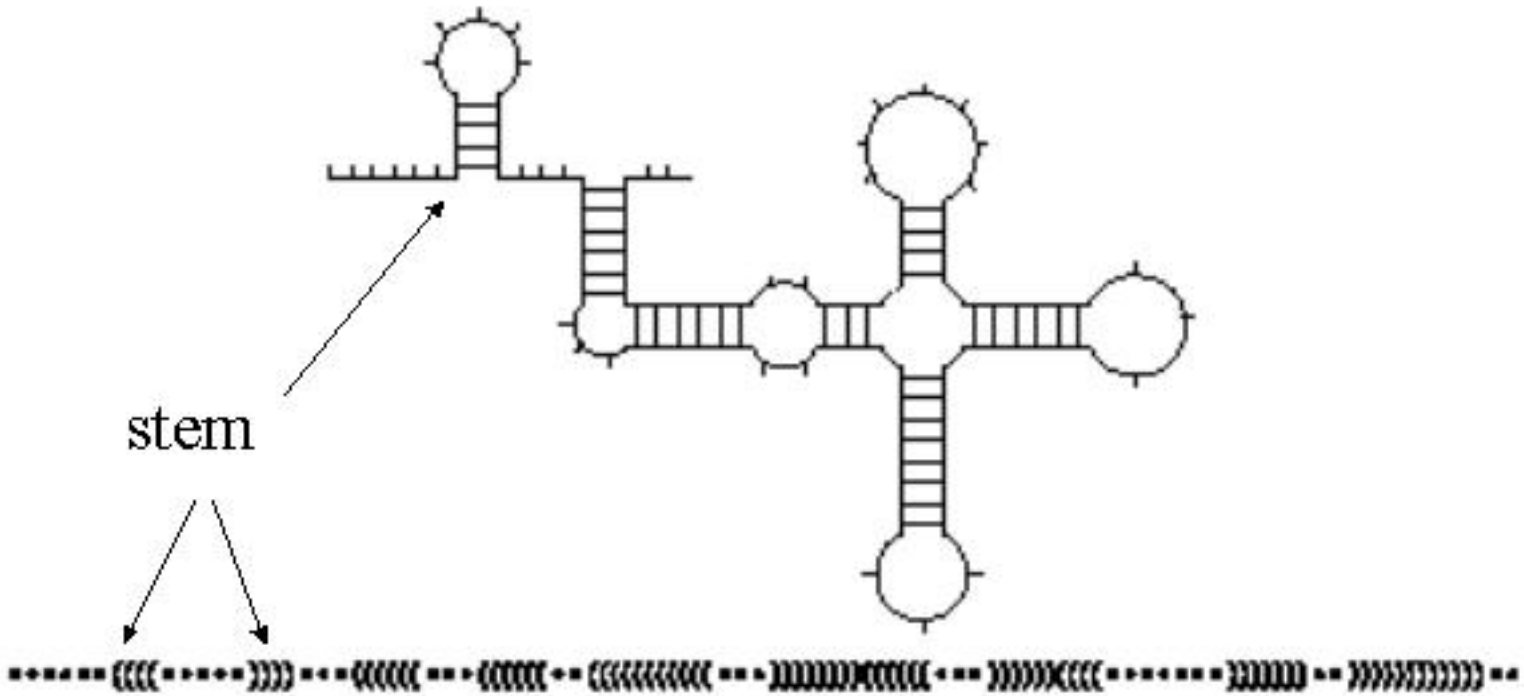
Nussinov and Jacobson, 1980; Zuker and Stiegler, 1981

ribosomal RNA database
http://www.cme.msu.edu/RDPhtml/index.html

Предсказание пространственной структуры

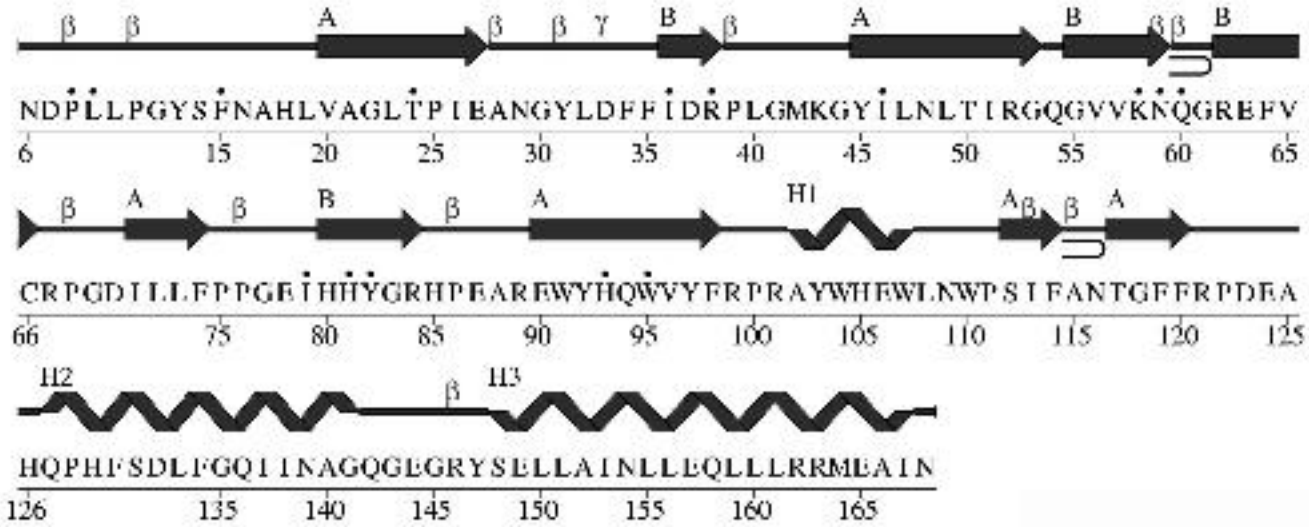


Представление (визуализация) вторичной структуры – матричное, линейное, графическое.



Линейное представление структуры – запись с помощью скобок (бракет)

б) анализ аминокислотных последовательностей белков, предсказание вторичной структуры, функциональных сайтов и доменов глобулярных белков по их аминокислотным последовательностям;



Уровни структурной организации белка на примере регуляторного белка arac E.coli (идентификатор PDB 2aac). На рисунке показаны: а) первичная структура белка – аминокислотная последовательность; и разметка вторичной структуры белка, выполненная программой PDBSum (Laskowski, 2001; *http://www.biochem.ucl.ac.uk/bsm/pdbsum/2aac/main.html*); б) пространственная структура димера, мономеры показаны разными оттенками серого цвета, элементы вторичной структуры показаны схематически в виде цилиндров (α-спирали) и стрелок (α-нитей).

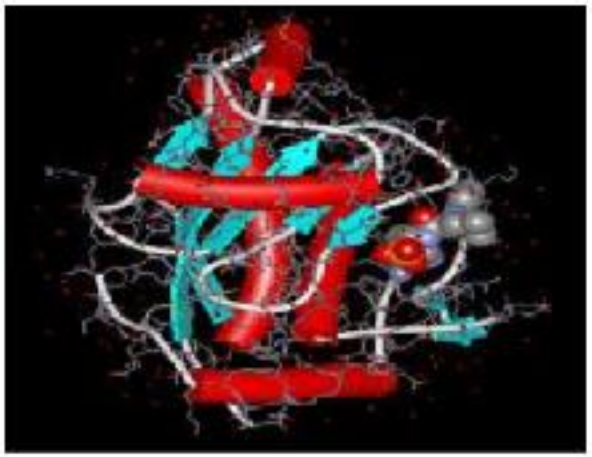
Построение последовательности белка по ДНК.

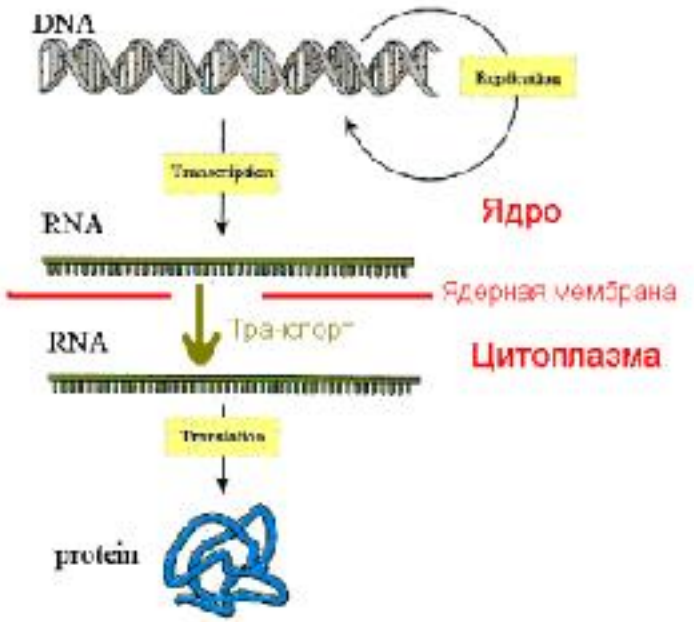
Предсказание вторичной структуры белка. Постановка задачи.

Полипептидная цепь –

Вторичная структура –

Трехмерная структура

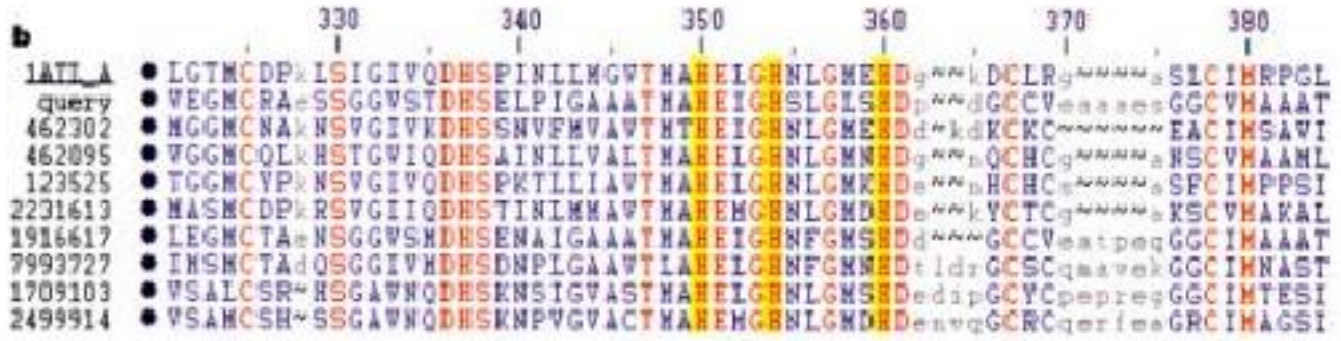




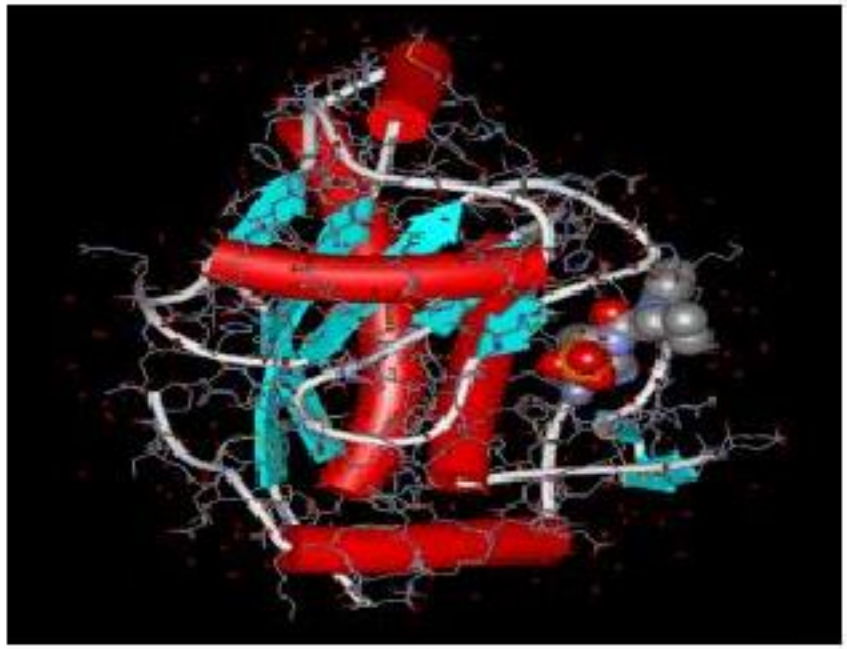
Адаптировано из:
URL Address: http://www.cytochemistry.net/Cell-biology/ribosome.htm
Gwen V. Childs, Ph.D.

Предсказание структуры белка

Расшифрованы аминокислотные последовательности сотен тысяч белков



Расшифрованы пространственные структуры десятков тысяч белков

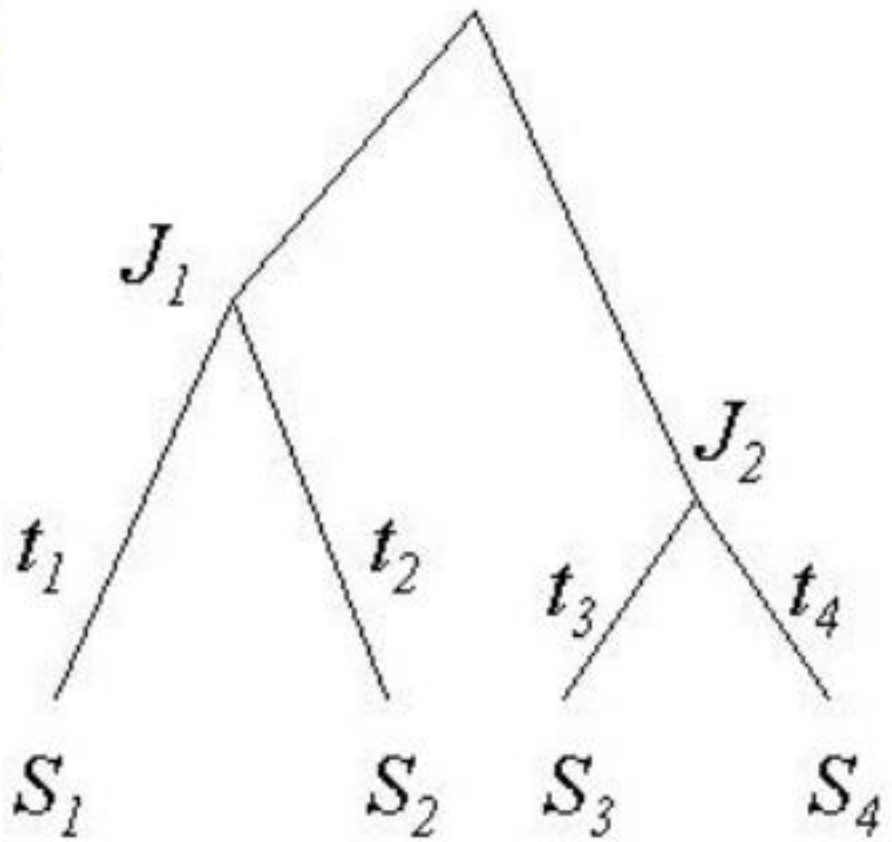


Задачи анализа белков будут рассмотрены в отдельном курсе к.б.н. Афонникова Д.А.

7) филогенетические сравнения

Рассмотрим задачу восстановления филогенетического построения дерева по набору последовательностей (ДНК или белков). В качестве исходных данных выступают современные последовательности S $(S_1, S_2 \ldots S_n)$. Как правило, рассматриваются последовательности гомологичны и имеется множественное выравнивание.

Результат построения – дерево T, висячие вершины которого представляют современные последовательности S $(S_1 S_2 \ldots S_n)$, а внутренние J $(J_1 J_2 \ldots J_n)$ – гипотетические предковые.



Ребра (ветви) дерева имеют длину t $\{t_1, t_2, \ldots, t_n\}$. При этом дерево удовлетворяет ряду требований, зависящих от конкретной методики, с целью получение топологии дерева, максимально близкой к истинной.

Заметим, что даже верно восстановленное дерево, построенное для генов, не всегда соответствует видовому дереву, построенному по фенотипическим характеристикам.

Установление эволюционных взаимосвязей по последовательностям

Рибосомальные РНК, белки

Метод Фитча и Марголиаша

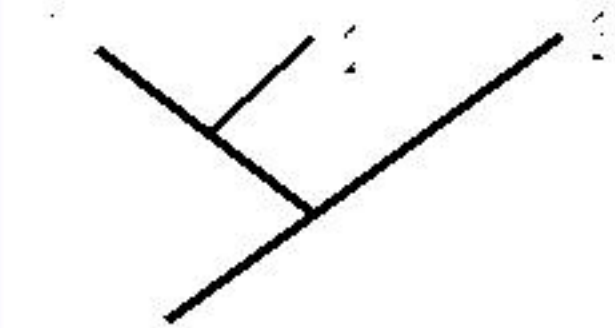
Fitch and Margoliash, 1987

Метод объединения соседей

Saitou and Nei, 1987

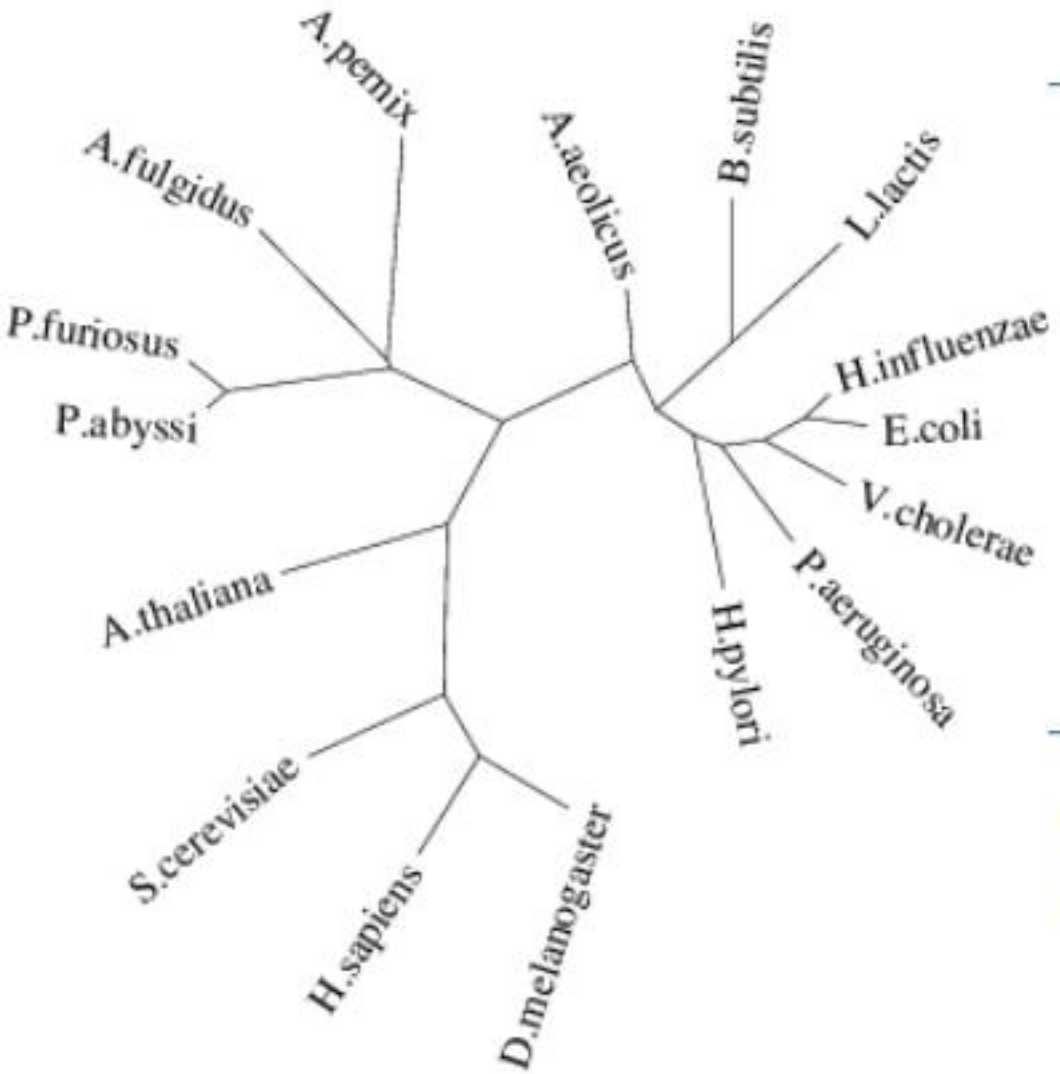
Метод максимальной парсимонии

Felsenstein, 1988



```
Organism1 AWTVASAVRTSI
Organism2 AYTVAAAVRTSI
Organism3 AWTVAAAVLTSI
```

Дерево построено на основе сходства групп ортологичных генов

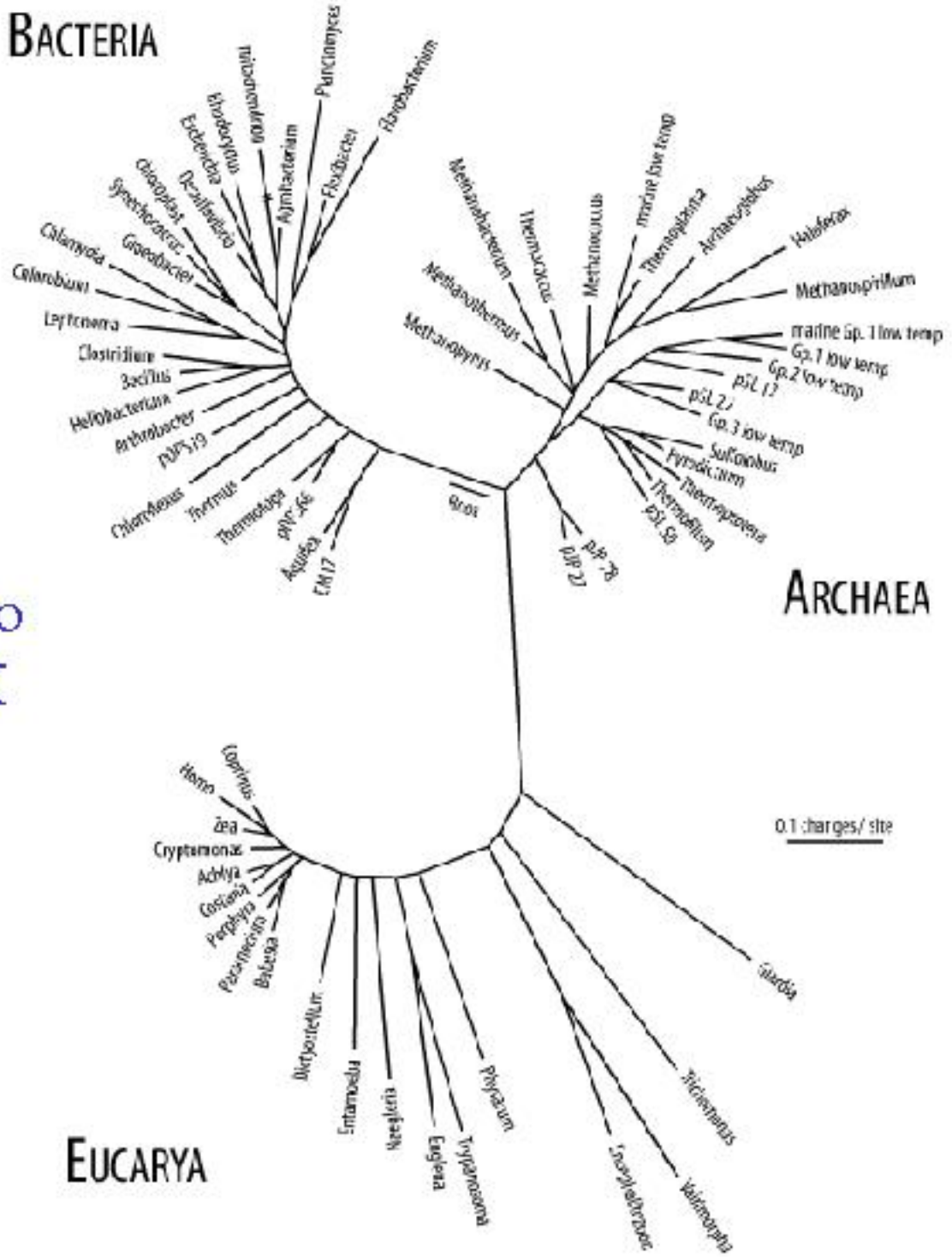


SHOT - Shared Ortholog and Gene Order Tree Reconstruction Tool

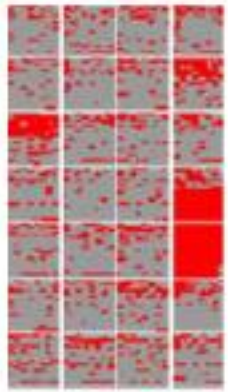
http://www.bork.embl-heidelberg.de/~korbel/SHOT/

Дерево Жизни построено по последовательностям рРНК

(Diagram courtesy of Norman Pace)

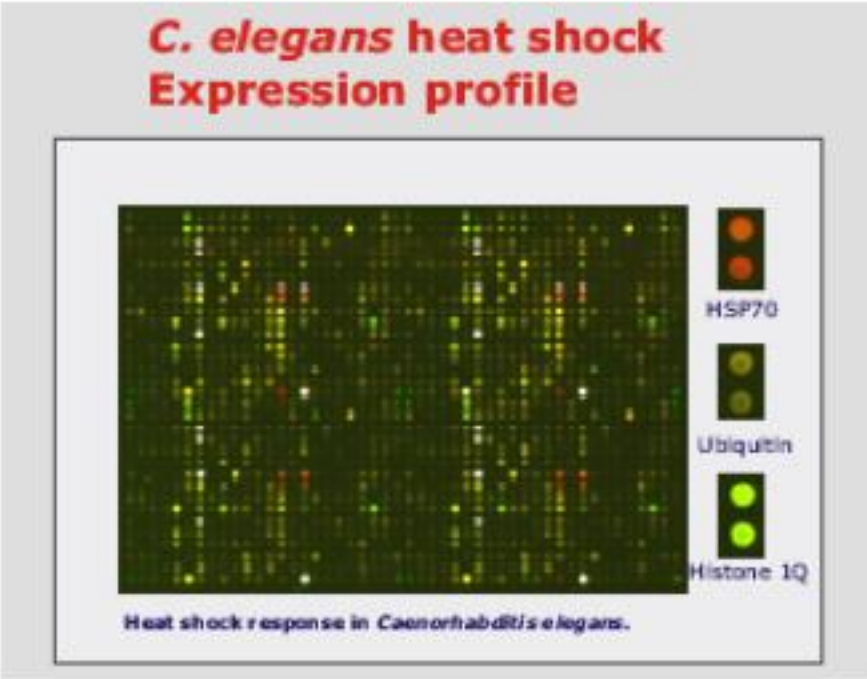


8) ДНК-чипы – экспрессионные кривые



Выявление и анализ закодированных в последовательностях ДНК функциональных сигналов требует применения современных методов распознавания образов, статистических подходов и применения специальных оптимизированных алгоритмов для преодоления вычислительных трудностей, связанных с обработкой огромных массивов информации.

Экспериментальная технология – закрепление олигонуклеотидов (30 нт) на стекле, флюоресцентное мечение гибридизации.



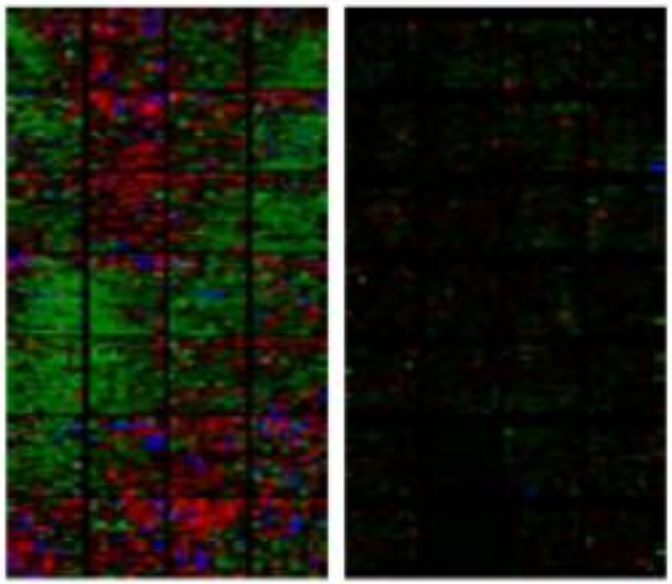
Компьютерные задачи – выбор оптимальных олигонуклеотидов для проб.
Восстановление последовательности по олигонуклеотидам и,
Актуальное направление сегодня –
Анализ кривых генной экспрессии в зависимости от условий эксперимента (ткани, времени, стадии клеточного деления, и т.д.)

http://oligo.lnatools.com/expression/

ДНК-чипы

Микропробы ДНК (микроэррэй – microarray) широко используются в биологических исследованиях. По анализу различной гибридизации на одной пластине с точечно нанесенными пробами можно определить изменения в уровнях экспрессии мРНК, вариации в числе копий ДНК и расположение сайтов связывания транскрипционных факторов в геномной шкале.

При выполнении экспериментов наибольшая проблема – обработка больших, зашумленных данных с цель. Определения специфического набора элементов, которые действительно гибридизуются различным образом. Такая обработка требует объединения различных методов и программ в единую технологическую линию.



before after

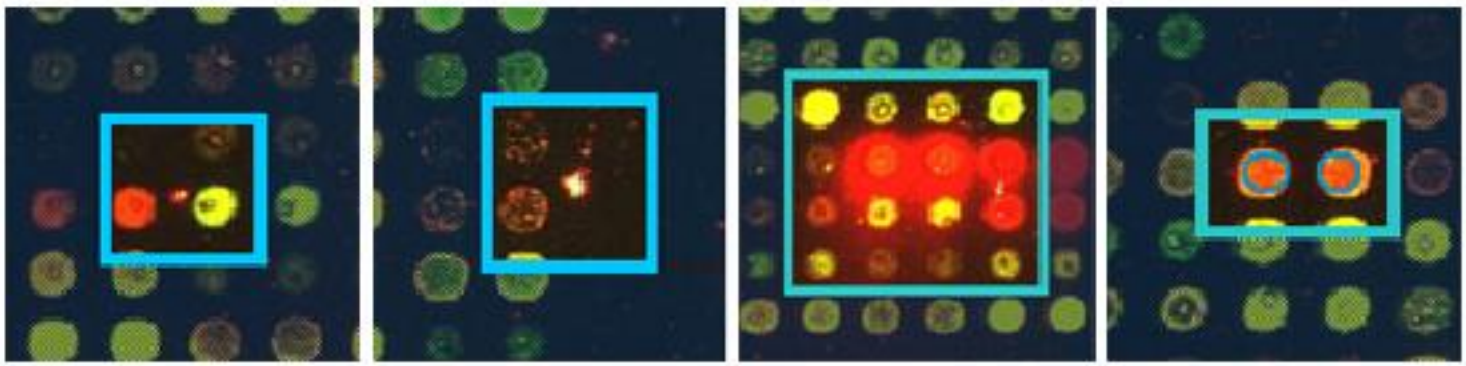
http://array.mbb.yale.edu/analysis/

http://www.cbs.dtu.dk/services/GenePublisher/

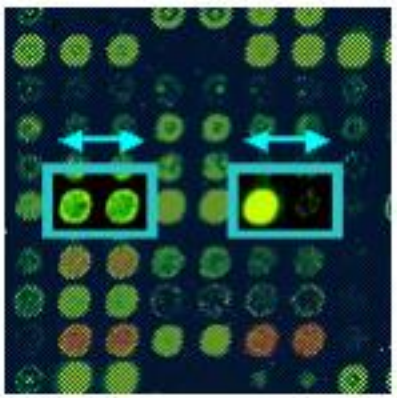
Хотя технология микроэррэй относительно нова, многие аспекты анализа данных после стадии эксперимента хорошо определены.

Это измерение интенсивности флюоресцентности, оцифровка изображения микрочипа с помощью компьютерных алгоритмов, кластеризация сходно экспрессирующихся генов и интеграция данных эксперимента с геномной информацией (базами данных).

Научная проблема – как трактовать численные данные, полученные сразу после сканирования и оцифровки изображения. Этой цели служит обработка данных (процессинг): (i) определение и минимизация уровня шума, связанного с экспериментом, (ii) оценка качества данных, полученных в эксперименте и (iii) идентификация элементов чипа, которые действительно по-разному гибридизуются.



Примеры точек (проб) гибридизации



9) задачи оперирования с большими массивами информации и управления (Интернет-навигации) разрозненными специализированными базами данных

Список вопросов к лекции:

2) **поиск гомологии и выравнивание генетических текстов**

(Понятие дот-матрицы и ее элементов, основные методы поиска гомологии)

2) **статистический анализ генетических текстов** **(структура повторов, классификация повторов)**

3) **предсказание кодирующих участков генов** и открытых рамок считывания **(постановка задачи выбора оптимального варианта белок-кодирующей последовательности)**

4) **предсказание функциональных сигналов** (функциональных сайтов и регуляторных районов) **(понятие консенсуса и паттерна)**

5) **анализ вторичной структуры РНК**

6) **анализ аминокислотных последовательностей**

7) **филогенетические сравнения**

8) **ДНК-чипы**

9) **Интернет-навигация**