

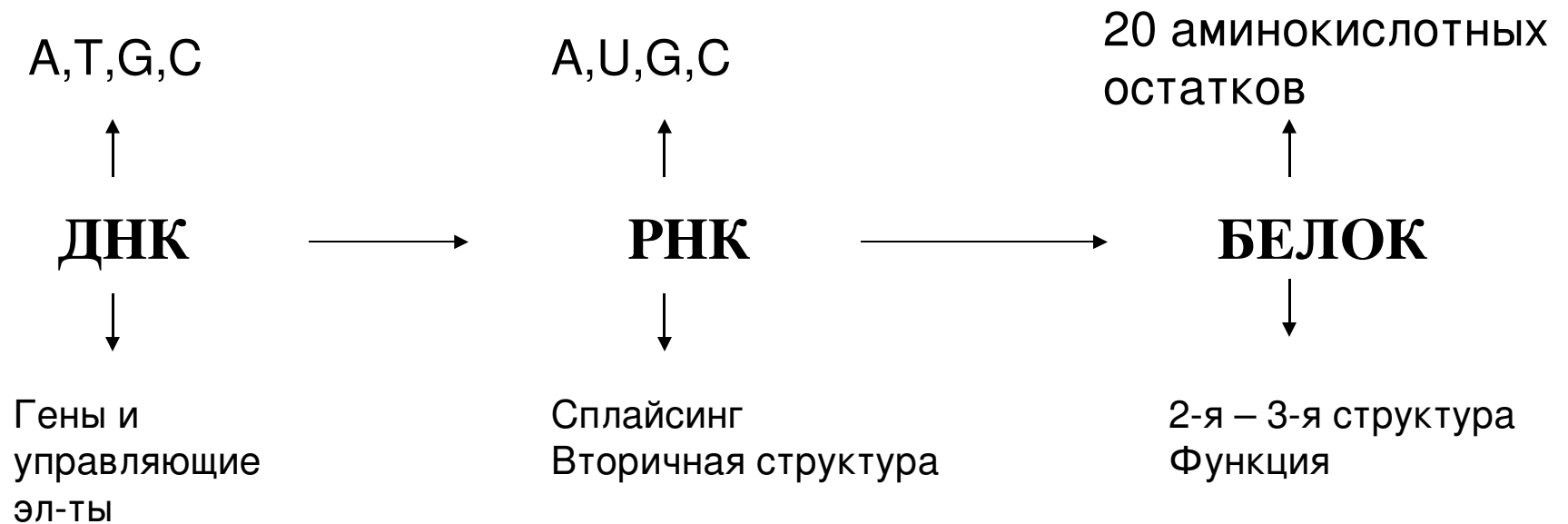
Биоинформатика и анализ биологических текстов

принципы и области применения

Пятницкий Михаил

НИИ Биомедицинской Химии РАМН

БИОМОЛЕКУЛЯРНЫЕ ТЕКСТЫ



ТЕКСТ – **ПОСЛЕДОВАТЕЛЬНОСТЬ** СИМВОЛОВ КОНЕЧНОГО АЛФАВИТА, НЕСУЩАЯ СМЫСЛОВУЮ НАГРУЗКУ

АЛФАВИТ

(МНОЖЕСТВО, ЗАДАВАЕМОЕ
ПЕРЕЧИСЛЕНИЕМ)

ПРАВИЛА

(ГРАММАТИКА, СЕМАНТИКА)

ТЕКСТ

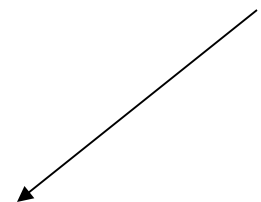
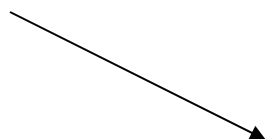
(ПОСЛЕДОВАТЕЛЬНОСТЬ **ВЫБОРОВ** С
ПОВТОРЕНИЯМИ ИЗ АЛФАВИТА)

ИНФОРМАЦИЯ

(УМЕНЬШЕНИЕ НЕОПРЕДЕЛЕННОСТИ)

СМЫСЛОВАЯ НАГРУЗКА

(«ЧТЕНИЕ»)



TGATGATGAAGACATCAGCATTGAAGGGCTGATGGAACACATCCCGGGGCCGGAC
TTCCCGACGGCGGCAATCATTAAACGGTCGTCGCGGTATTGAAGAAGCTTACCGTA
CCGGTCGCGGCAAGGTGTATATCCGCGCTCGCGCAGAAGTGGAAGTTGACGCCAA
AACCGGTCGTGAAACCATTATCGTCCACGAAATTCGGTATCAGGTAAACAAAGCG
CGCCTGATCGAGAAGATTGCGGAACTGGTAAAAGAAAAACGCGTGGAAGGCATCA
GCGCGCTGCGTGACGAGTCTGACAAAGACGGTATGCGCATCGTGATTGAAGTGAA
ACGCGATGCGGTGCGTGAAGTTGTGCTCAACAACCTCTACTCCCAGACCCAGTTG
CAGGTTTCTTTCGGTATCAACATGGTGGCATTGCACCATGGTCAGCCGAAGATCA
TGAACCTGAAAGACATCATCGCGGCGTTTGTTCGTCACCGCCGTGAAGTGGTGAC
CCGTCGTAATAATTTTCGAACTGCGTAAAGCTCGCGATCGTGCTCATATCCTTGAA
GCATTAGCCGTGGCGCTGGCGAACATCGACCCGATCATCGAACTGATCCGTCATG
CGCCGACGCCTGCAGAAGCGAAAACCTGCGCTGGTTGCTAATCCGTGGCAGCTGGG
CAACGTTGCCGCGATGCTCGAACGTGCTGGCGACGATGCTGCGCGTCCGGAATGG
CTGGAGCCAGAGTTCGGCGTGCGTGATGGTCTGTACTACCTGACCGAACAGCAAG
CTCAGGCGATTCTGGATCTGCGTTTGCAGAAACTGACCGGTCTTGAGCACGAAAA
ACTGCTCGACGAATACAAAGAGCTGCTGGATCAGATCGCGGAACTGTTGCGTATT
CTTGGTAGCGCCGATCGTCTGATGGAAGTGATCCGTGAAGAGCTGGAGCTGGTTC
GTGAACAGTTCGGTGACAAACGTCGTAATAACCGCCAACAGCGCAGACAT
CAACCTGGAAGATCTGATCACCCAGGAAGATGTGGTCGTGACGCTCTCTCACCAG
GGCTACGTTAAGTATCAGCCGCTTTCTGAATACGAAGCGCAGCGTCGTGGCGGGA

Форматы хранения последовательностей

- Файлы последовательностей должны быть исключительно текстовыми (ASCII) и не содержать специальных символов, т.е. их не стоит редактировать в Microsoft Word.
- Разные базы данных используют слегка отличающиеся форматы для хранения последовательностей
- Важно правильно использовать и при необходимости конвертировать один формат в другой
- Существует более 20 форматов для хранения последовательностей. Наиболее распространенные: GenBank и FASTA

GenBank file

```
FT      CDS                522..1985
FT                                  /codon_start=1
FT                                  /db_xref="GOA:P00179"
FT                                  /db_xref="SWISS-PROT:P00179"
FT                                  /gene="CYP2C5"
FT                                  /product="progesterone 21-hydroxylase"
FT                                  /protein_id="AAA63461.1"
FT                                  /translation="MDPVVVLVLGLCCLLLLSIWKQNSGRGKLPPGPTPFPIIGNILQI
FT DAKDISKSLTKFSECYGPVFTVYLGMKPTVVLHGYEAVKEALVDLGEFAGRGSVP
FT KVSKGLGIAFSNAKTWKEMRRFSLMTLRNFGMGKRSIEDRIQEEARCLVEELRKT
FT NASP
FT CDPTFILGCAPCNVICSVIFHNRFDYKDEEFLKLMESLNENVRIILSSPWLQVY
FT NFPAL
FT LDYFPGIHKTL LKNADYIKNFIMEKVKEHQKLLDVNNPRDFIDCFLIKMEQEN
FT NLEFTL
FT ESLVIAVSDLFGAGTETTSTTLRYSLLLLLKHPEVAARVQEEIERVIGRHRSP
FT CMOQRS
FT RMPYTD A VIHEIQRFIDLLPTNLPHAVTRDVRFRNYFIPKGTDIITSLTSVL
FT HDEKAFP
FT NPKVFDPGHFLDESGNFKKSDYFMPFSAGKRM CVGEG LARME LFLFLTSILQ
FT NFKLQSL
FT VEPKDL DITAVVNGFVSVPPSYQLCFIPI«
SQ Sequence 1641 BP; 518 A; 294 C; 343 G; 486 T; 0 other;
  1  gttgcactca  tgatattaag  gaagaatctt  aaaaaacctg  actcaattcc  taatatacca
 60  ccagggccat  ggaaactacc  aataatagga  agcatacccc  atctcgttgg  ttctccacca
120  cacagaaaat  taagagattt  ggccaaaaaa  tatggcccct  tgatgcacct  tcaacttgga
180  gaggtcatct  tcatcattgt  ttctcagca  gagtatgcta  aggaagtcac  gaaaacccat
240  gatgtcacat  ttgcatccag  gcctcgttct  cttttcacag  atatagtgtt  ttatgggtcc
300  acagacatag  gcttttcacc  ttatggtgat  tactggagac  aagttcgaaa  gatttgcaat
360  gtagagcttc  taagtatgaa  acgtgtccag  tctttatggc  caatcagggg  ggaagagggtg
420  aaaaatctaa  tccaacgcat  tgcatacaga  gaaggggccg  tcgtcaatct  ttctcaagct
480  attgattcat  tgattttcac  aatcacttca  aggtctgctt  ttggcaagag  atacatggag
540  caagaagagt  tcatatcatg  cgtaagagaa  gttatgaagc  tagctggagg  tttcaacata
```

FASTA file

>gi|1786197|gb|AAC73126.1| chaperone Hsp40, co-chaperone with DnaK
MAKQDYEEILGVSKTAEEREIRKAYKRLAMKYHPDRNQGDKEAEAKFKEIKEAYEVLTD SQKRAAYDQYGH
AAFEQGGMGGGGFGGGADFSDIFGDVFGDIFGGGRGRQRAARGADLRYNMELTLEEAVRGVTKEIRIPTLE
ECDVCHGSGAKPGTQPQTCPTCHGSGQVQMRQGGFAVQQTCPHCQGRGTLIKDPCNKCHGHGRVERS KTLS
VKIPAGVDTGDRIRLAGEGEAGEHGAPAGDLYVQVQVKQHPIFEREGNNLYCEVPINFAMAALGGEIEVPT
LDGRVCLKVPGETQTGKLFMRGKGVKSVRGGAGDQLLCRVVETPVGLNERQKQLLQELQESFGGPTGEH
NSPRSKSFFDGVKKFFDDLTR

>gi|1786198|gb|AAC73127.1| IS186/IS421 transposase
MNYSHDNWSAILAHIGKPEELDTSARNAGALTRRREIRDAATLLRLGLAYGPGGMSLREVTAWAQLHDVAT
LSDVALLKRLRNAADWFGILAAQTLAVRAAVTGCTSGKRLRLVDGTASAPGGGSAEWRLHMGYDPHTCQF
TDFELTDSRDAERLDRFAQTAD EIRIADRGFGSRPECIRSLAFGEADYIVRVHWRGLRWLTAEGMRFDMMG
FLRGLDCGKNGETTVMIGNSGNKKAGAPFPARLIAVSLPPEKALISKTRLLSENRRKGRVVQAETLEAAGH
VLLLTSLPEDEYSAEQVADCYRLRWQIELAFKRLKSLHLHDALRAKEPELAKAWIFANLLAAFLIDDI IQP
SLDFPPRSAGSEKKN

>gi|1786200|gb|AAC73129.1|regulatory protein for HokC, overlaps hokC
MLNTRCVPLTDRKVKEKRAMKQHKAMIVALIVICITAVVAALVTRKDLCEVHIRTGQTEVAVFTAYESE

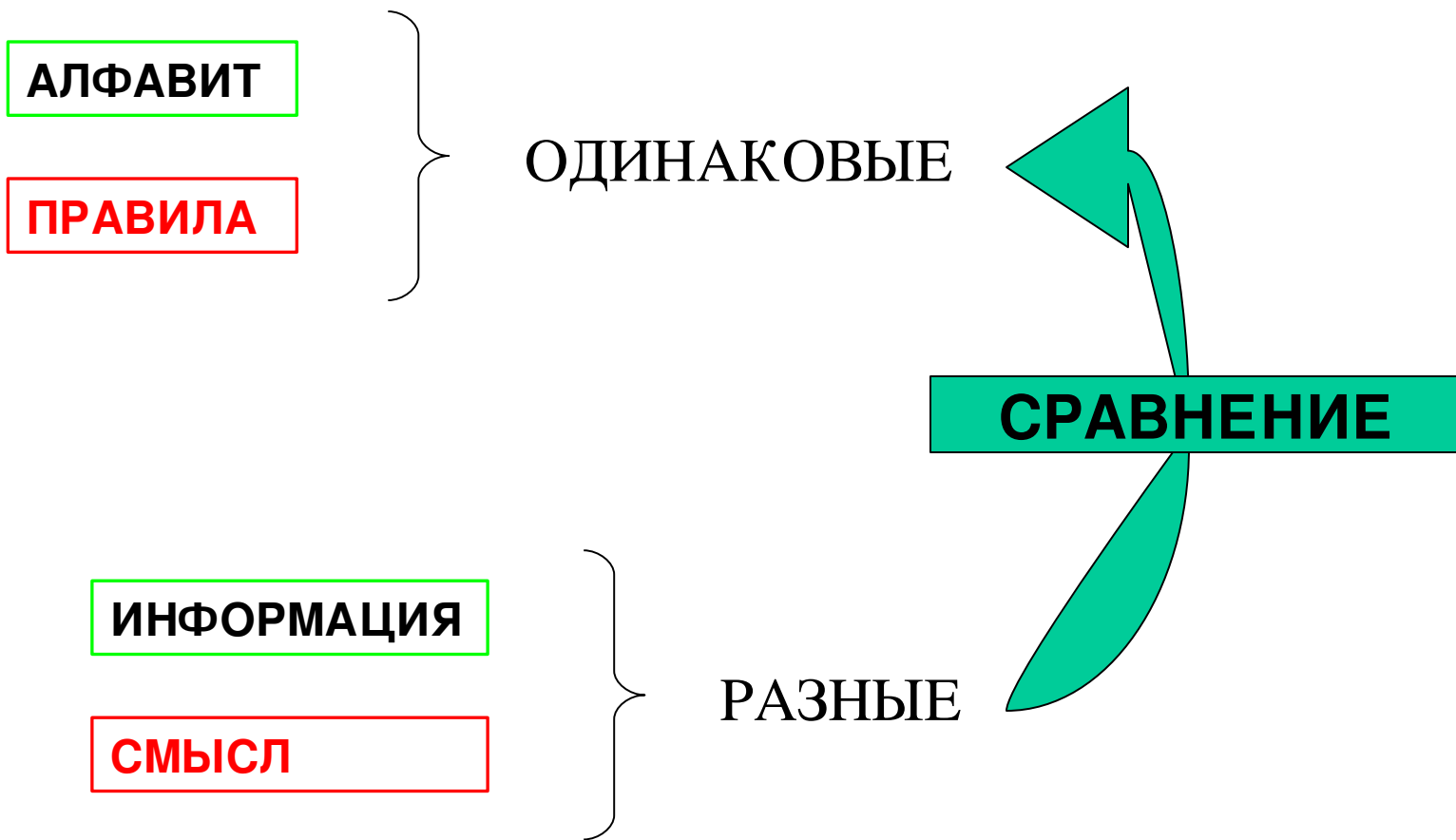
>gi|48994874|gb|AAT48122.1| toxic membrane protein
MKQHKAMIVALIVICITAVVAALVTRKDLCEVHIRTGQTEVAVFTAYESE

TGATGATGAAGACATCAGCATTGAAGGGCTGATGGAACACATCCCGGGGCCGGAC
TTCCCGACGGCGGCAATCATTAAACGGTCGTCGCGGTATTGAAGAAGCTTACCGTA
CCGGTCGCGGCAAGGTGTATATCCGCGCTCGCGCAGAAGTGGAAGTTGACGCCAA
AACCGGTCGTGAAACCATTATCGTCCACGAAATTCGGTATCAGGTAAACAAAGCG
CGCCTGATCGAGAAGATTGCGGAACTGGTAAAAGAAAAACGCGTGGAAGGCATCA
GCGCGCTGCGTGACGAGTCTGACAAAGACGGTATGCGCATCGTGATTGAAGTGAA
ACGCGATGCGGTGCGTGAAGTTGTGCTCAACAACCTCTACTCCCAGACCCAGTTG
CAGGTTTCTTTCGGTATCAACATGGTGGCATTGCACCATGGTCAGCCGAAGATCA
TGAACCTGAAAGACATCATCGCGGCGTTTGTTCGTCACCGCCGTGAAGTGGTGAC
CCGTCGTAATAATTTTCGAACTGCGTAAAGCTCGCGATCGTGCTCATATCCTTGAA
GCATTAGCCGTGGCGCTGGCGAACATCGACCCGATCATCGAACTGATCCGTCATG
CGCCGACGCCTGCAGAAGCGAAAACCTGCGCTGGTTGCTAATCCGTGGCAGCTGGG
CAACGTTGCCGCGATGCTCGAACGTGCTGGCGACGATGCTGCGCGTCCGGAATGG
CTGGAGCCAGAGTTCGGCGTGCGTGATGGTCTGTACTACCTGACCGAACAGCAAG
CTCAGGCGATTCTGGATCTGCGTTTGCAGAAACTGACCGGTCTTGAGCACGAAAA
ACTGCTCGACGAATACAAAGAGCTGCTGGATCAGATCGCGGAACTGTTGCGTATT
CTTGGTAGCGCCGATCGTCTGATGGAAGTGATCCGTGAAGAGCTGGAGCTGGTTC
GTGAACAGTTCGGTGACAAACGTCGTAATAACCGCCAACAGCGCAGACAT
CAACCTGGAAGATCTGATCACCCAGGAAGATGTGGTCGTGACGCTCTCTCACCAG
GGCTACGTTAAGTATCAGCCGCTTTCTGAATACGAAGCGCAGCGTCGTGGCGGGA

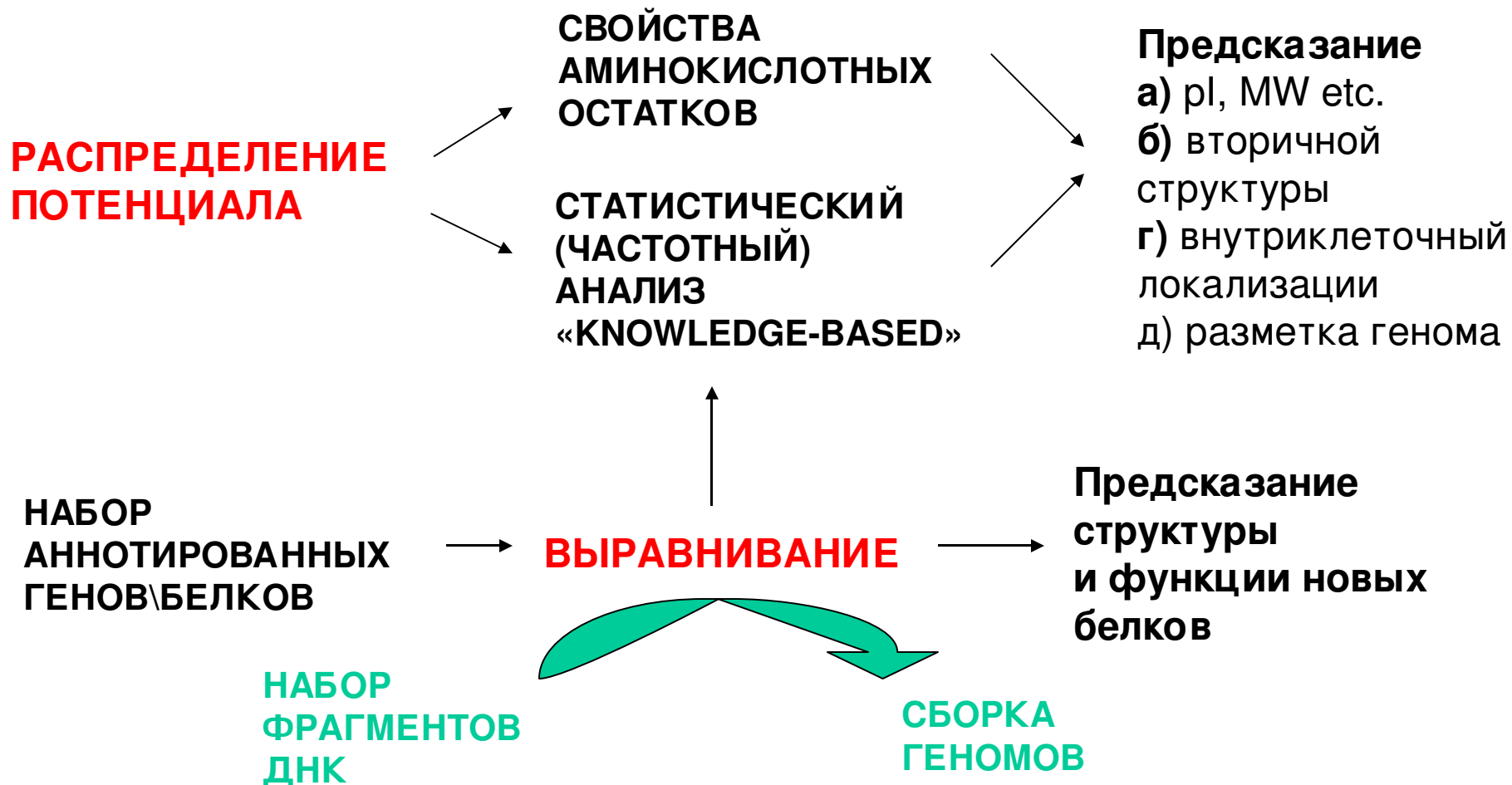
Возможные вопросы

- Можно ли определить из какого организма эта ДНК?
- Отличается ли эта последовательность от другой ДНК этого организма?
- Какие параметры характеризуют эту последовательность?
- Последовательность – кодирует белок? Транспозон? Регулятор?

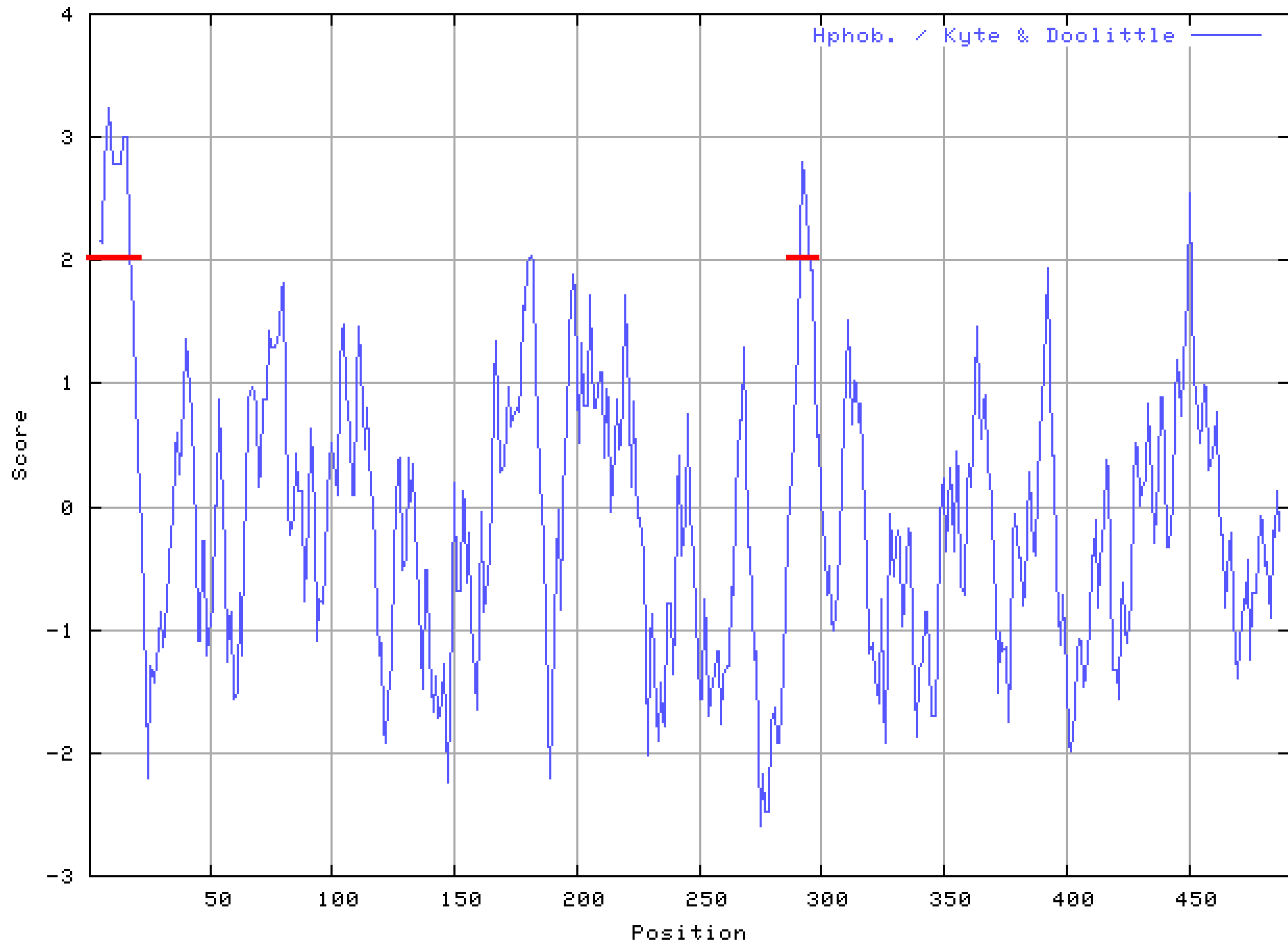
НАБОР ТЕКСТОВ



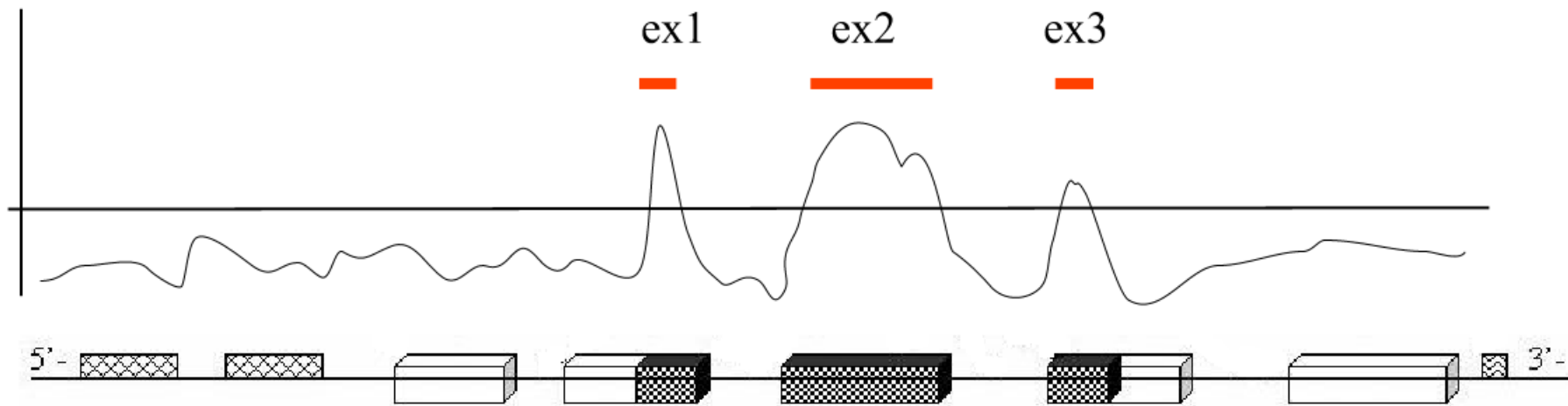
МЕТОДЫ АНАЛИЗА ГЕНЕТИЧЕСКИХ И БЕЛКОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ



ProtScale output for user sequence



КОДИРУЮЩИЙ ПОТЕНЦИАЛ



- **Позиционно-специфичные свойства**

i $i+1$ $i+2$ $i+17$ $i+28$
atcgatcagtatcgat GT ctgagctatgag

Потенциал в позиции $i = (\text{остаток } (i) = ? \text{ 'a'}) + (\text{остаток } (i + 17) = ? \text{ 'c'})$

Поиски схожих последовательностей в базах данных

- До 70х - 80х годов не было ясности в степени сходства последовательностей ДНК и белков на молекулярном уровне.
- Экспоненциальный рост баз данных по последовательностям – необходимость развития эффективных методов для поиска сходных последовательностей.
- Были разработаны первые программы для поиска сходных последовательностей :

BLAST (Altschul et al. 1990) Basic Local Alignment Search Tool

FASTA (1988)

Clustal

Поиск сходных последовательностей

- Первый шаг – **выравнивание (alignment)** двух или более последовательностей
- Результат поиска – список последовательностей из базы данных с которыми может быть выровнена исходная последовательность. Список отсортирован по уровню меры сходства
- Пример – поиск гена, похожего на заданный. Найденный сходный ген может дать дополнительную информацию о возможной функции.
- Пример – последовательность с известной функцией ищется в другом геноме (поиск гомологов)
- Поиск должен быть **быстрым и чувствительным!**
- К сожалению, обычно это взаимоисключающие понятия

Оценка сходства последовательностей (счет выравнивания)

- Рассматриваем каждый сайт выравнивания
- Счет(score) в каждом сайте
 - Положительный если совпадают
 - Отрицательный, если не совпадают
- Общий счет – сумма по всем сайтам
 - Значимость зависит от длины последовательности

GTAGTC

CTAGCG

Только замены

За совпадение +2, за несовпадение -1

TTCGTCGTAGTCGGCTCGACCTG
GTACGTCTAGCGAGCGTGATCCT

9 совпадений **+18**

14 несовпадений **-14**

} Общий счет **+4**

Включение пропусков (gaps)

– обычно, сравниваемые последовательности разной длины

За совпадение +2, за несовпадение –1, за вставку -1

TT-CGTCGTAGTCG-GC-TCGACC-TG
GTACGTC-TAG-CGAGCGT-GATCCT-

17 совпадений +34

2 несовпадения - 2

8 вставок - 8

} Общий счет +24

Выбор выравнивания

- Возможны различные варианты выравниваний!
 - Нужно перебрать **все возможные варианты!!**
 - Выберем выравнивание с наилучшим счетом
 - Таких выравниваний может быть несколько

TT-CGTCGTAGTCG-GC-TCGACC-TG	+24
GTACGTC-TAG-CGAGCGT-GATCCT-	
-TTCGT-CGTAGTC-GGCTCG-ACCTG	0
GTAC-GTCTA-GCGAGCGT-GATCC-T	

Почему это трудно ?

Выравнивание (без вставок) требует алгоритма который производит число сравнений \sim квадрату длины последовательности

При учете возможности вставок, число вариантов становится астрономическим!

**Решение – алгоритмы основанные на
ДИНАМИЧЕСКОМ ПРОГРАММИРОВАНИИ**

Глобальное выравнивание vs локальное выравнивание

глобальное

ATTGCAGTG–TCGAGCGTCAGGCT

ATTGCGTCGATCGCAC–GCACGCT

локальное

CATATTGCAGTGGTCCCGCGTCAGGCT

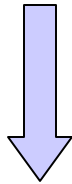
TAAATTGCGT–GGTCGCACTGCACGCT

ВЫРАВНИВАНИЕ ЭМУЛИРУЕТ СОБЫТИЯ МОЛЕКУЛЯРНОЙ ЭВОЛЮЦИИ

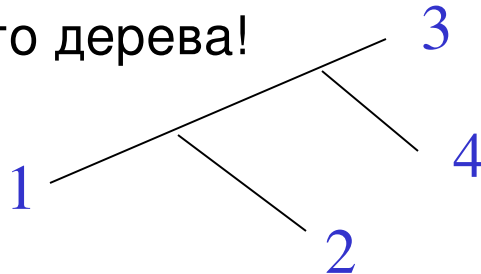
1 АТАГСТАГСАТGCATCATCTTC
2 АТАССТАГСАТGCATCATCATC
3 АТАГСТАГСАТGGATCATCATC
4 АТАГСТААСАТGGATC-TCATC

Несовпаде
ние

(«замена»)
Вставка(г
ар)



Оценка филогенетического дерева!



ЗАМЕНЫ АМИНОКИСЛОТНЫХ ОСТАТКОВ НЕРАВНОЗНАЧНЫ

Замена S на T

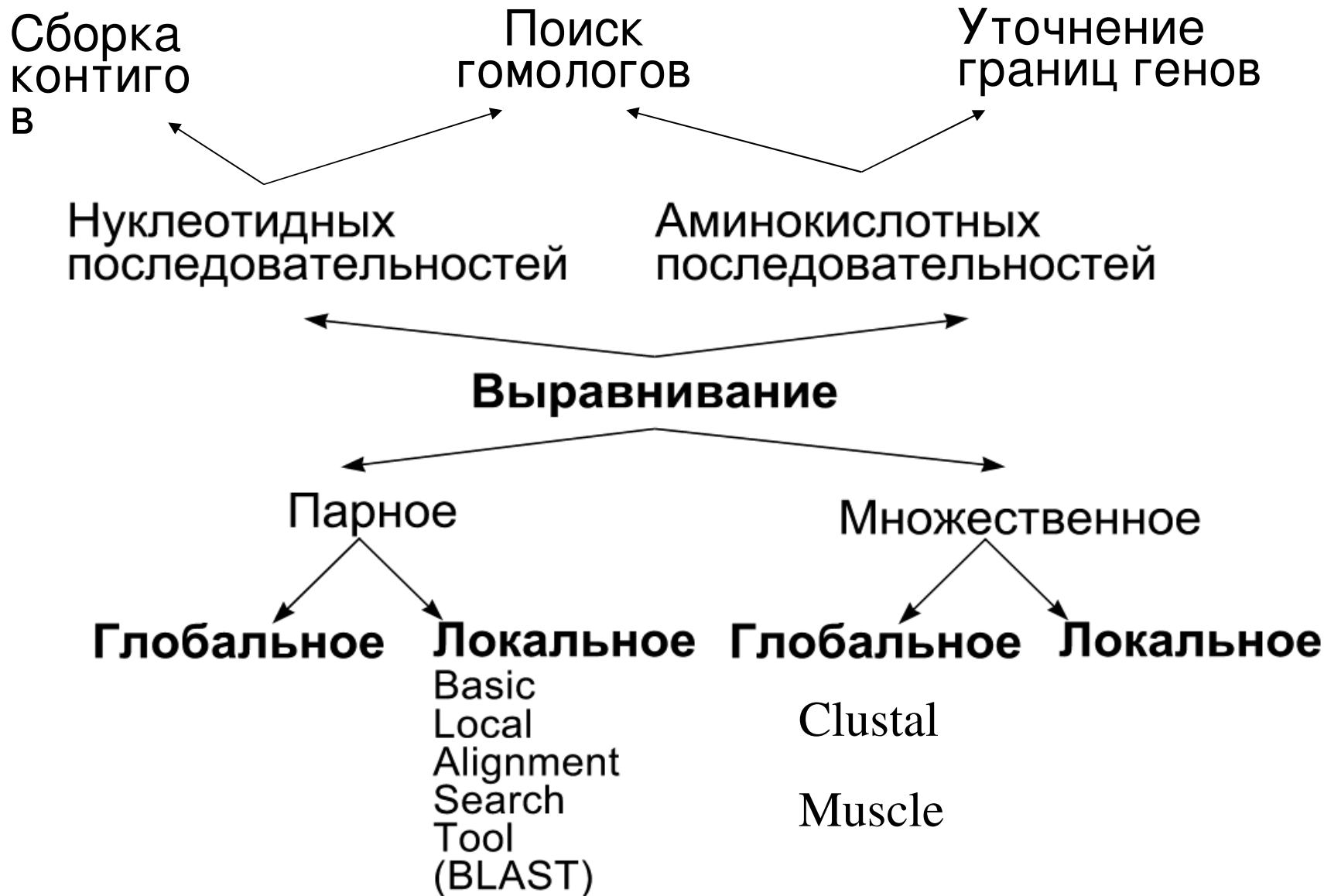
L-PPGPFPLPVL-N

--PPGPRPLPV-N

Замена V на L

- ⇒ МАТРИЦЫ ЗАМЕН АМИНОКИСЛОТНЫХ ОСТАТКОВ
- ЭВОЛЮЦИОННЫЕ («KNOWLEDGE-BASED»)
 - ПО ФИЗИКО-ХИМИЧЕСКИМ СВОЙСТВАМ

КЛАССИФИКАЦИЯ АЛГОРИТМОВ ВЫРАВНИВАНИЯ



КОНСЕНСУСНАЯ ПОСЛЕДОВАТЕЛЬНОСТЬ СОДЕРЖИТ ОБЩУЮ ЧАСТЬ НЕСКОЛЬКИХ ПЕРВИЧНЫХ СТРУКТУР

МНОЖ. ВЫРАВНИВАНИЕ

L-PPGPSPLPVL-N

--PPGPTPLPVV-N

RYPPGPLPLPGIGN

L-PPGPRPLSIL--

.-PPGP.PL...-

КОНСЕНСУС

ИНВАРИАНТНАЯ ПОЗИЦИЯ

ВАРИАБЕЛЬНАЯ ПОЗИЦИЯ

ТЕРМИНОЛОГИЯ ВЫРАВНИВАНИЯ

- Вставка\делеция insertion\deletion ‘-’
- Замена, матрица замен (substitution matrix)
- Консенсусная последовательность (consensus)
- Счет выравнивания (score)

Результаты работы BLAST: www.ncbi.nlm.nih.gov/blast

RID=1077749476-12278-195445991153.BLASTQ3, - Netscape

File Edit View Go Communicator Help

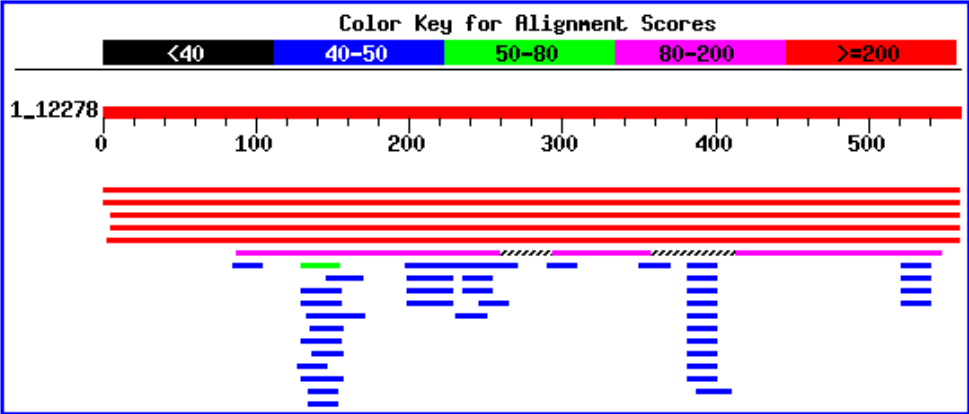
Back Forward Reload Home Search Netscape Print Security Shop Stop

Instant Message WebMail Calendar Radio People Yellow Pages Download Customize... RealPlayer

Bookmarks Location: <http://www.ncbi.nlm.nih.gov:80/BLAST/Blast.cgi> What's Related

Distribution of 49 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments



Color Key for Alignment Scores

Score Range	Color
<40	Black
40-50	Blue
50-80	Green
80-200	Magenta
>=200	Red

1_12278

Sequences producing significant alignments:

Accession	Description	Score (bits)	E Value
gi 42662196 dbj AB162821.1	<i>Pseudomonas japonica</i> apd1 gene ...	1108	0.0
gi 9947973 gb AE004624.1	<i>Pseudomonas aeruginosa</i> PAO1, sect...	551	e-154
gi 26557028 gb AE016784.1	<i>Pseudomonas putida</i> KT2440 sectio...	551	e-154
gi 8671335 emb AJ009858.2 PAE9858	<i>Pseudomonas aeruginosa</i> ex...	543	e-151
gi 18446632 gb AY048591.1	<i>Pseudomonas putida</i> PQQ-linked al...	531	e-148
gi 13194685 gb AF326086.1	<i>Pseudomonas butanovora</i> 1-butanol...	151	2e-33
gi 22779358 dbj AB091400.1	<i>Pseudomonas putida</i> aldA, qbdB, ...	52	0.001
gi 9255864 gb AF277373.1 AF277373	<i>Ralstonia eutropha</i> terahy...	48	0.018
gi 39652705 emb BX572593.1	<i>Rhodopseudomonas palustris</i> CGA0...	46	0.072
gi 24426505 emb AL939115.1 SC0939115	<i>Streptomyces coelicolo...</i>	46	0.072
gi 2055285 dbj D86375.1	<i>Gluconobacter suboxydans</i> DNA for d...	46	0.072

Document: Done

ПРАКТИЧЕСКИЕ НАВЫКИ

- формат FASTA

- <http://www.ncbi.nlm.nih.gov/blast>

- <http://au.expasy.org/tools/>

Compute pI/Mw

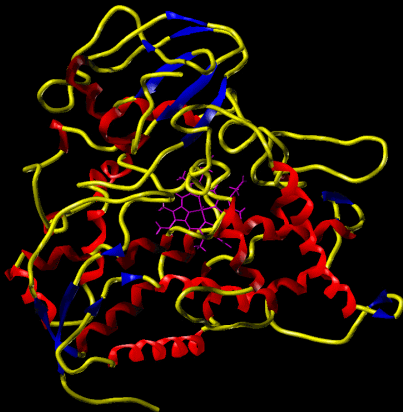
ProtScale

ClustalW

Спасибо за внимание!

КОНСТРУИРОВАНИЕ 3-х МЕРНОЙ СТРУКТУРЫ БЕЛКА ПО ГОМОЛОГИИ

ШАБЛОН



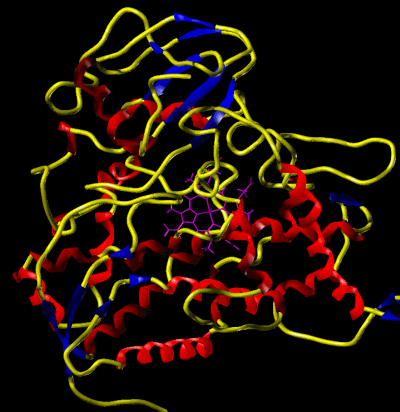
ВЫРАВНИВАНИЕ

```
1 MEALGFLKLEVNGPMVTVALSVALLALLKWYSTSAFSRLEKLGLRHPKPSPFIGNLTFFR
1 MMTTSLIW----GIAIAACCLWLILGIRRRQTGE-PPLEN-GL-----IPYLG----CA

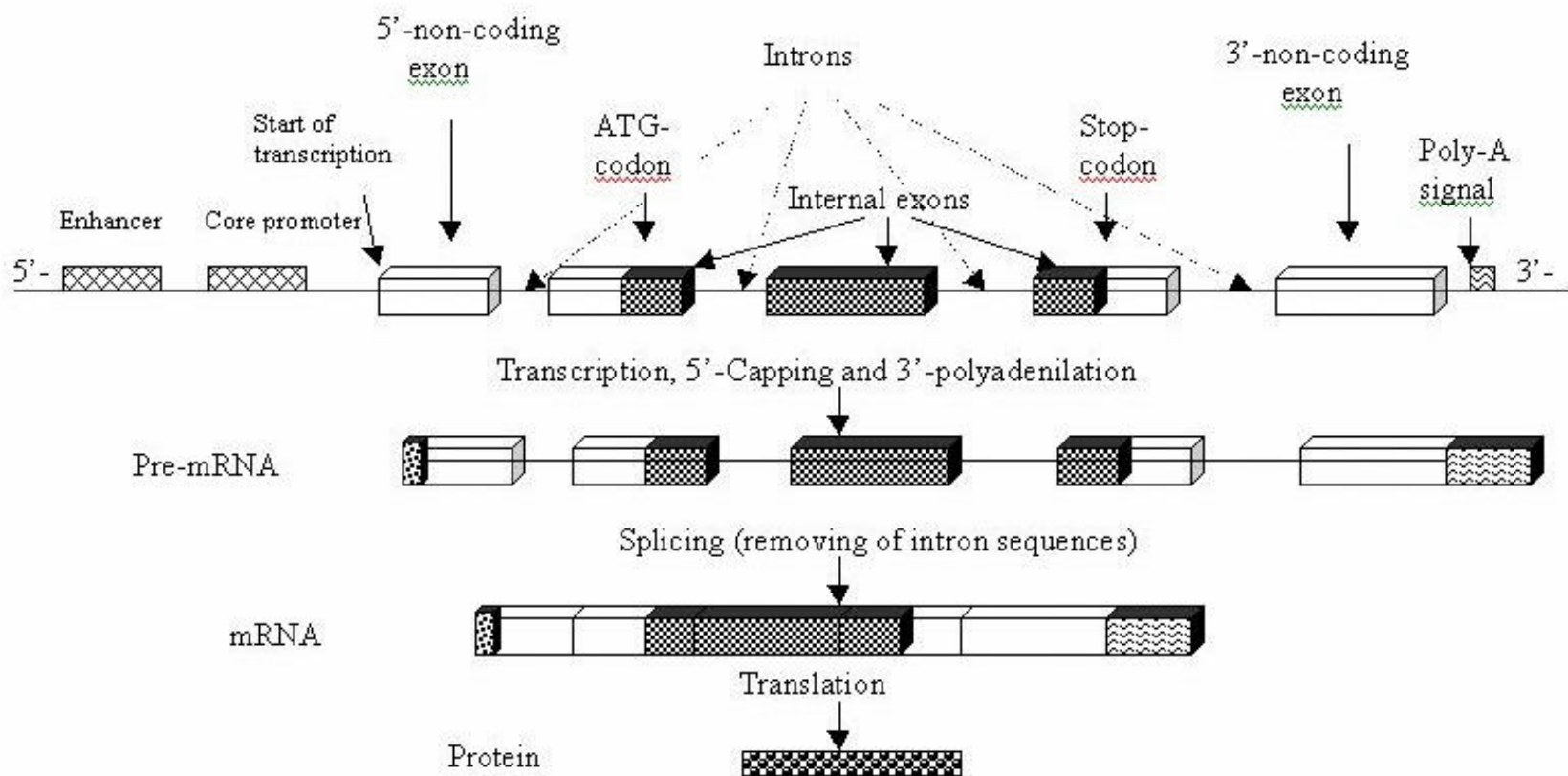
61 QGFWESQMELRKLYGPLCGYYLGRRMFIVISEPDMIKQVLVENFSNFTNRMASGLEFKSV
46 LQFGANPLEFLRANQRKHGHVFTCK-----LMGKYVHF---ITNPLSYHKV

121 ADSVLFLRDKRWEEVR-
89 LCHGKYF---DWKKFH
```

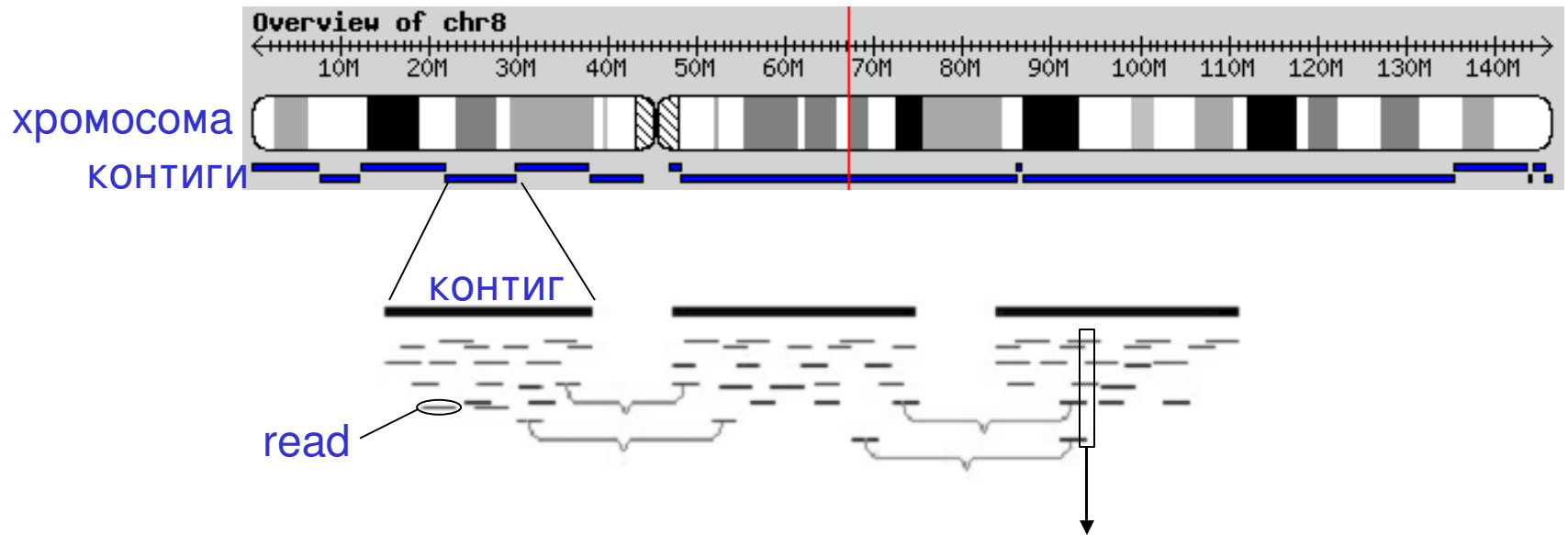
МИШЕНЬ



СТРУКТУРА ГЕНА



СБОРКА КОНТИГОВ



```
TAGCTTACACAGATTACTGC
TAGATAACACAGATTACTGA
TAG TTACACAGAGTATTGC
TAGATAACAC GATTACTGA
TAGATTACACAGACTACTGA
```

SNP

Хромосома	>100 MB
Контиг	1-50 MB
Read	1-10 kB

ЗАДАЧА АНАЛИЗА БИМОЛЕКУЛЯРНЫХ ТЕКСТОВ

