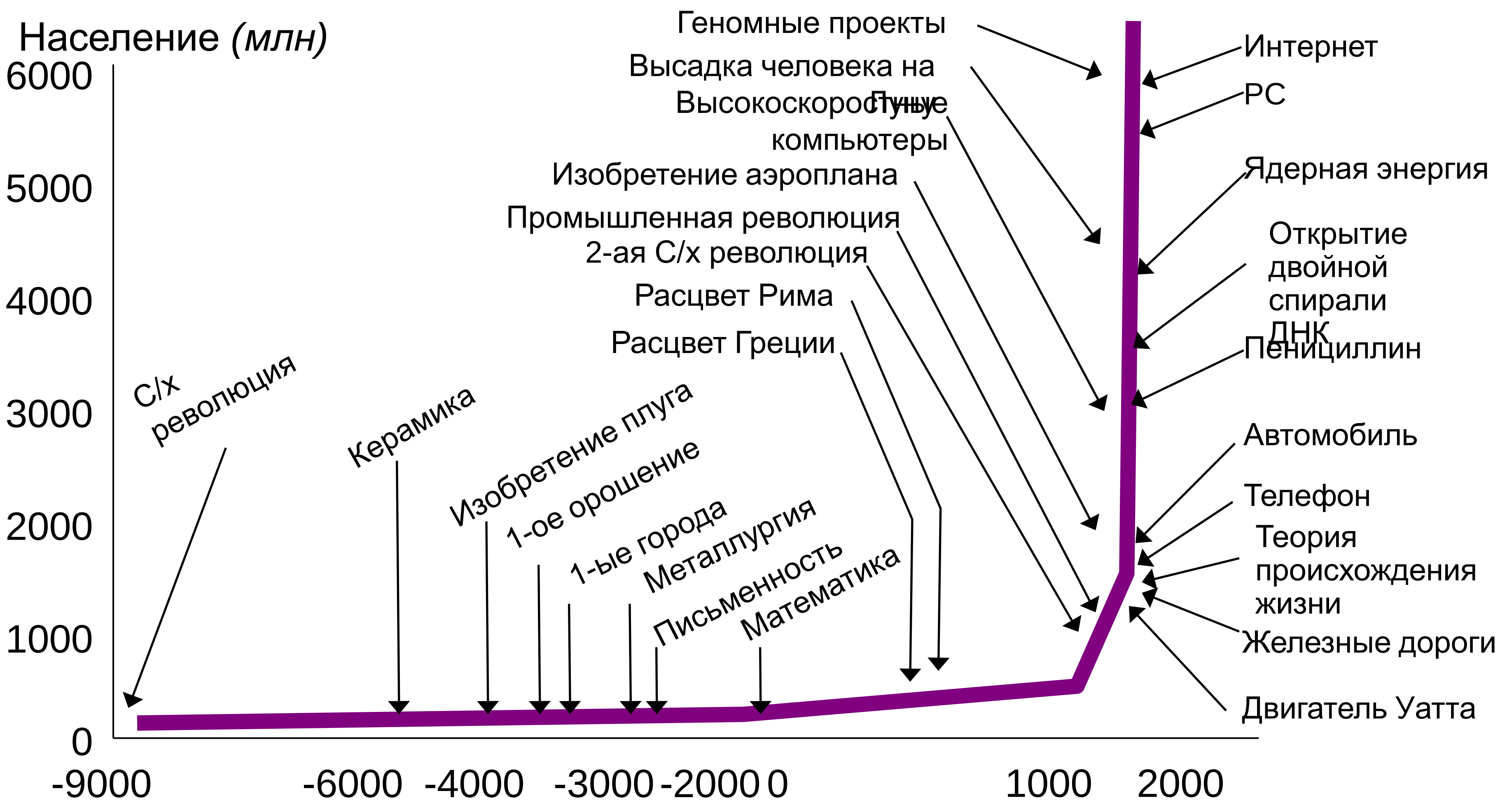
Биоинформатика

Коротков Евгений Вадимович
Институт Биоинженерии, ФИЦ Биотехнологии РАН

bioinf@yandex.ru

ИСТОРИЯ ТЕХНОЛОГИЙ



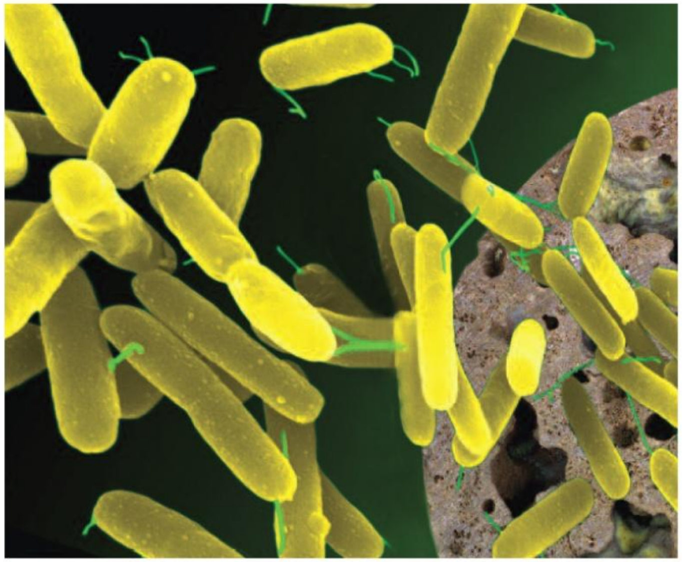
Ист.: Milken Institute, Robert Fogel/University of Chicago

**Чтение и анализ
генетических текстов**

**Манипулирование известными
и создание новых,
ранее не существовавших в природе,
генетических текстов**

**Создание органов и организмов
с рукотворными
генетическими программами**

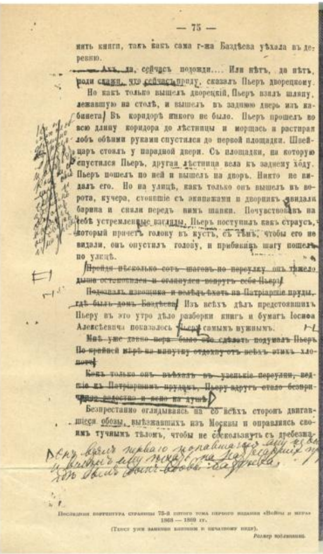
число букв в геноме микроба



Геном бактерии
2 500 000 букв

=

число букв в романе Л.Н.Толстого “Война и Мир”



— 75 —

нять книги, такъ какъ сама г-жа Баздѣева уѣхала въ деревню.

~~— Ахъ, да,~~ сейчасъ подожди.... Или нѣтъ, да нѣтъ, поди скажи, что сейчасъ приду, сказалъ Пьеръ дворецкому.

Но какъ только вышелъ дворецкій, Пьеръ взялъ шляпу, лежавшую на столѣ, и вышелъ въ заднюю дверь изъ кабинета. Въ коридорѣ никого не было. Пьеръ прошелъ во всю длину коридора до лѣстницы и морщась и растирая лобъ обѣими руками спустился до первой площадки. Швейцаръ стоялъ у парадной двери. Съ площадки, на которую спустился Пьеръ, другая лѣстница вела къ заднему ходу. Пьеръ пошелъ по ней и вышелъ на дворъ. Никто не видалъ его. Но на улицѣ, какъ только онъ вышелъ въ ворота, кучера, стоявшіе съ экипажами и дворникъ увидали барина и сняли передъ нимъ шапки. Почувствовавъ на себѣ устремленные взгляды, Пьеръ поступилъ какъ страусъ, который прячетъ голову въ кустъ, съ тѣмъ, чтобы его не видали, онъ опустилъ голову, и прибавивъ шагу пошелъ по улицѣ.

~~Пройдя нѣсколько сотъ шаговъ по переулку онъ тяжело дыша остановился и оглянулся вокругъ себя Пьеръ.~~

~~Подозвалъ извощика и велѣлъ ѣхать на Патріаршіе пруды, гдѣ былъ домъ Баздѣева.~~ Изъ всѣхъ дѣлъ предстоявшихъ Пьеру въ это утро дѣло разборки книгъ и бумагъ Іосифа Алексѣевича показалось ~~Пьеру~~ самымъ нужнымъ.

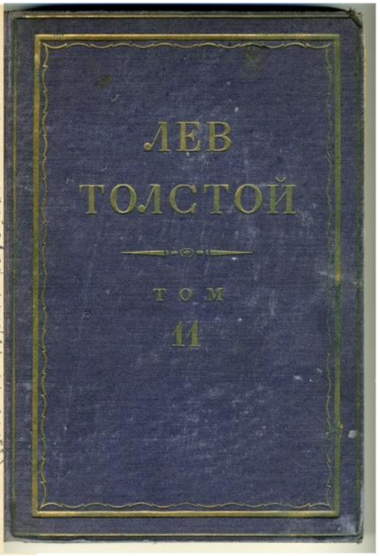
~~Мнѣ уже давно пора было это сдѣлать подумалъ Пьеръ. По крайней мѣрѣ на минутку отдохну отъ всѣхъ этихъ хлопотъ.~~

~~Какъ только онъ въѣхалъ въ узенькіе переулки, ведшіе къ Патріаршимъ прудамъ, Пьеру вдругъ стало безпричинно радостно и весело на душѣ.~~

Безпрестанно оглядываясь на со всѣхъ сторонъ двигавшіеся обозы, выѣзжавшихъ изъ Москвы и оправляясь своимъ тучнымъ тѣломъ, чтобы не соскользнуть съ дребезжа-

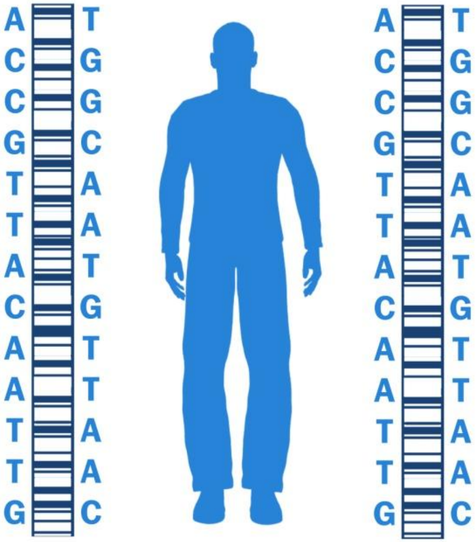
Онъ взялъ перваго попавшагося ему [illegible] и велѣлъ ему ѣхать на Патріаршіе [illegible] гдѣ былъ домъ вдовы Баздѣева.

Последняя корректура страницы 75-й пятого тома первого издания «Войны и мира» 1868 — 1869 гг.
(Текст уже [illegible] к печатному виду).
Размер подлинника.



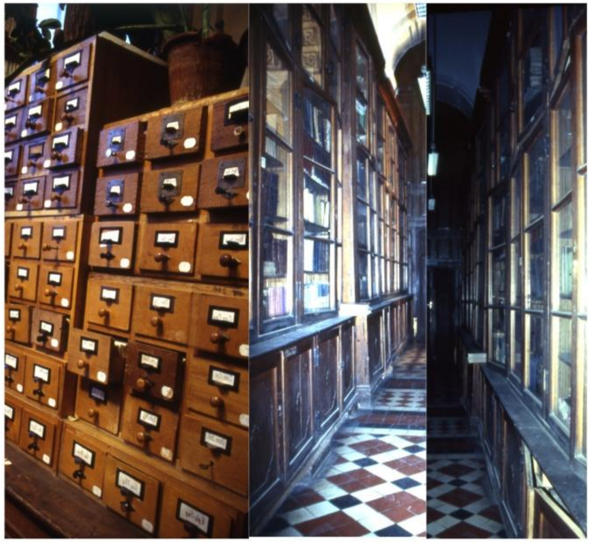
роман Л.Н.Толстого
“Война и Мир”
2 500 000 букв

число букв
в индивидуальном
геноме человека


A C C G T T A C A A T T G
T G G C A A T G T T A A C
A C C G T T A C A A T T G
T G G C A A T G T T A A C

Геном человека
6 000 000 000 букв

число букв в книгах
библиотеки Л.Н.Толстого
в Ясной Поляне

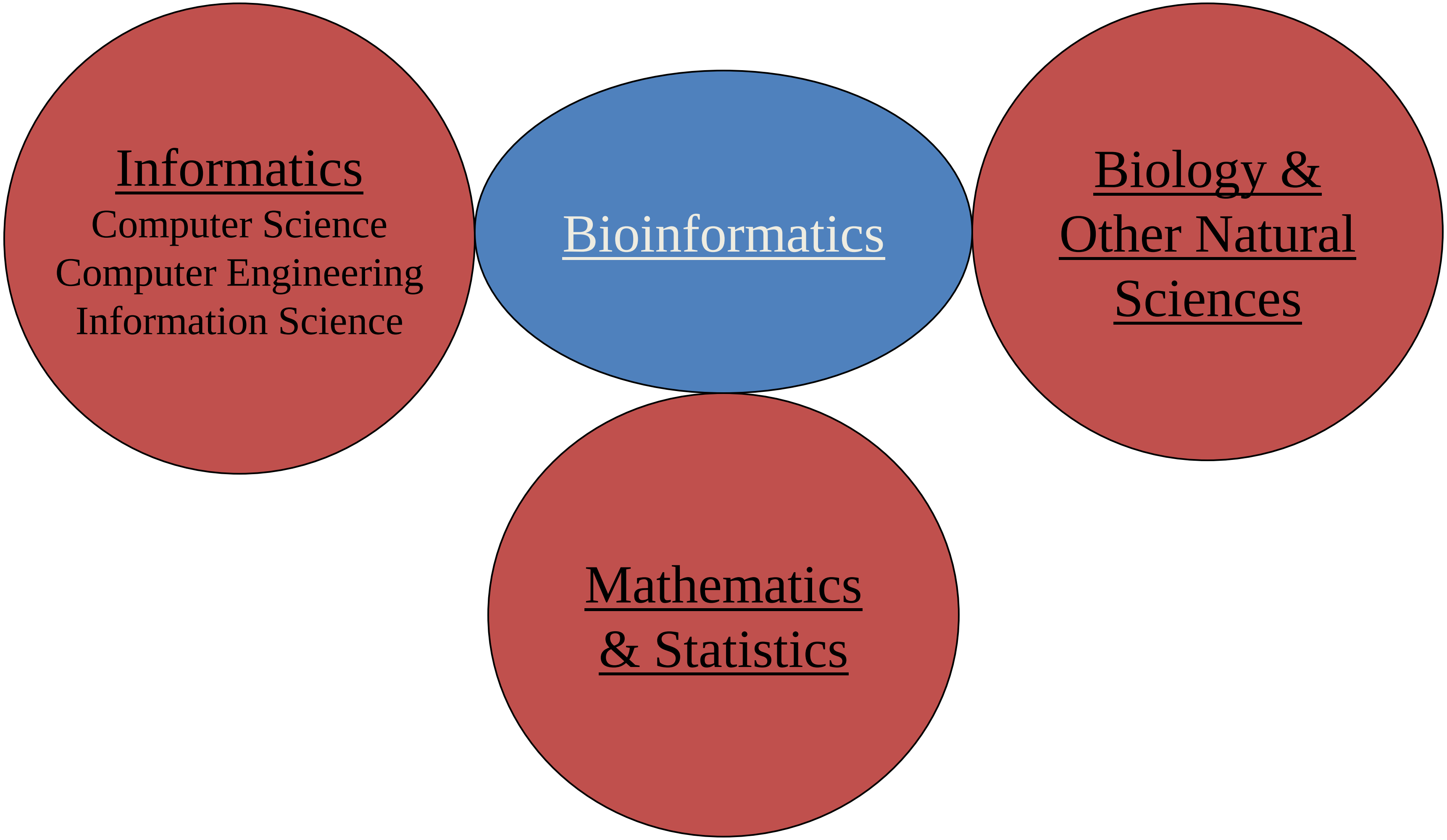


библиотека Л.Н.Толстого
в Ясной Поляне
6 000 000 000 букв

- В мире идет процесс накопления генетической информации: данные биобанков удваиваются в объеме, примерно, каждые 7 месяцев. Лаборатории по расшифровке ДНК получают несколько петабайт секвенированных данных в год (1 терабайт содержит около 1 трлн субъединиц ДНК).

- Во многих странах создание банков биологической информации выливается в проекты национального масштаба, они становятся системообразующими для мировой науки.

What is Bioinformatics?



Informatics
Computer Science
Computer Engineering
Information Science

Bioinformatics

Biology & Other Natural Sciences

Mathematics & Statistics

Bioinformatics Related Fields

- Computational biology
- Computational molecular biology
- Biomolecular informatics
- Computational genomics
- …

Составные части биоинформатики

- 1D и 3D биология
- Разработка биологических баз данных
- Генетические сети и их использование
- Геномика
- Протеомика

Рентгеноструктурный анализ (РСА) макромолекул

Индикаторы качества модели макромолекулы, построенной по данным РСА

Алгоритмы вычисления поверхности макромолекулы

Алгоритмы нахождения гидрофобного ядра молекулы белка

Алгоритмы нахождения структурных доменов белков

Пространственное выравнивание структур белков

Структурные классификации доменов SCOP и CATH

Молекулярная динамика

Biological Data

- Genomes
  - DNA Sequences of A, T, C, G
  - Annotated with function, “interesting” features
- Proteins
  - Amino Acid Sequences
    - Sequences of 20 letters
  - Annotated with structure, function, etc.

Biological Data

- Gene Expression
  - Dynamic behavior of genes
- Protein Expression
  - Dynamic behavior of proteins
- Structural Features
  - RNA and proteins
- …

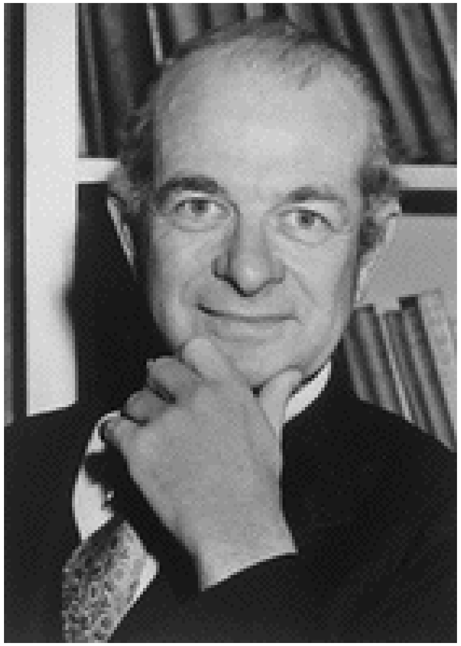
Biological Data

Sus scrofa agouti-related protein gene

```
   1 ggcacattct cctgttgagc caggctatgc tgaccacaat gttgctgagc tgtgccctac
  61 tgctggcaat gcccaccatg ctgggggccc agataggctt ggccccccctg gagggtatcg
 121 gaaggcttga ccaagccttg ttcccagaac tccaaggtca gtgcgggcag gagtgggttg
 181 ggtggggctt ggacatcctc tggccacaaa gtattctgct tgtatgagcc ctttcttccc
 241 cttcccaatc ccaggcctgg gaggtgggtg ttttgtgcat gggtggttct gccctcacat
 301 catctgtccc agatctaggc ctgcagcccc cactgaagag gacaactgca gaacgggcag
 361 aagaggctct gctgcagcag gccgaggcca aggccttggc agaggtaaca gctcagggaa
 421 agggctgagg ccacaagtct tgagtgggtg tgtcaagcat caacctctat ctgtgcttgg
 481 agttgccact gtggtacaac gggattggcg gtgtcttggg agcgctggga cgtggtttca
 541 tccccggcca gcacaagtgg gttaaggatc tggccttgcc atcccttcag cttaggctga
 601 gactgtggct tggagctgat ctctgaccgg aagctccata tgctctgggg tgaccaaaaa
 661 tggaaaaaca aacatacaaa acacctctac ctgcacttcc tgacccccctc acccggggcg
 721 acactgcaga ccatcccgtt cacgctccac ttccatcctg ccttgatctg gcgcattcca
 781 tgaatgtgct tttggaagtc cttgtttccc aacccttgta ggtgctagat cctgaaggac
 841 gcaaggcacg ctccccacgt cgctgcgtaa ggctgcacga atcctgtctg ggacaccagg
 901 taccatgctg cgacccatgt gctacatgct actgccgttt cttcaacgcc ttctgctact
 961 gccgcaagct gggtactgcc acgaacccct gcagccgcac ctagctggcc agccaatgtc
1021 gtcg
```

Пионеры биоинформатики

1962



Лайнус Полинг

- Анализ аминокислотных последовательностей глобинов нескольких позвоночных
- Гипотеза **молекулярных часов**

Zuckerkandl, E., and L. Pauling. **1962**. Molecular disease, evolution, and genic heterogeneity. Horizons in Biochemistry, Academic Press, New York, 189-225.

Zuckerkandl, E., and L. Pauling. **1965**. Evolutionary divergence and convergence in proteins. Evolving Genes and Proteins, Academic Press, New York, 97-166.

Пионеры биоинформатики

1965

Маргарет Дейхофф

- Однобуквенный код аминокислот A,C,D,E,F,G,H…
- Матрицы аминокислотных замен PAM (Point Accepted Mutation)

Атлас последовательностей белков и их структур (1965)

Секвенирование

- ✓ 1977 г. Maxam-Gilbert and Sanger Sequencing
- ✓ 2005 г. Next-Generation Sequencing

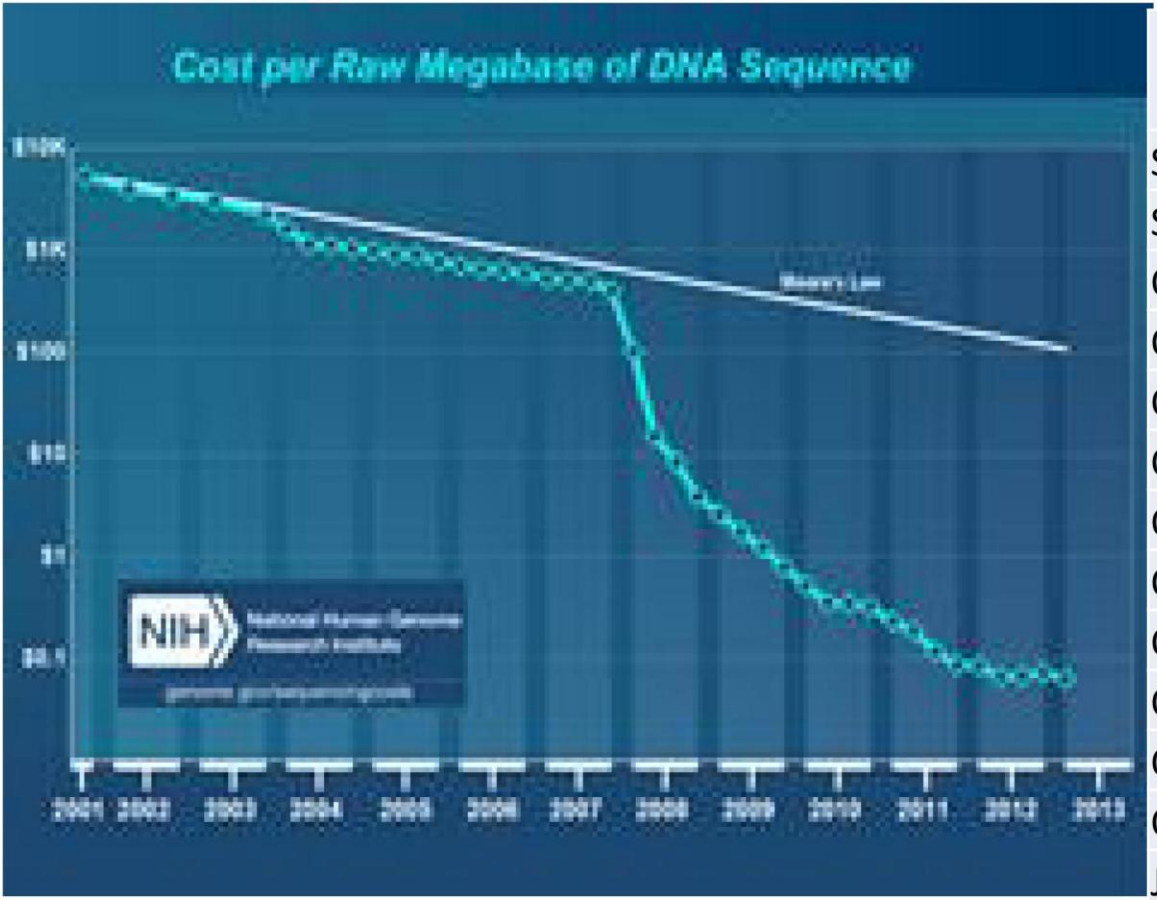
- ➢ Virus – 3222 (Bacteriophage phiX 174, 5386 пн – 1977 г.)
- ➢ Bacteria – 2289 (*Haemophilus influenza,* 1.8 x 10^6 пн – 1995 г.)
- ➢ Eukarya – 168 (*S. cerevisiae* 1.2 x 10^7 пн – 1995 г; *H. sapien*, 3 x 10^9 пн -2001 г.)
- ➢ Archaea – 152 (*Methanococcus jannaschi* , 1.7 x 10^6 nt – 1996 г.)

1953 : Discovery of DNA structure by Watson and Crick

1973 : First sequence of 24 bp published
1977 : Sanger sequencing method published
1980 : Nobel Prize Wally Gilbert and Fred Sanger
1982 : Genbank started
1983 : Development of PCR
1987 : 1st automated sequencer : Applied Biosystems Prism 373

1996 : Capillary sequencer : ABI 310
1998 : Genome of Caenorhabditis elegans sequenced
2000 : Human genome sequenced
2005 : 1st 454 Life Sciences Next Generation Sequencing system : GS 20 System
2006 : 1st Solexa Next Generation Sequencer : Genome Analyzer
2007 : 1st Applied Biosystems Next Generation Sequencer : SOLiD
2009 : 1st Helicos single molecule sequencer : Helicos Genetic Analyser System
2011 : 1st Ion Torrent Next Generation Sequencer : PGM
2011 : 1st Pacific Biosciences single molecule sequencer : PacBio RS Systems
2012 : Oxford Nanopore Technologies demonstrates ultra long single molecule reads

Стоимость секвенирования с развитием технологий NGS



Date	Cost per Mb	Cost per Genome
Sep-01	$5,292.39	$95,263,072
Sep-02	$3,413.80	$61,448,422
Oct-03	$2,230.98	$40,157,554
Oct-04	$1,028.85	$18,519,312
Oct-05	$766.73	$13,801,124
Oct-06	$581.92	$10,474,556
Oct-07	$397.09	$7,147,571
Oct-08	$3.81	$342,502
Oct-09	$0.78	$70,333
Oct-10	$0.32	$29,092
Oct-11	$0.09	$7,743
Oct-12	$0.07	$6,618
Jan-13	$0.06	$5,671

Общий принцип

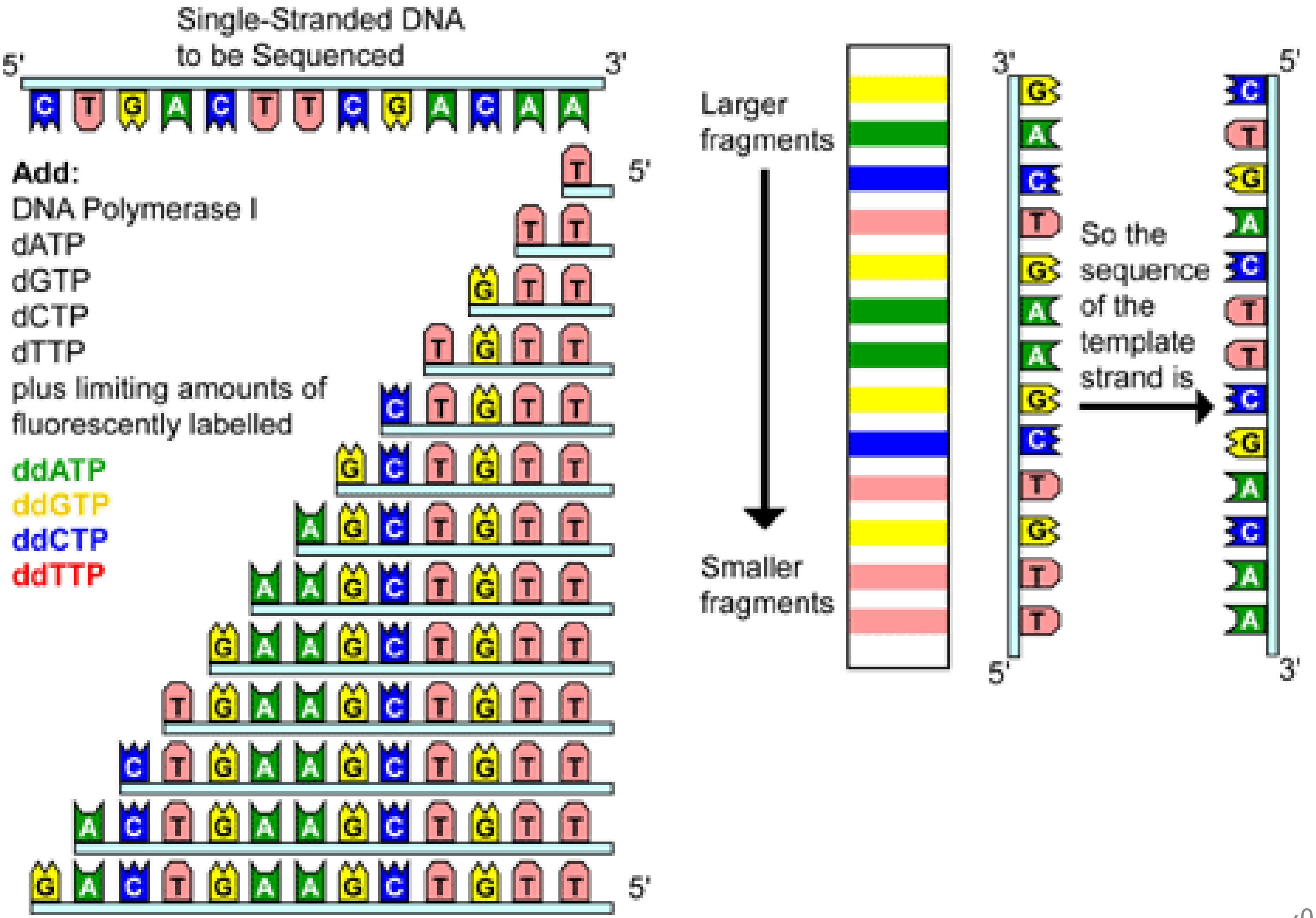
ДНК нарезается на фрагменты определенной длины

К ним лигируются адаптеры

Амплификация каждого отдельного фрагмента в изолированных от других условиях

Анализ последовательности амплифицированных клонов ДНК

A schematic of sequencing



Single-Stranded DNA to be Sequenced

Add:

DNA Polymerase I

dATP

dGTP

dCTP

dTTP

plus limiting amounts of fluorescently labelled

ddATP

ddGTP

ddCTP

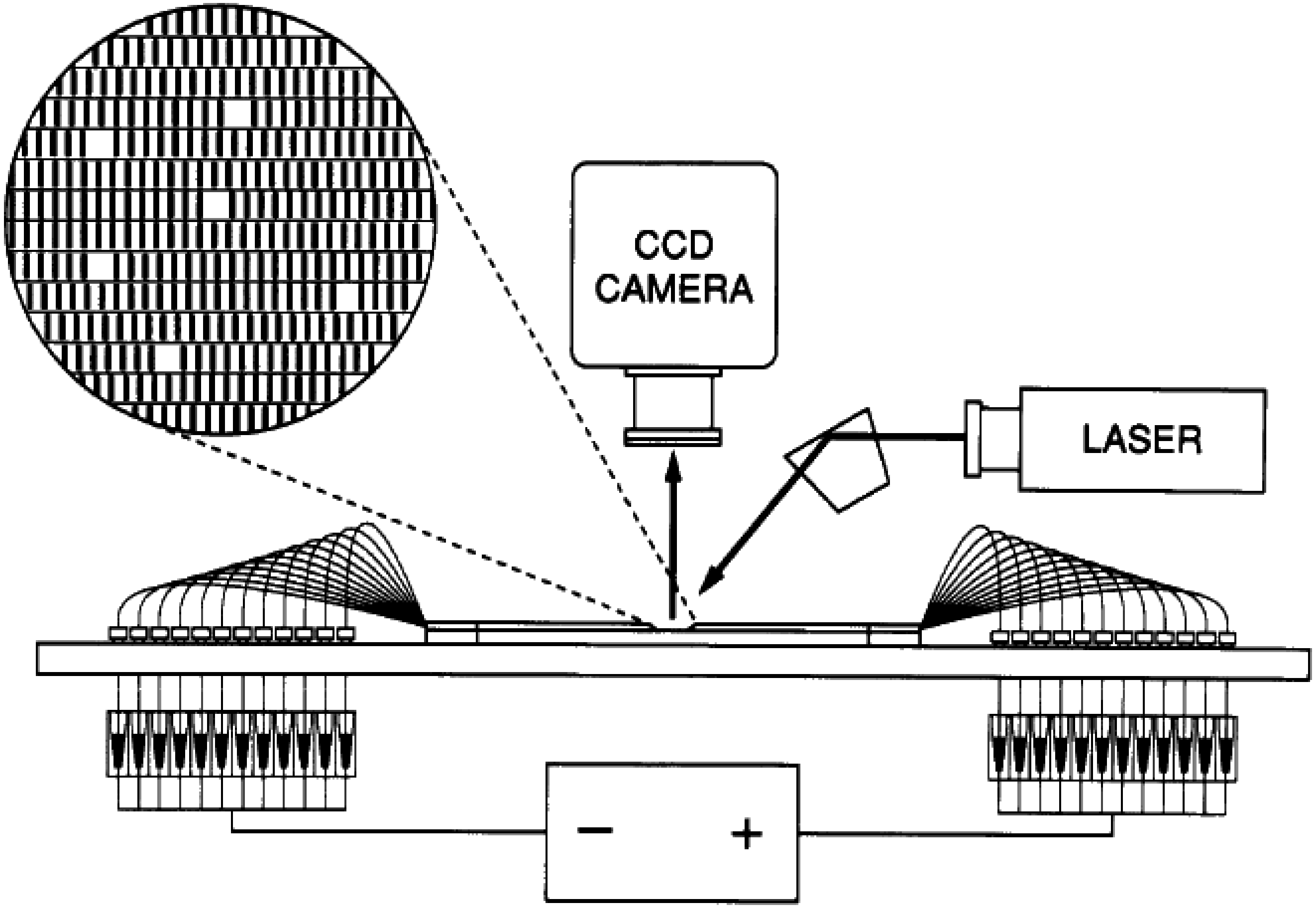
ddTTP

Larger fragments

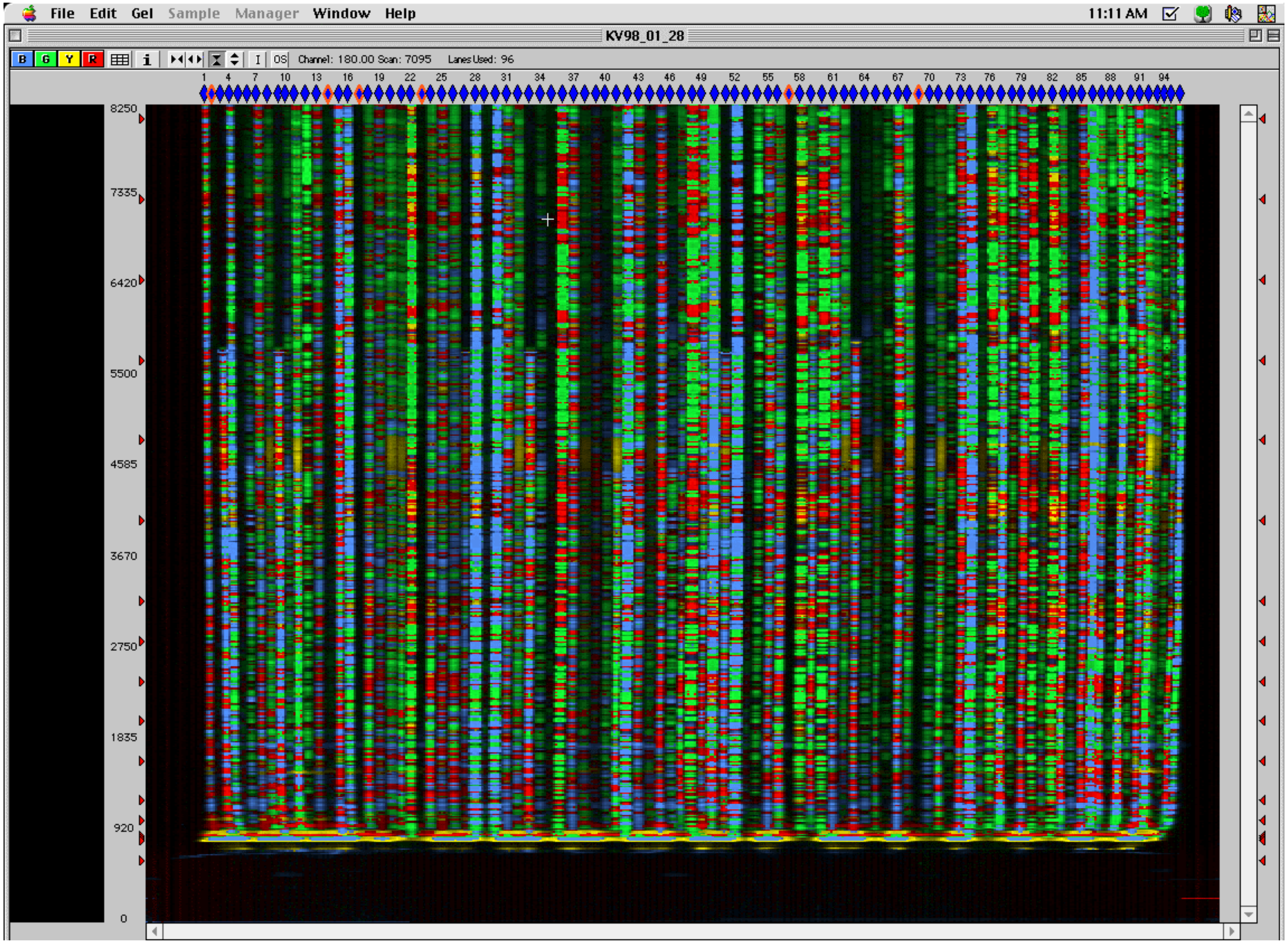
Smaller fragments

So the sequence of the template strand is

Laser Dye Based Sequencing

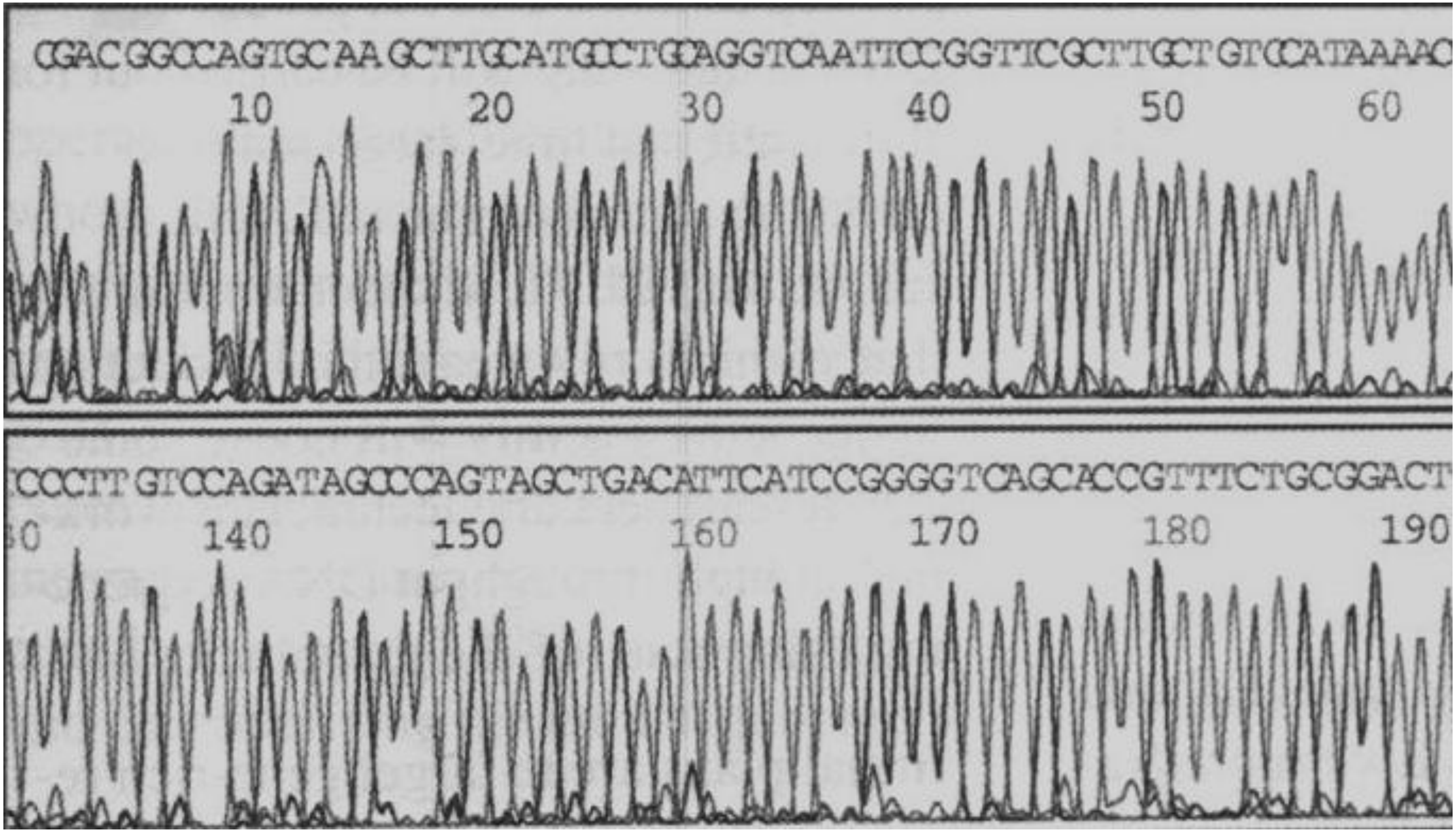


Four-Color Sequencing



Channel: 180.00 Scan: 7095 Lanes Used: 96

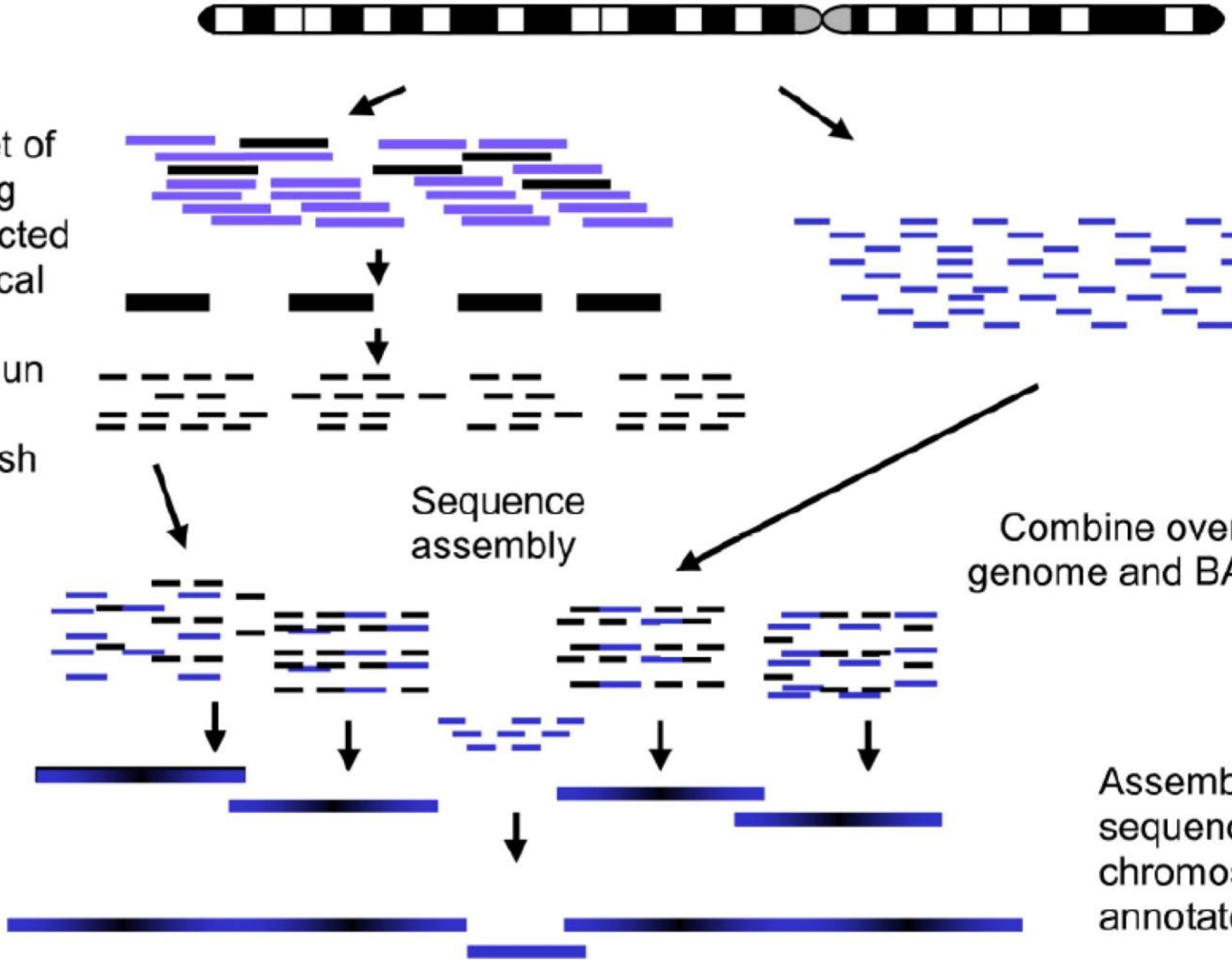
Automated Base Calling



A Biology Lab?

Human Genome Project

C:\Users\Korostin\Desktop\DN/



Minimal set of overlapping BACs selected from physical

BAC shotgun reads
4x - prefinish

Sequence assembly

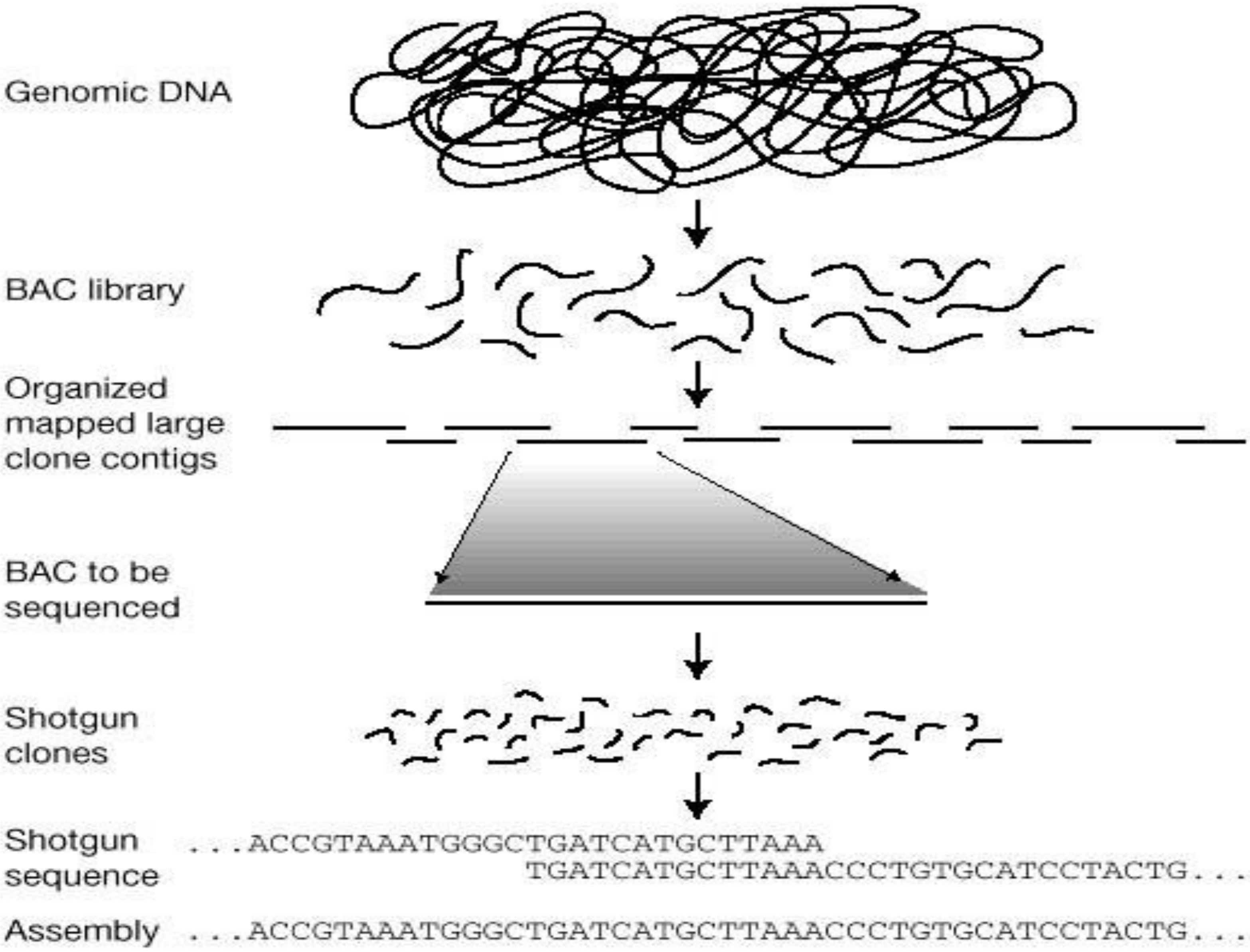
Whole- genome shotgun reads, including next-gen sequence

Combine overlapping whole-genome and BAC-derived reads

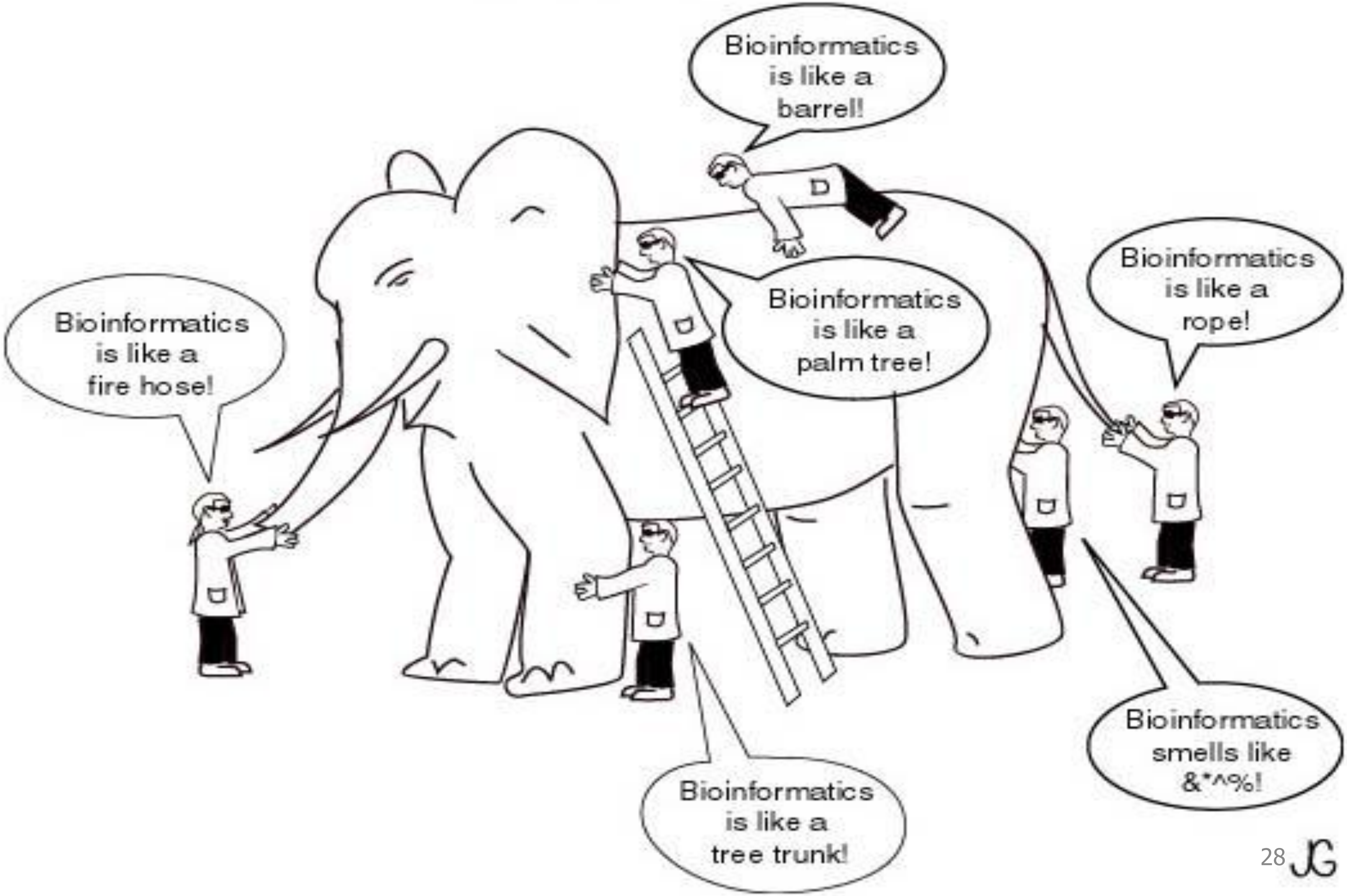
Assemble clone sequences to represent chromosomes and annotate using Ensembl

Human Genome Sequencing

Hierarchical shotgun sequencing


Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

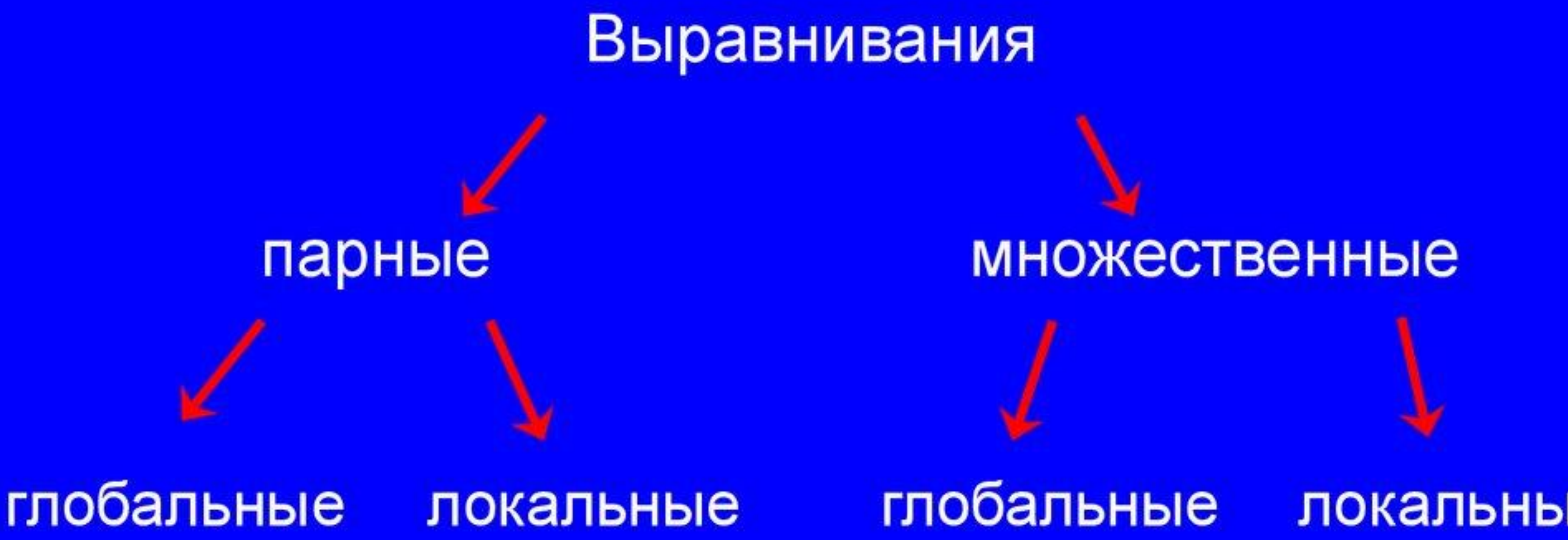
A group of scientists were asked to describe Bioinformatics...


Bioinformatics is like a barrel!
Bioinformatics is like a fire hose!
Bioinformatics is like a palm tree!
Bioinformatics is like a rope!
Bioinformatics is like a tree trunk!
Bioinformatics smells like &*^%!
JG

Fundamental Problems in Bioinformatics

- Pairwise Sequence Alignment
- Multiple Sequence Alignment
- Phylogenetic Analysis
- Sequence Based Database Searches
- Gene Prediction
- Structure Prediction (RNA and Protein)
- Protein Classification
- Gene Expression
- Genetic nets

Какие бывают выравнивания?



Выравнивания

парные

множественные

глобальные

локальные

глобальные

локальны

Почему нам интересует парное сходство последовательностей?

Функцию, структуру и многие свойства белка/ДНК определяет последовательность

Родственные белки имеют похожие свойства. Молекулы, похожие по свойствам, похожи по последовательностям.

Свойсва можно предсказать, если мы найдем последовательности похожие на данную.

Pairwise Sequence Alignment

- Given two DNA or AA sequences, find the best way to “line them up”
  - Biology allows for variation
  - Gaps, mismatches, etc..

```
HEAGAWGHEE
PAWHEAE
```

```
HEAGAWGHE-E
  |  | || |
P-A--W-HEAE
```

```
HEAGAWGHE-E
    || || |
--P-AW-HEAE
```

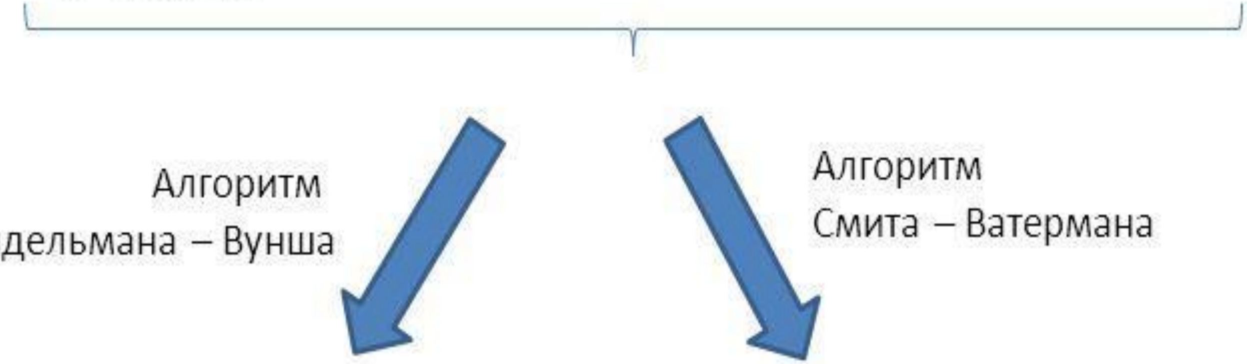
Парное выравнивание: вес

Две последовательности:

```
>P1
ALGTEEIC
>P2
ALGTIAA
```

Параметры:

- матрица замен
- штрафы за пропуски



Алгоритм
Нидельмана – Вунша

Алгоритм
Смита – Ватермана

```
P1 ALGTEEIC-
P2 ALGT--IAA
```

Оптимальное полное выравнивание

```
P1 ALGT
P2 ALGT
```

Оптимальное частичное выравнивание

>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVA
DALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLA
SVSTVLTSKYR

Бэта-2 субъединицей гемоглобина Rattus norvegicus (Серая крыса).
hemoglobin subunit beta-2 [Rattus norvegicus]

Score = 115 bits (288), Expect = 1e-24, Method: Compositional matrix adjust.
Identities = 63/145 (44%), Positives = 87/145 (60%), Gaps = 8/145 (5%)

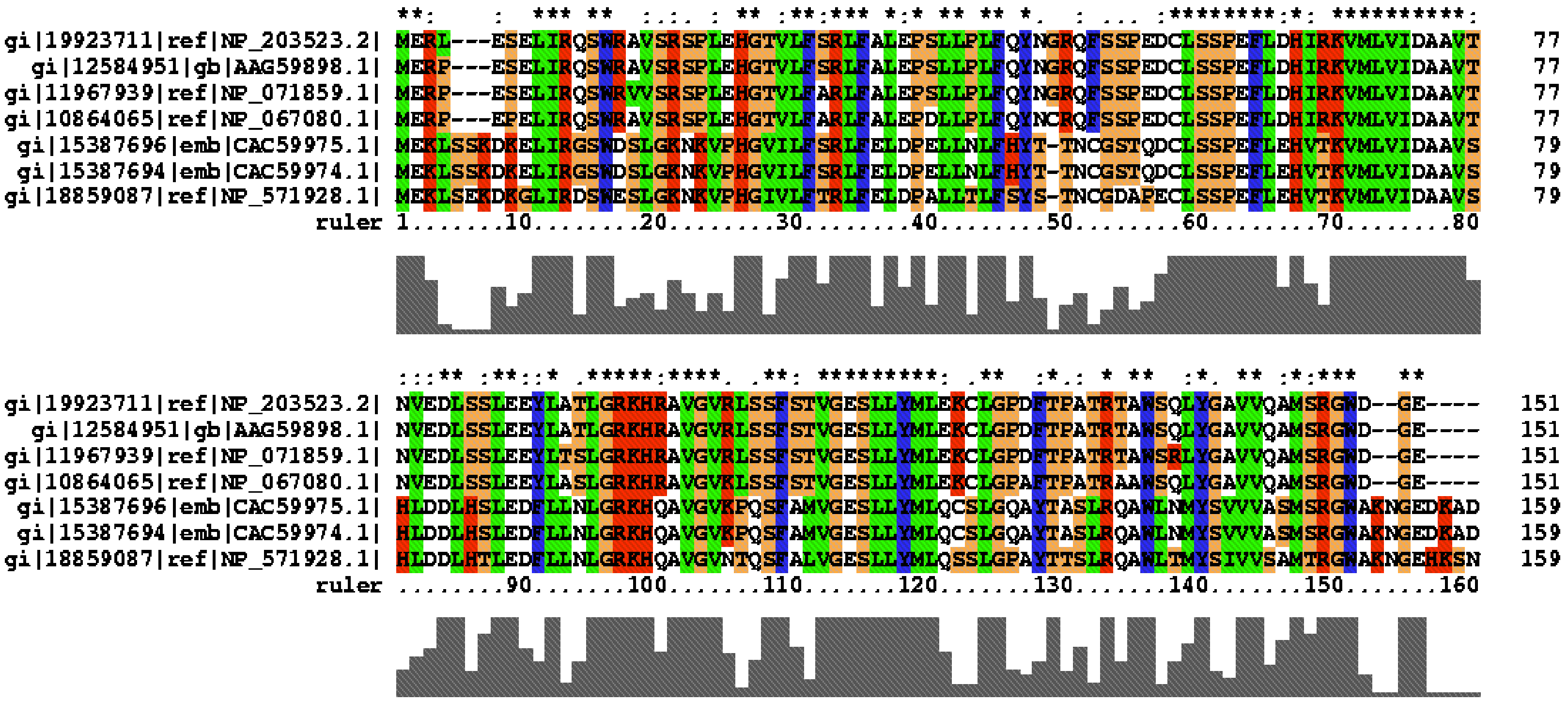
```
Query 3 LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSHGSA-----QV 56
L+ A+K V WGKV +A GAEAL R+ + +P T+ YF F DLS SA QV
Sbjct 4 LTDAEKATVSGLWGKV--NADNVGAEALGRLLVVYPWTQRYFSKFGDLSSASAIMGNPQV 61

Query 57 KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA 116
K HGKKV +A + + H+D++ + LS+LH KL VDP NF+LL + +++ L HL
Sbjct 62 KAHGKKVINAFNDGLKHLDNLKGTFAHLSELHCDKLHVDPENFRLLGNMIVIVLGHHLGK 121

Query 117 EFTPAVHASLDKFLASVSTVLTSKY 141
EFTP A+ K +A V++ L KY
Sbjct 122 EFTPCAQAAFQKVVAGVASALAHKY 1
```

Sequence Based Database Searches

- Keyword
  - Find all sequences named “cytochrome c”
- Sequence
  - Find all sequences similar to `HEAGAWGHEE`
  - Remember, there are gigabytes to search, and I’m not about to wait two days for an answer!
- BLAST, FASTA, …

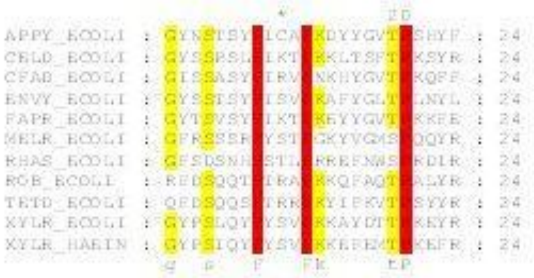
Multiple Sequence Alignment

- Extend pairwise problem to multiple sequences



```
                             **:    :  *** **  :.:. : ** :**:*** *:* ** ** *.  : ... :********:*: **********:
gi|19923711|ref|NP_203523.2| MERL---ESELIRQSWRAVSRSPLEHGTVLFSRLFALEPSLLPLFQYNGRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVT    77
  gi|12584951|gb|AAG59898.1| MERP---ESELIRQSWRAVSRSPLEHGTVLFSRLFALEPSLLPLFQYNGRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVT    77
gi|11967939|ref|NP_071859.1| MERP---ESELIRQSWRVVSRSPLEHGTVLFARLFALEPSLLPLFQYNGRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVT    77
gi|10864065|ref|NP_067080.1| MERP---EPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVT    77
 gi|15387696|emb|CAC59975.1| MEKLSSKDKELIRGSWDSLGKNKVPHGVILFSRLFELDPELLNLFHYT-TNCGSTQDCLSSPEFLEHVTKVMLVIDAAVS    79
 gi|15387694|emb|CAC59974.1| MEKLSSKDKELIRGSWDSLGKNKVPHGVILFSRLFELDPELLNLFHYT-TNCGSTQDCLSSPEFLEHVTKVMLVIDAAVS    79
gi|18859087|ref|NP_571928.1| MEKLSEKDKGLIRDSWESLGKNKVPHGIVLFTRLFELDPALLTLFSYS-TNCGDAPECLSSPEFLEHVTKVMLVIDAAVS    79
                       ruler 1.......10.......20.......30.......40.......50.......60.......70.......80

                             :::** :**::* .*****:****. .**: *********: .**  :*.: * **  :*. ** :*:***   **
gi|19923711|ref|NP_203523.2| NVEDLSSLEEYLATLGRKHRAVGVRLSSFSTVGESLLYMLEKCLGPDFTPATRTAWSQLYGAVVQAMSRGWD--GE----   151
  gi|12584951|gb|AAG59898.1| NVEDLSSLEEYLATLGRKHRAVGVRLSSFSTVGESLLYMLEKCLGPDFTPATRTAWSQLYGAVVQAMSRGWD--GE----   151
gi|11967939|ref|NP_071859.1| NVEDLSSLEEYLTSLGRKHRAVGVRLSSFSTVGESLLYMLEKCLGPDFTPATRTAWSRLYGAVVQAMSRGWD--GE----   151
gi|10864065|ref|NP_067080.1| NVEDLSSLEEYLASLGRKHRAVGVKLSSFSTVGESLLYMLEKCLGPAFTPATRAAWSQLYGAVVQAMSRGWD--GE----   151
 gi|15387696|emb|CAC59975.1| HLDDLHSLEDFLLNLGRKHQAVGVKPQSFAMVGESLLYMLQCSLGQAYTASLRQAWLNMYSVVVASMSRGWAKNGEDKAD   159
 gi|15387694|emb|CAC59974.1| HLDDLHSLEDFLLNLGRKHQAVGVKPQSFAMVGESLLYMLQCSLGQAYTASLRQAWLNMYSVVVASMSRGWAKNGEDKAD   159
gi|18859087|ref|NP_571928.1| HLDDLHTLEDFLLNLGRKHQAVGVNTQSFALVGESLLYMLQSSLGPAYTTSLRQAWLTMYSIVVSAMTRGWAKNGEHKSN   159
                       ruler ........90.......100.......110.......120.......130.......140.......150.......160
```

Для чего строят множественные выравнивания?



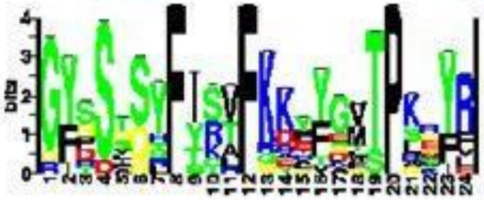
```
                           *         20
APPY_ECOLI : GYNSTSYFICAFKDYYGVTPSHYF : 24
CELD_ECOLI : GYSSPSLFIKTFKKLTSFTPKSYR : 24
CFAB_ECOLI : GISSASYFIRVFNKHYGVTPKQFF : 24
ENVY_ECOLI : GYSSTSYFISVFKAFYGLTPLNYL : 24
FAPR_ECOLI : GYTSVSYFIKTFKEYYGVTPKKFE : 24
MELR_ECOLI : GFKSSERFYSTFGKYVGMSPQQYR : 24
RHAS_ECOLI : GFSDSNHFSTLFKREFNWSPKDIR : 24
ROB_ECOLI  : REDSQQTFIRAFKKQFAQTPALYR : 24
TETD_ECOLI : QEDSQQSFTRKFKYIFKVTPSYYR : 24
XYLR_ECOLI : GYPSLQYFYSVFKKAYDTTPKEYR : 24
XYLR_HAEIN : GYPSIQYFYSVFKKEFEMTPKEFR : 24
             g  s   F   FK     tP
```

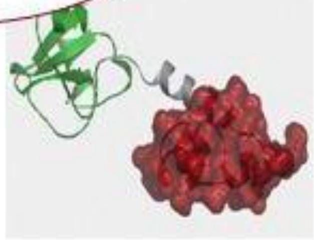
позволяет найти общее

мотивы, паттерны, профили

поиск активного центра

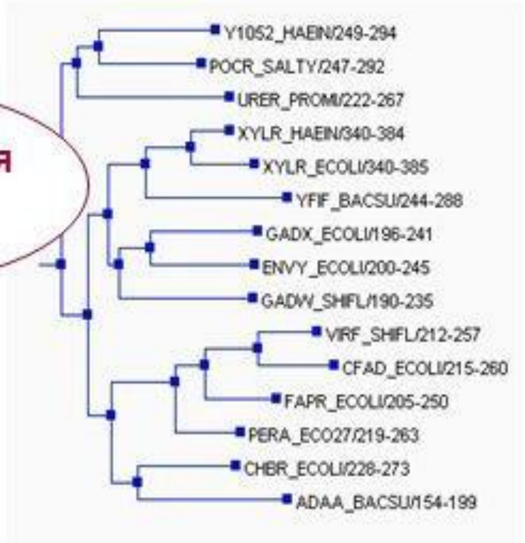
предсказание 3D-структуры





позволяет оценить эволюционные отношения

реконструкция эволюции



Построение множественных выравниваний — необходимый этап решения многих задач молекулярной биологии

Множественное выравнивание

Можно определить вес (хотя ситуация со штрафом за пропуски сложнее)

Но не существует приемлемого алгоритма, гарантирующего нахождение оптимального по данному весу выравнивания

Аналог алгоритма Нидельмана – Вунша имеет приемлемое время работы лишь для очень малого числа последовательностей (до 4–5)

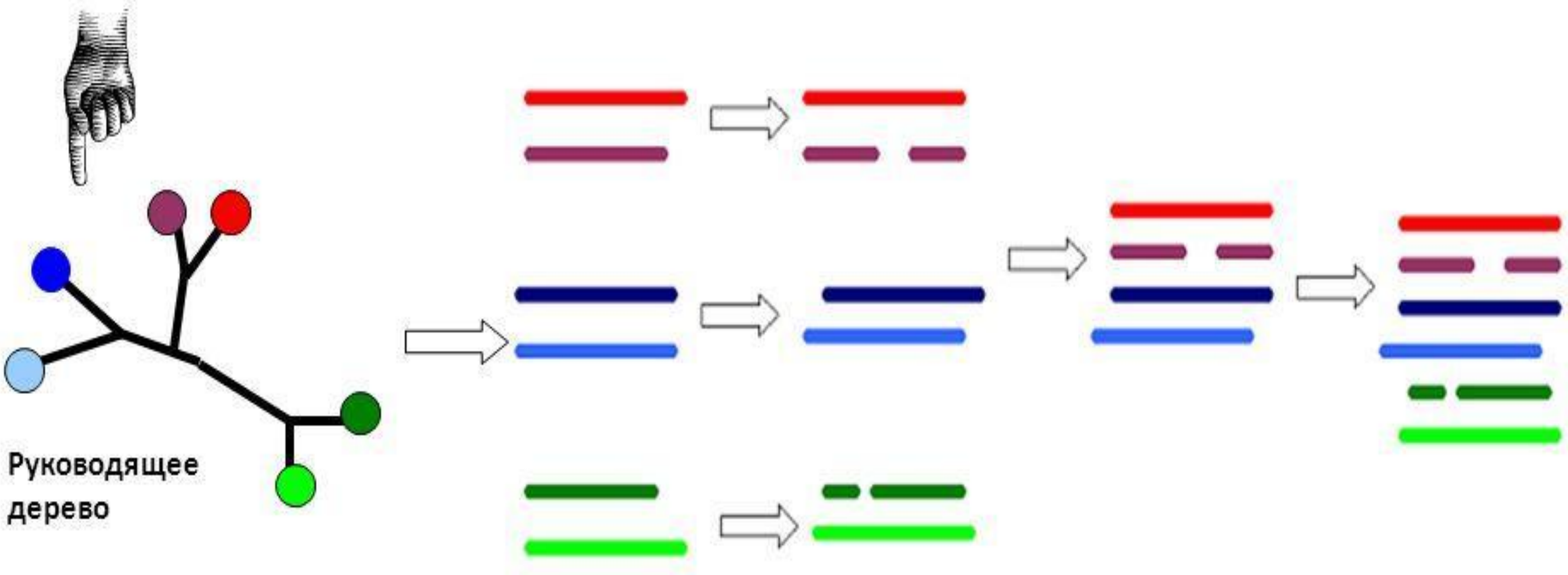
Программы множественного выравнивания

- **ClustalW** – к настоящему времени явно устарела, но по-прежнему очень популярна
- **Muscle** – пожалуй, на текущий момент программа первого выбора
- **MAFFT** – тоже очень популярная программа
- **DiAlign**
- **T-Coffee**
- **Kalign**
- **ProbCons**

Всё это программы **полного** выравнивания

Единственная популярная программа частичного множественного выравнивания – **MEME** (ищет блоки, то есть выравнивания без пропусков)

Алгоритм ClustalW – пример эвристического прогрессивного алгоритма



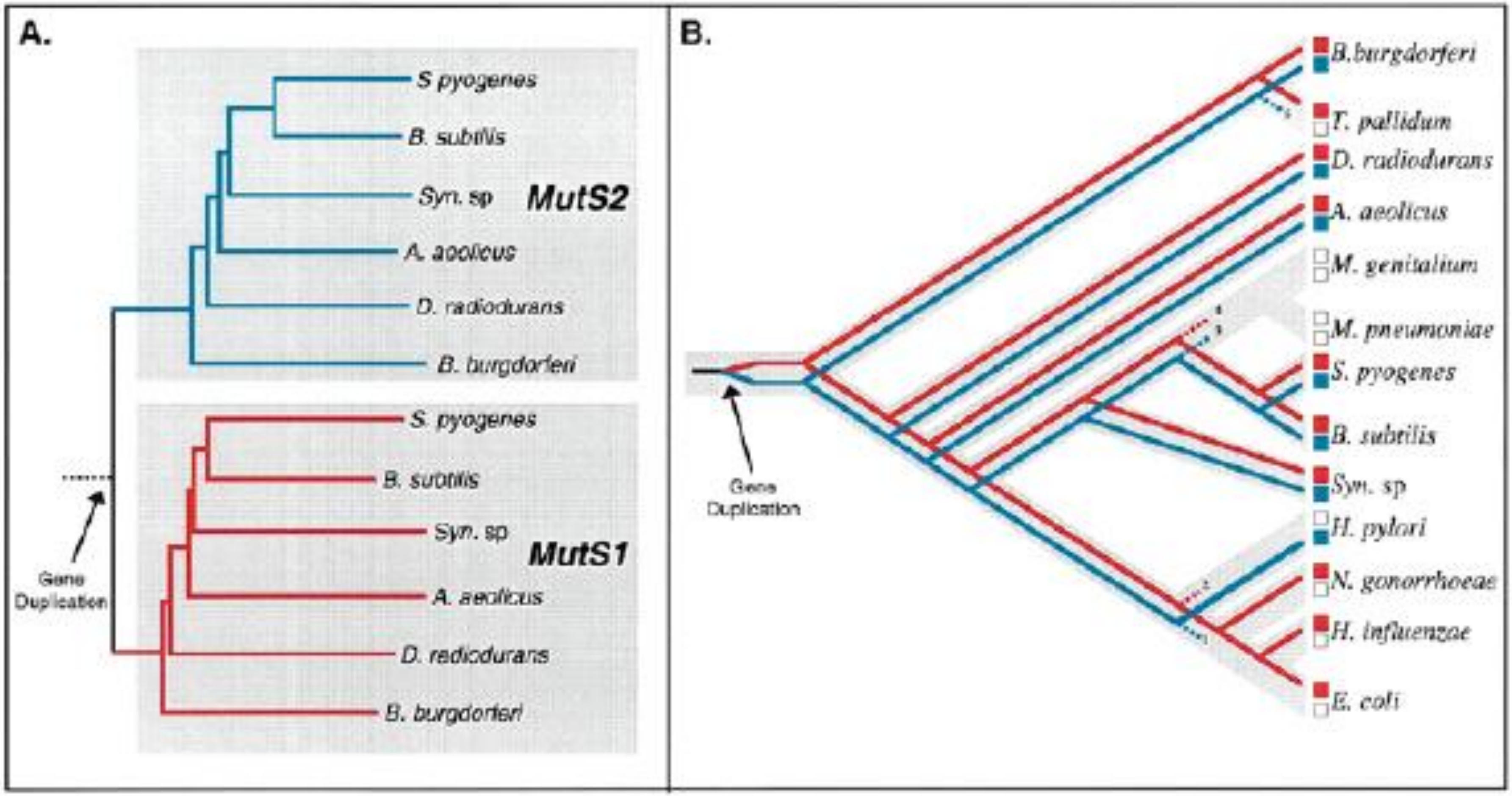
Очевидные недостатки :

- результат зависит от порядка выравниваний;
- «один раз гэп – навсегда гэп»

Phylogenetic Analysis

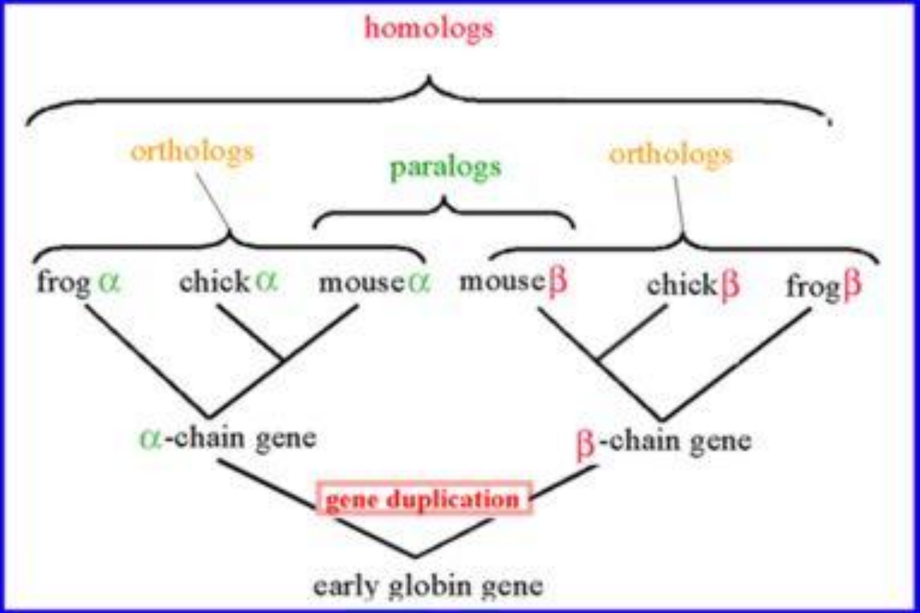
- Study relationships between organisms
  - Characteristic similarity
  - Sequence similarity
  - Whole genome comparison
  - …

Phylogenetic Analysis



Ортологи — последовательности, возникшие из одного общего предшественника в процессе видообразования. Ортологи, как правило, имеют одну и ту же функцию

Паралоги — последовательности, возникшие из одного общего предшественника в результате дупликации одного гена в одном организме. Паралоги, как правило, имеют разные функции.



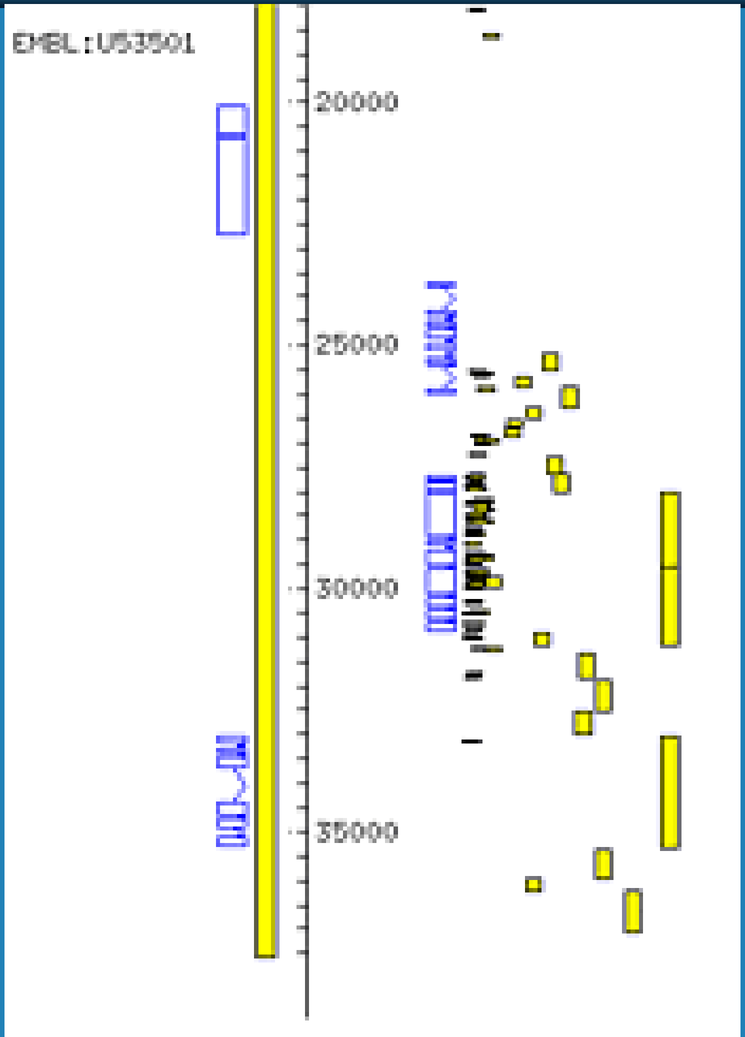
Gene Prediction

- Does the following sequence contain a gene?

```
TTGTAATCTCCTCTGTGACTATAATGACTAGTCTCAGGCCTGCCTTCCCCAGAAACCTCTCTTTTGGCTATTTCTCTTTC
TAGTTCTCTGTTTAAACAAAATTTATTCTATATATCTATCTATCTGTCTATCTATCTATCTATCTATCTATCTATCTATC
TATCTATCTATCTATCATCTACTTATCATCTGTCTAGCCATTTGAAGCATCTTTGTGTTTTAGGTCCTGTTAGATTCTCC
TTTCAGCCAGTGGAGGATCTGGACAGAGCTATTTCTTAGCTTCCCCTAAGCCATGTTGTTAGAACGAATCCCCCACACCT
CCTCTGAGTGCTACGTCTCCGTCAAGAATTATGTATGTGGGATCCAGATGGCCCAGTGGATAAAACTGCAAGTGTCATGA
CCATGACCTGACTTCAAGGGATTGTGTAGAAAGGGAGTTATCACAGTGTGAGGGACAGGGCTAAGGACACTAACCCGTAT
GTTGAGGGGCACAGACGCTAGCAACAACAGTGAAGTGTTTAAAAAGGCAAAAATCATGTTTCTAGAAGTCAGGAAGAGCC
TAACTTGTGGACAAGGACCAACAGGCAGCAGTTGTAATGGGGCAGGGCAGAGGGAGAGCGGACACGCAGCTTTTGGCATC
AAACACACCCAGAGTGTGGATAGAGAGTAGGGAAATACTCTAGTCTCTGGCTAGGATACTCCCCTCTCTTTTTGACATTT
CTCATTGGCAGCCCCAAGTGGTCACTGGAGAGCCAGGAAGCCTAAAGGACACAGTTAGTAGCAGCCAGCTCCTTTGGTGG
AATTTTGGGGACATGGTGGGGTGACTTGGCTCTATCCAGGCCAGGGCTGGGTGTGAGTATACACTTAGTGACTGGCCTTC
```

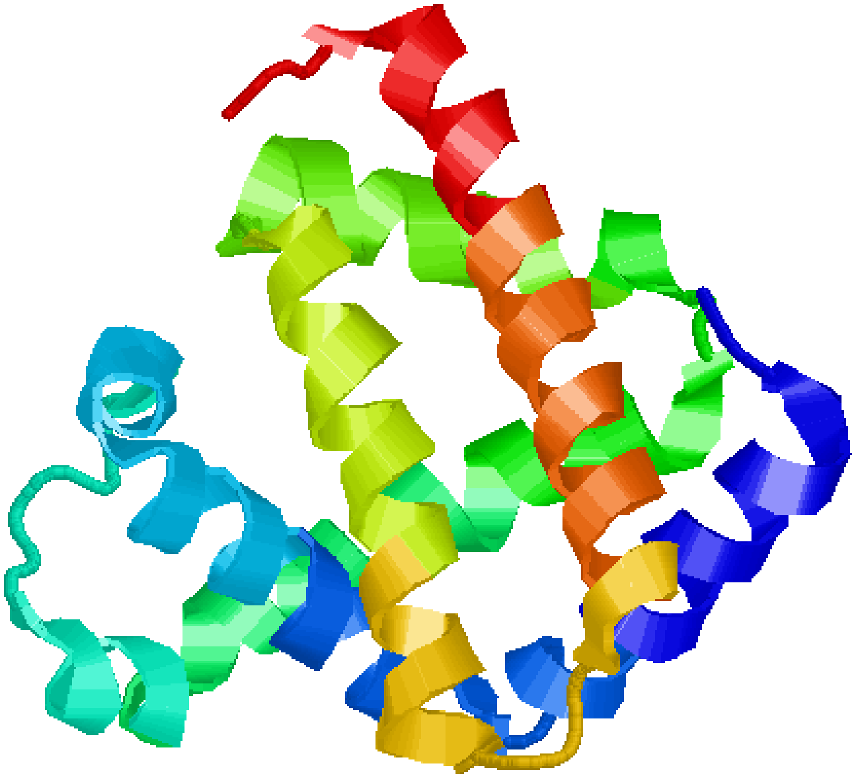
- How many introns? Exons? Promoters? Other features?

Genome annotation



Structure Prediction (RNA, Protein)

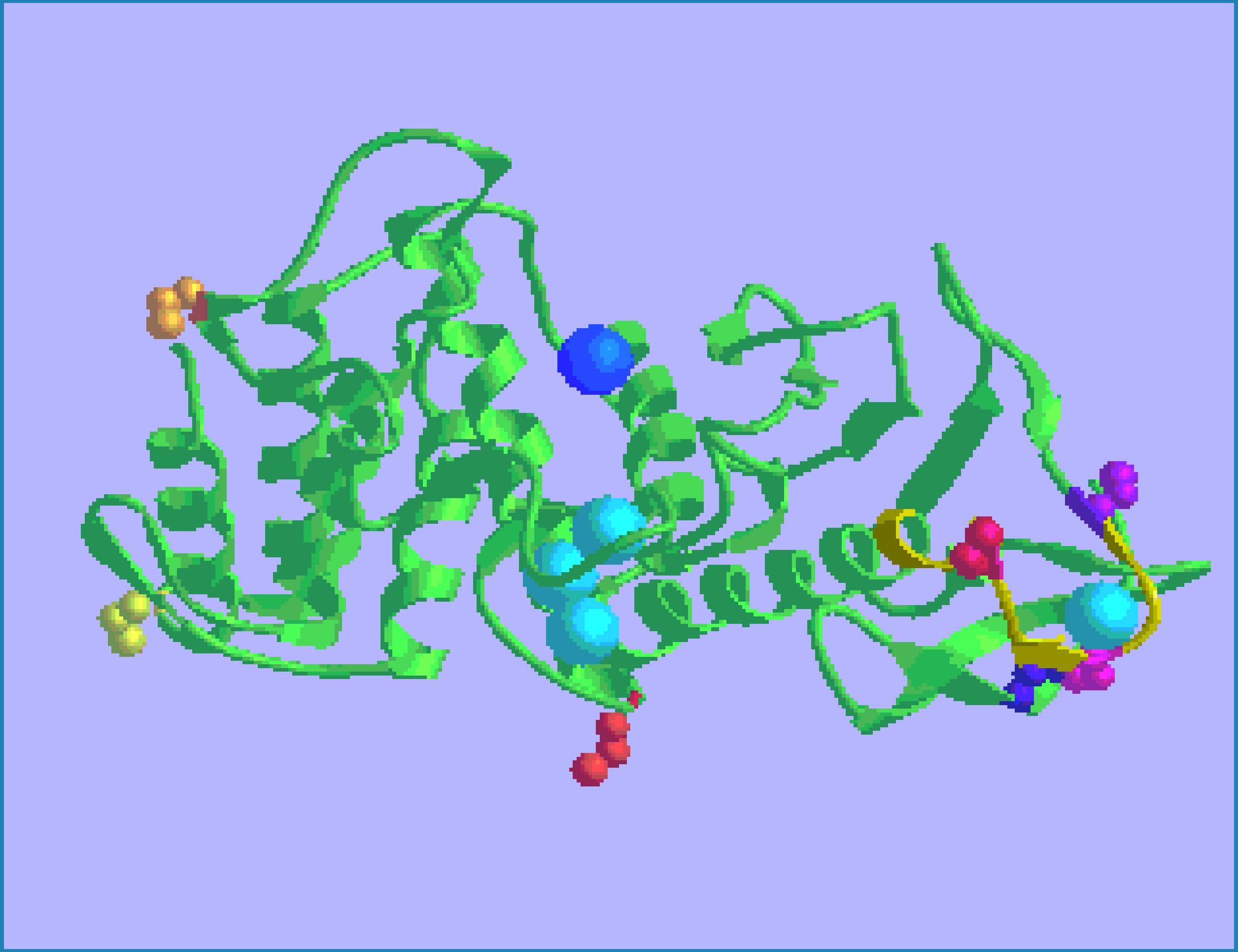
- From sequence, predict 2 and 3D structures.



Protein Classification

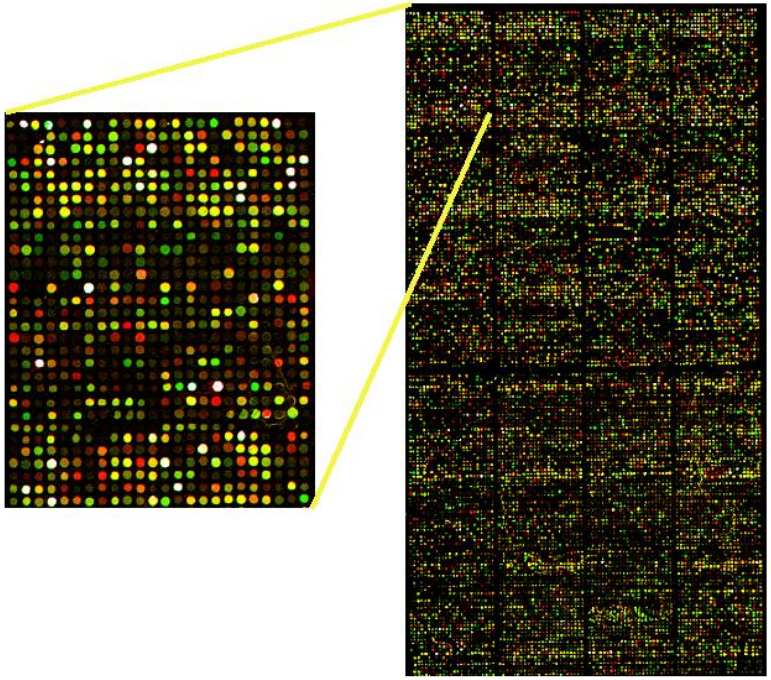
- From sequence, identify characteristics of a protein
  - Active sites
  - Families (e.g. globin)
  - Blocks
  - Domains
  - Folds
  - Motifs
  - Etc.

Protein engineering



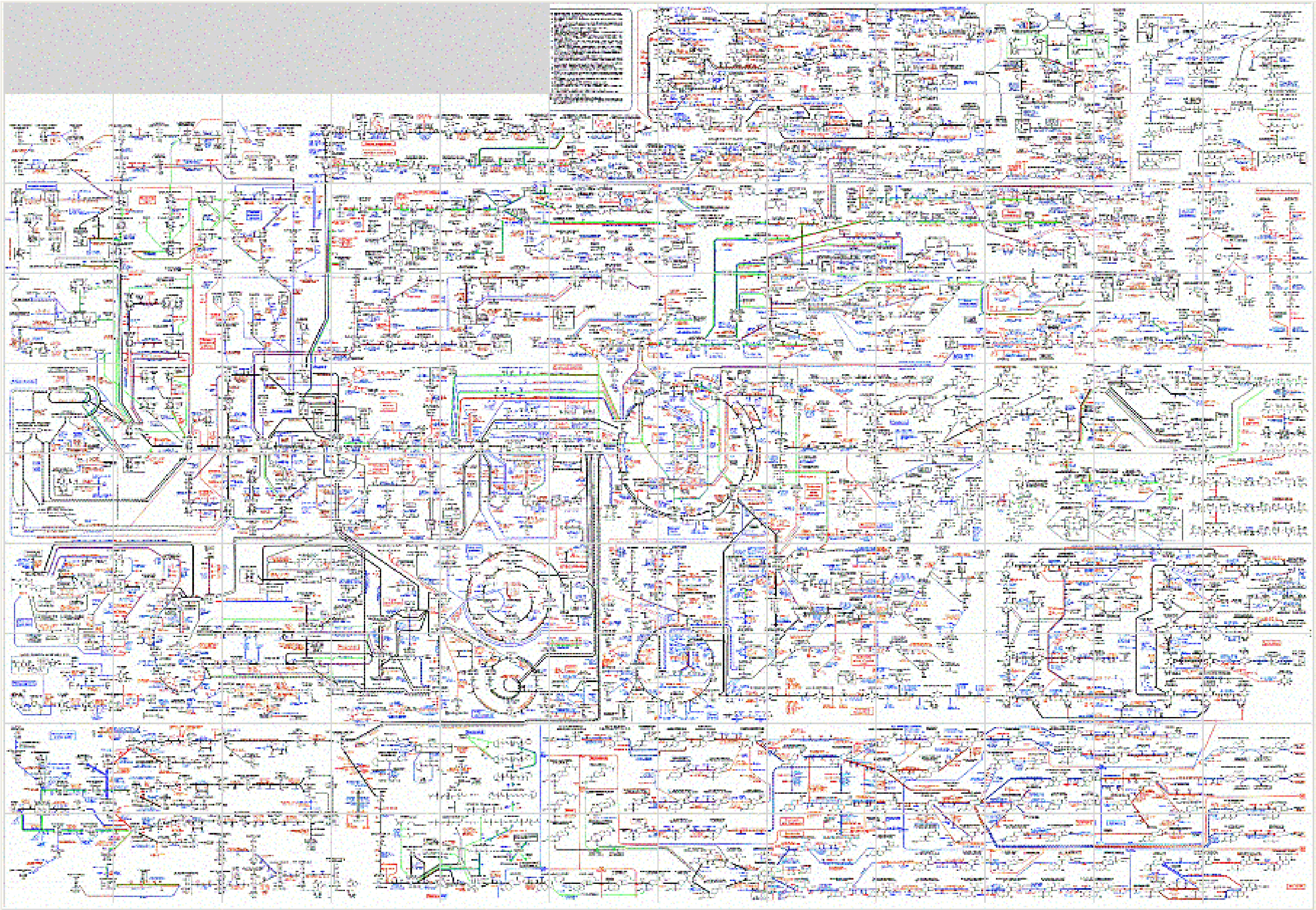
Gene Expression

- Study of gene activity under experimental conditions
  - Large scale studies with microarrays



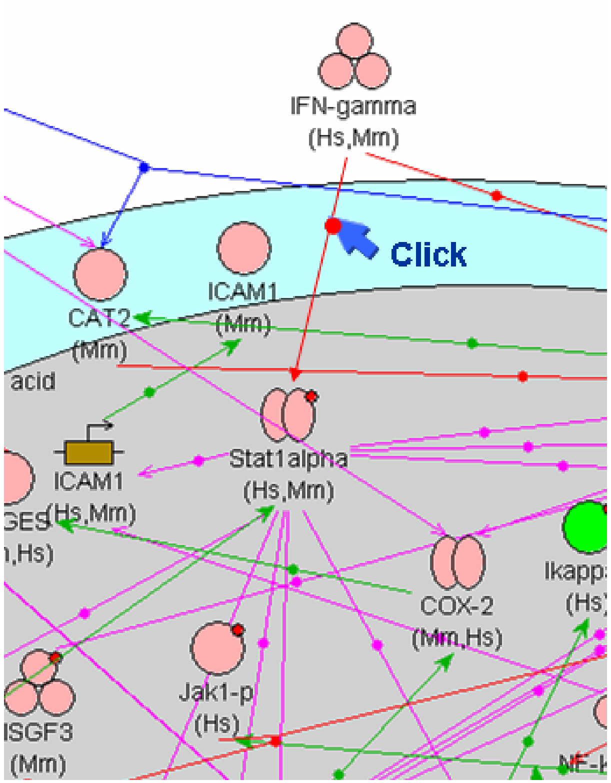
Фрагмент одной из карт метаболических путей.

Современная биология стала источником огромных объемов экспериментальной информации, осмысливание которых невозможно без использования эффективных информационных технологий и методов математического моделирования

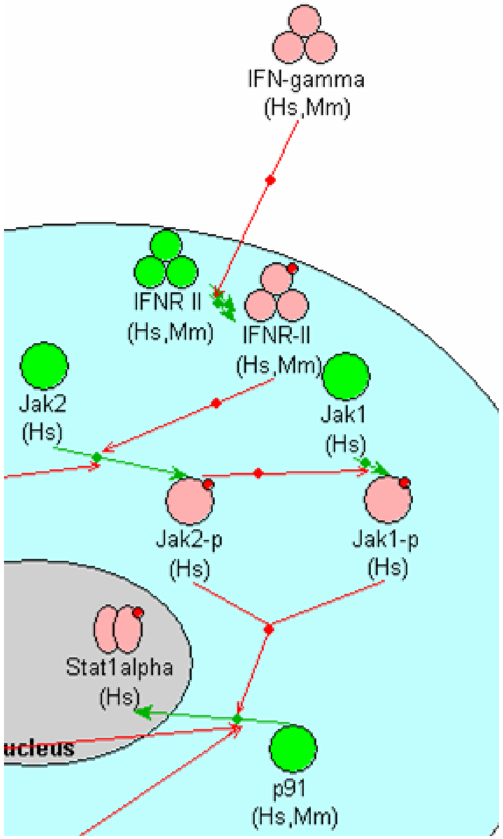


Компьютерная технология формализованного описания, конструирования и визуализации генных сетей

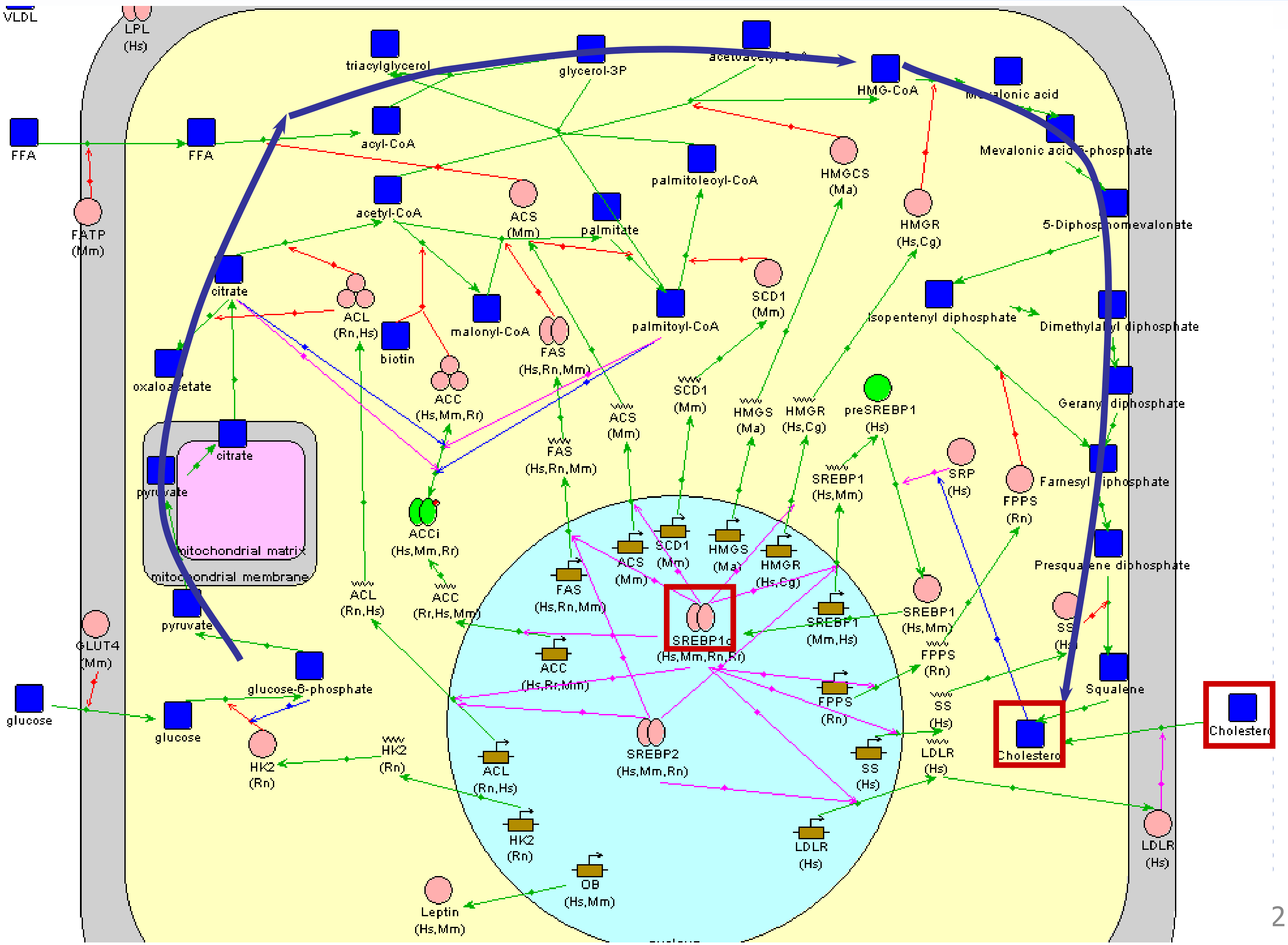
Gene network
"Macrophage activation"
(E. Nedosekina, E. Ananko)



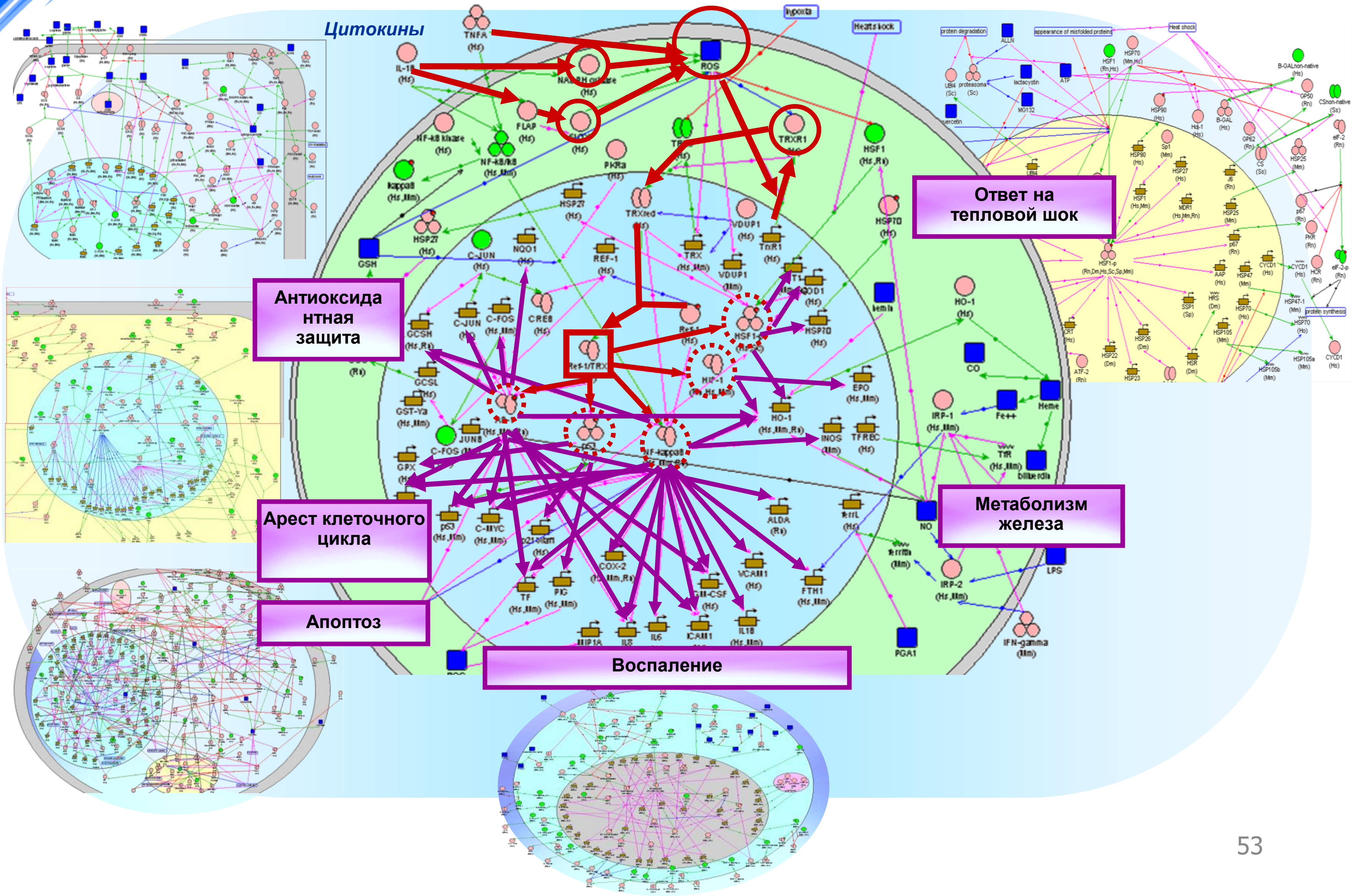
Subscheme "Jak-Stat signal transduction payhway" *E. Nedosekina, E. Ananko)*



МЕТАБОЛИЧЕСКИЕ ПУТИ – ОБЯЗАТЕЛЬНЫЕ ЭЛЕМЕНТЫ ГЕННЫХ СЕТЕЙ. Адипоцит: мевалонатный путь биосинтеза холестерина в клетке.



Интеграция генных сетей при противовоспалительном ответе



Активные формы кислорода

Цитокины

Ответ на тепловой шок

Антиоксидантная защита

Метаболизм железа

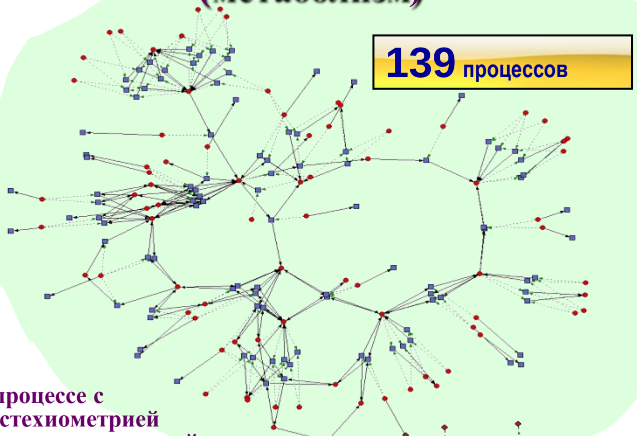
Арест клеточного цикла

Апоптоз

Воспаление

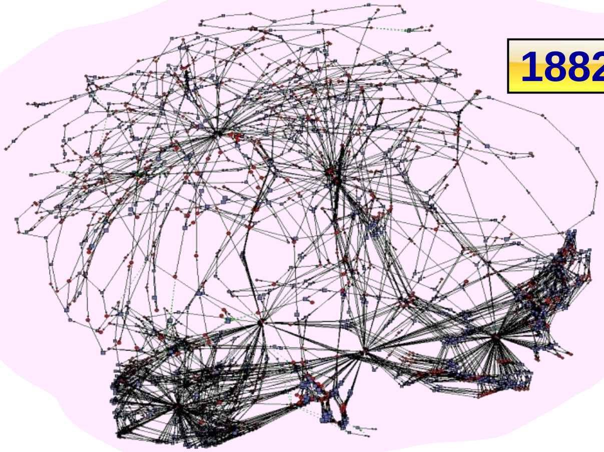
Соотношение метаболической и регуляторной компонент цикла трикарбоновых кислот E. Coli K-12:

Исполняющая компонента (метаболизм)



139 процессов

Регуляторная компонента (управление метаболизмом)

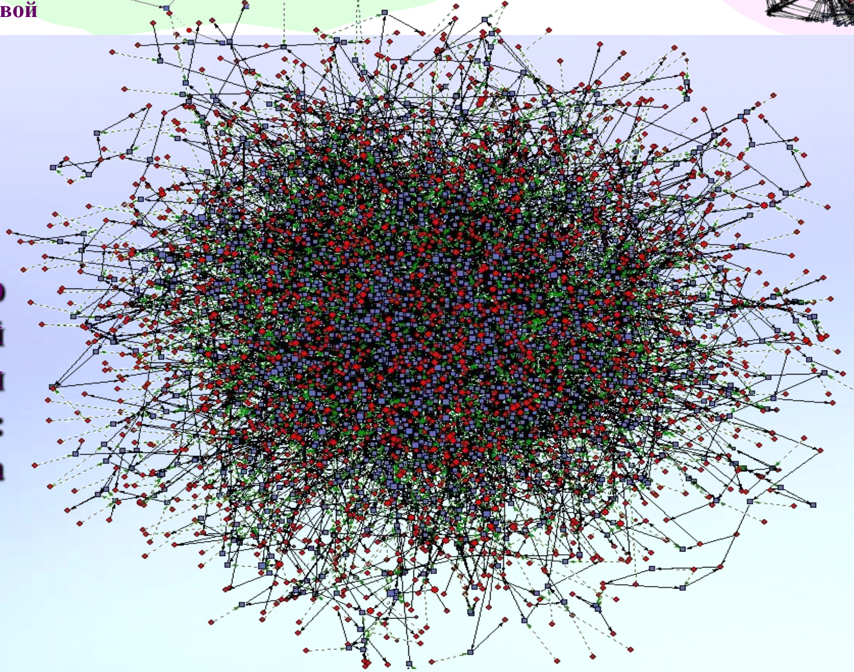


1882 процессов

■ - ПРОЦЕСС

→ - участие в процессе с ненулевой стехиометрией

- - → - участие в процессе с нулевой стехиометрией



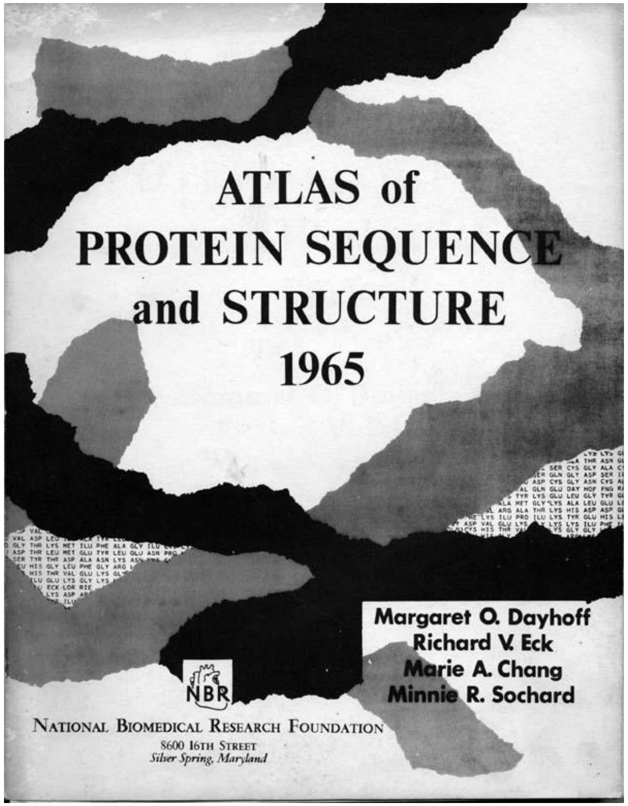
Полный граф метаболической компоненты E. COLI K-12: 3973 процесса

Нижние оценки сложности модели (без детального учета этапов матричного биосинтеза):
~ 60 000 – 100 000 процессов

Более детальная модель:
~ 1 000 000 процессов

Портретная модель:
не менее 10 000 000 процессов

Первый “банк данных”



ATLAS of
PROTEIN SEQUENCE
and STRUCTURE
1965

Margaret O. Dayhoff
Richard V. Eck
Marie A. Chang
Minnie R. Sochard

NBR

NATIONAL BIOMEDICAL RESEARCH FOUNDATION
8600 16TH STREET
Silver Spring, Maryland

1965 -1978

Атлас белковых последовательностей и их структур

Первая версия атласа содержала описание **65 !** последовательностей белков

Genome Sizes

Species	Genome Size
Bacteriophage MS2	3569 bp
Esherichia coli	4.7 million bp
Human	3.3 billion bp

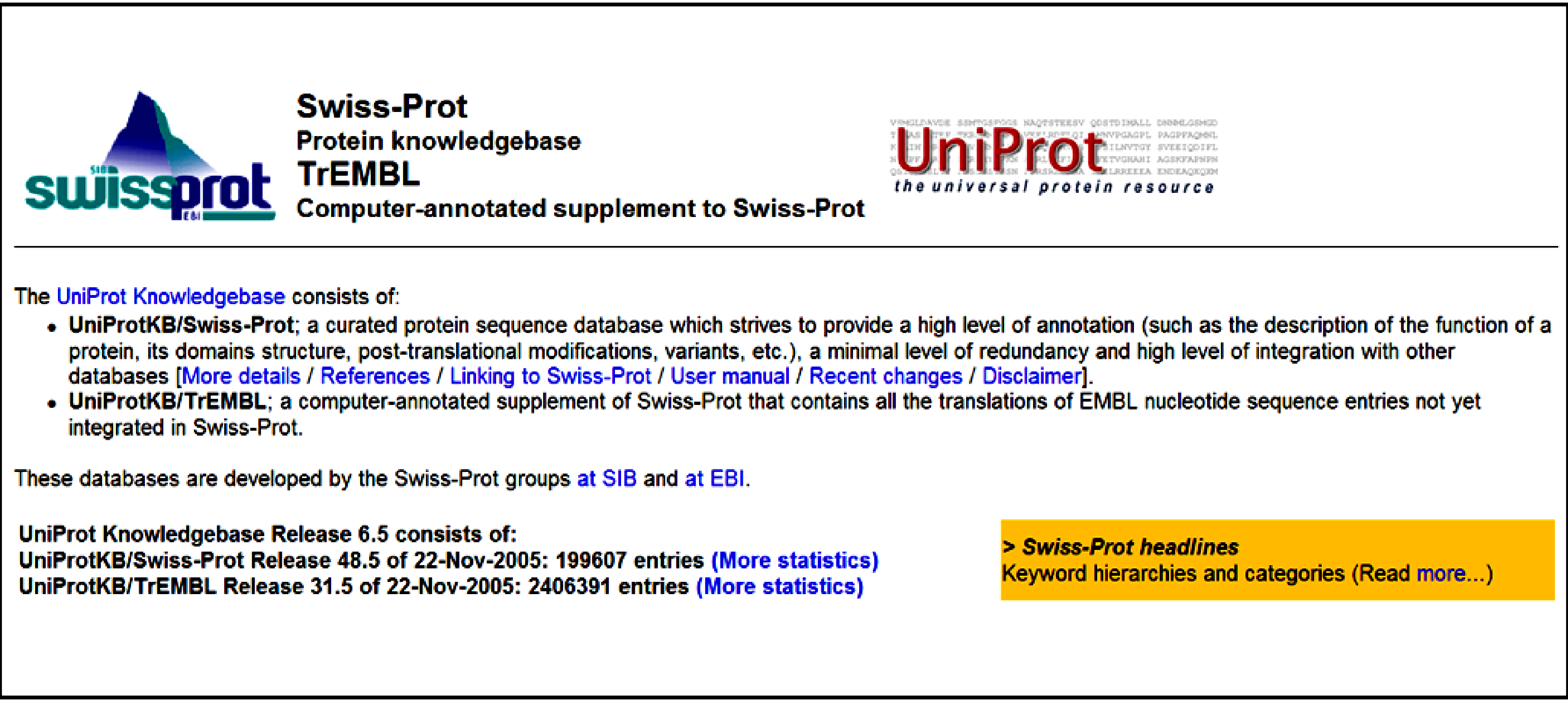
Nucleotide Sequence Databases

- **3 main databases**
  - **EMBL:** www.ebi.ac.uk/embl
  - **GenBank:** www.ncbi.nlm.nih.gov/GenBank
  - **DDBJ:** www.ddbj.nig.ac.jp

The 3 databases are synchronized on a daily basis, and the accession numbers are consistent.

There are no legal restriction in the usage of these databases. However, there are some patented sequences in the database

Protein Sequence Databases



Swiss-Prot
Protein knowledgebase
TrEMBL
Computer-annotated supplement to Swiss-Prot

The UniProt Knowledgebase consists of:

- **UniProtKB/Swiss-Prot**; a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases [More details / References / Linking to Swiss-Prot / User manual / Recent changes / Disclaimer].
- **UniProtKB/TrEMBL**; a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

These databases are developed by the Swiss-Prot groups at SIB and at EBI.

UniProt Knowledgebase Release 6.5 consists of:
UniProtKB/Swiss-Prot Release 48.5 of 22-Nov-2005: 199607 entries (More statistics)
UniProtKB/TrEMBL Release 31.5 of 22-Nov-2005: 2406391 entries (More statistics)

> Swiss-Prot headlines
Keyword hierarchies and categories (Read more...)

The SWISS-PROT database has **some legal restrictions**: the entries are copyrighted, but freely accessible by academic researchers.
Commercial companies must buy a license fee from SIB.

http://www.expasy.ch/sprot/

Анализ белковых последовательностей: Swiss-Prot

Swiss-Prot – одна из первых баз данных белковых последовательностей, “gold standard” белковой аннотации.

Аннотация выполнена вручную группой профессиональных экспертов на основе экспериментальной информации, описанной в научных статьях.

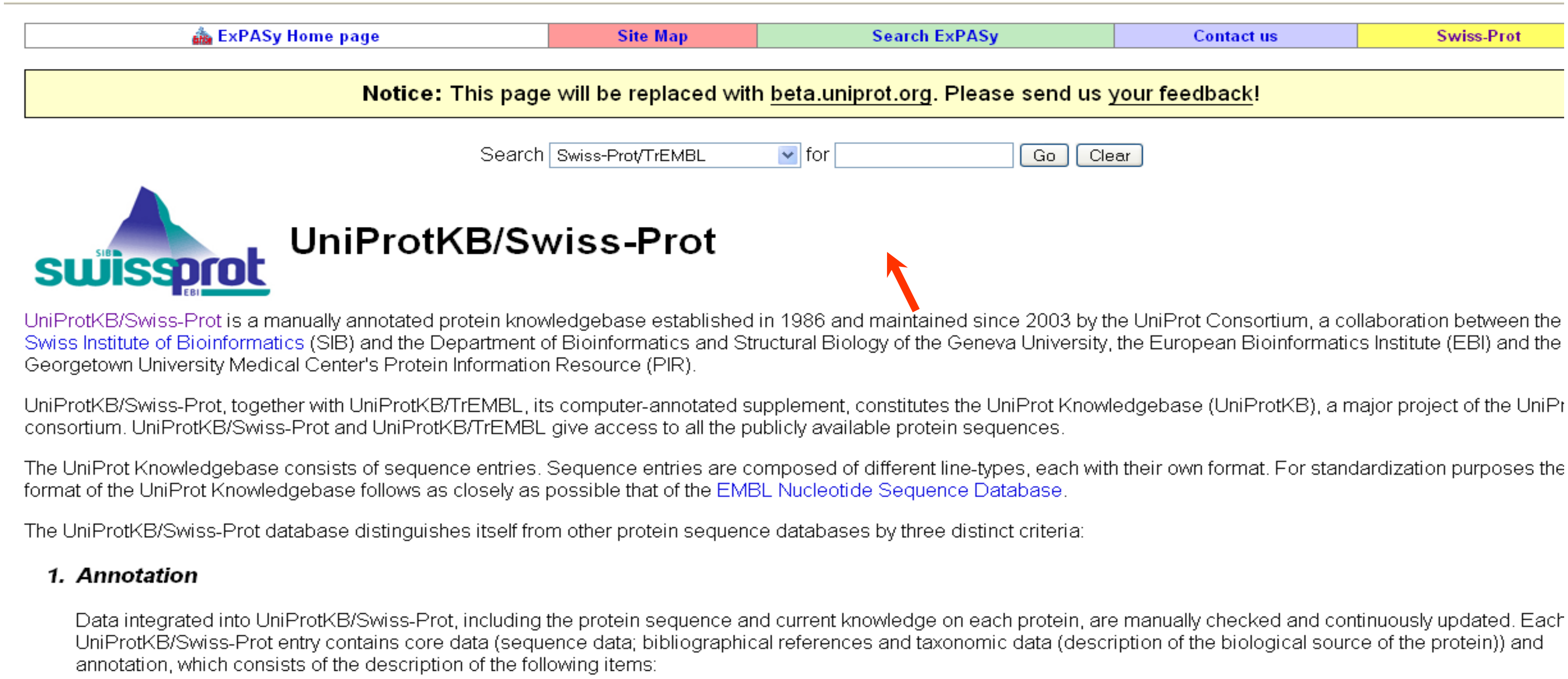
Организована в 1986 году – SIB+EBI+PIR+GU = prof. Amos Bairoch

На сегодняшний день – 556568 последовательностей

UniProt DB

UniProt = Swiss-Prot + TrEMBL (Translated EMBL sequence database)

TrEMBL – 107 427635 sequences



Notice: This page will be replaced with beta.uniprot.org. Please send us your feedback!

UniProtKB/Swiss-Prot

UniProtKB/Swiss-Prot is a manually annotated protein knowledgebase established in 1986 and maintained since 2003 by the UniProt Consortium, a collaboration between the Swiss Institute of Bioinformatics (SIB) and the Department of Bioinformatics and Structural Biology of the Geneva University, the European Bioinformatics Institute (EBI) and the Georgetown University Medical Center's Protein Information Resource (PIR).

UniProtKB/Swiss-Prot, together with UniProtKB/TrEMBL, its computer-annotated supplement, constitutes the UniProt Knowledgebase (UniProtKB), a major project of the UniPr consortium. UniProtKB/Swiss-Prot and UniProtKB/TrEMBL give access to all the publicly available protein sequences.

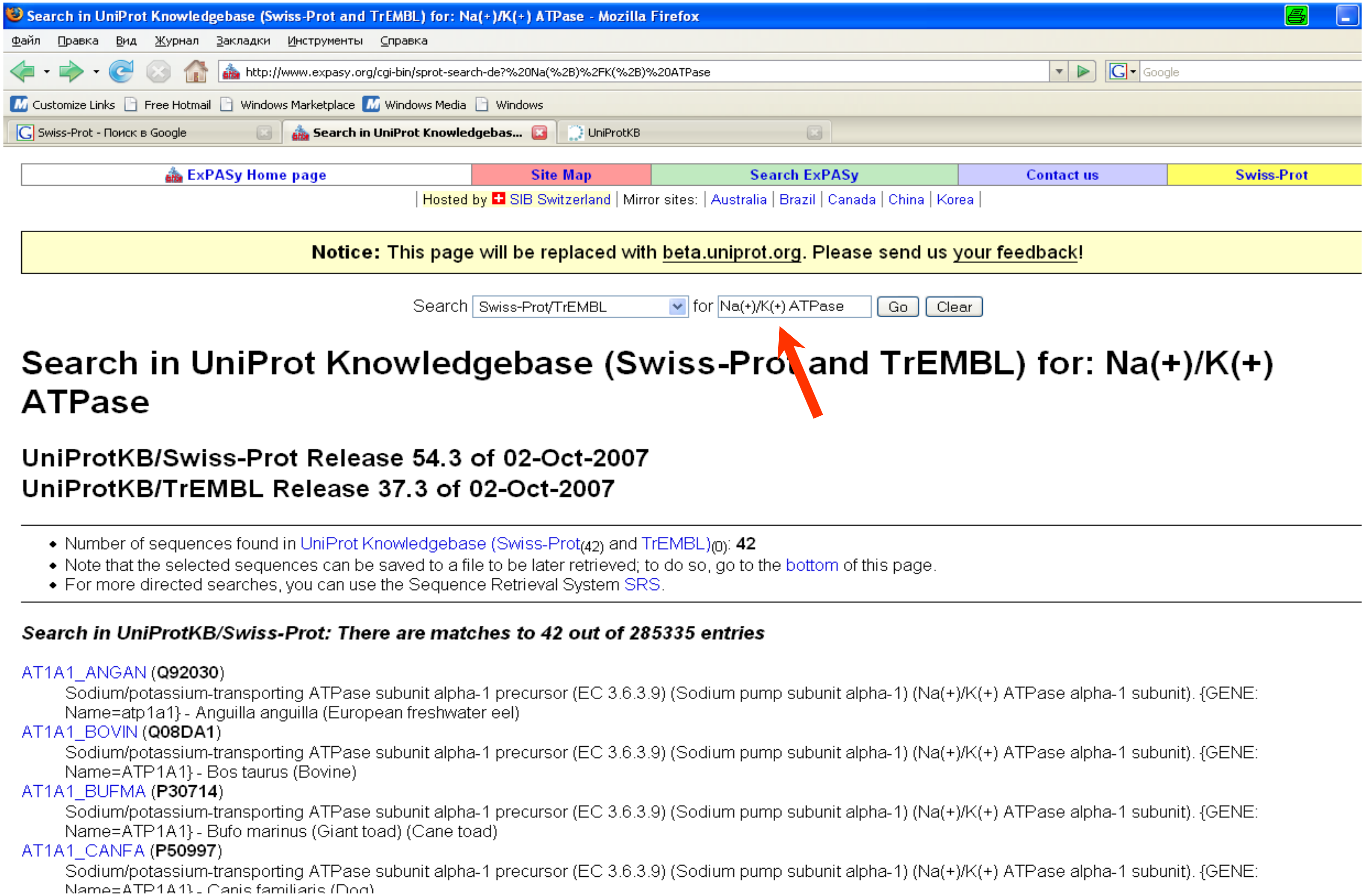
The UniProt Knowledgebase consists of sequence entries. Sequence entries are composed of different line-types, each with their own format. For standardization purposes the format of the UniProt Knowledgebase follows as closely as possible that of the EMBL Nucleotide Sequence Database.

The UniProtKB/Swiss-Prot database distinguishes itself from other protein sequence databases by three distinct criteria:

1. ***Annotation***

   Data integrated into UniProtKB/Swiss-Prot, including the protein sequence and current knowledge on each protein, are manually checked and continuously updated. Each UniProtKB/Swiss-Prot entry contains core data (sequence data; bibliographical references and taxonomic data (description of the biological source of the protein)) and annotation, which consists of the description of the following items:

Поиск белка в Swiss-Prot
(по названию)


Search in UniProt Knowledgebase (Swiss-Prot and TrEMBL) for: Na(+)/K(+) ATPase - Mozilla Firefox
Файл Правка Вид Журнал Закладки Инструменты Справка
http://www.expasy.org/cgi-bin/sprot-search-de?%20Na(%2B)%2FK(%2B)%20ATPase
Google
Customize Links
Free Hotmail
Windows Marketplace
Windows Media
Windows
Swiss-Prot - Поиск в Google
Search in UniProt Knowledgebas...
UniProtKB
ExPASy Home page
Site Map
Search ExPASy
Contact us
Swiss-Prot
Hosted by SIB Switzerland | Mirror sites: | Australia | Brazil | Canada | China | Korea |
Search
Swiss-Prot/TrEMBL
for
Na(+)/K(+) ATPase
Go
Clear

Notice: This page will be replaced with beta.uniprot.org. Please send us your feedback!

Search in UniProt Knowledgebase (Swiss-Prot and TrEMBL) for: Na(+)/K(+) ATPase

UniProtKB/Swiss-Prot Release 54.3 of 02-Oct-2007
UniProtKB/TrEMBL Release 37.3 of 02-Oct-2007

- Number of sequences found in UniProt Knowledgebase (Swiss-Prot(42) and TrEMBL)(0): **42**
- Note that the selected sequences can be saved to a file to be later retrieved; to do so, go to the bottom of this page.
- For more directed searches, you can use the Sequence Retrieval System SRS.

Search in UniProtKB/Swiss-Prot: There are matches to 42 out of 285335 entries

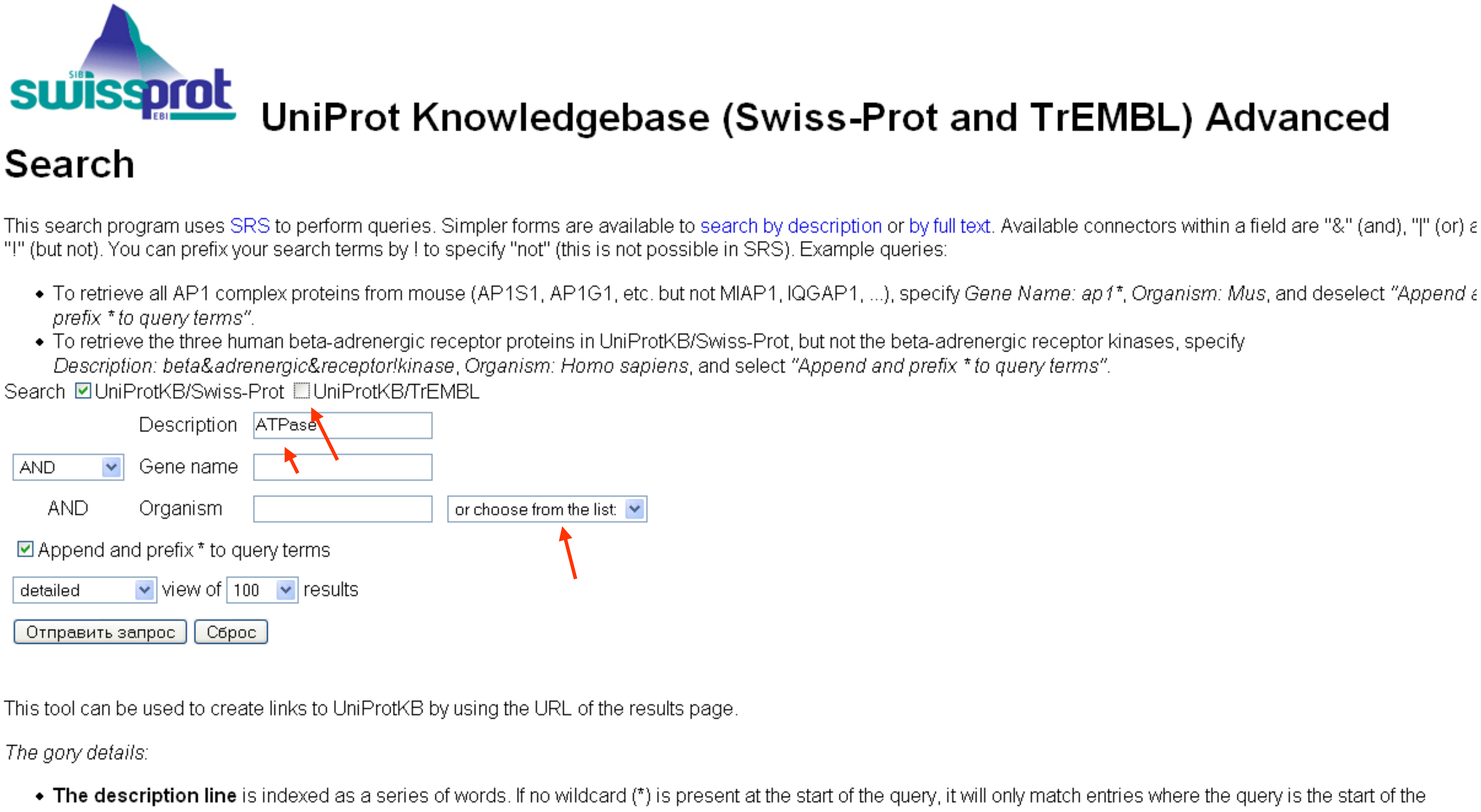
AT1A1_ANGAN (**Q92030**)
Sodium/potassium-transporting ATPase subunit alpha-1 precursor (EC 3.6.3.9) (Sodium pump subunit alpha-1) (Na(+)/K(+) ATPase alpha-1 subunit). {GENE: Name=atp1a1} - Anguilla anguilla (European freshwater eel)

AT1A1_BOVIN (**Q08DA1**)
Sodium/potassium-transporting ATPase subunit alpha-1 precursor (EC 3.6.3.9) (Sodium pump subunit alpha-1) (Na(+)/K(+) ATPase alpha-1 subunit). {GENE: Name=ATP1A1} - Bos taurus (Bovine)

AT1A1_BUFMA (**P30714**)
Sodium/potassium-transporting ATPase subunit alpha-1 precursor (EC 3.6.3.9) (Sodium pump subunit alpha-1) (Na(+)/K(+) ATPase alpha-1 subunit). {GENE: Name=ATP1A1} - Bufo marinus (Giant toad) (Cane toad)

AT1A1_CANFA (**P50997**)
Sodium/potassium-transporting ATPase subunit alpha-1 precursor (EC 3.6.3.9) (Sodium pump subunit alpha-1) (Na(+)/K(+) ATPase alpha-1 subunit). {GENE: Name=ATP1A1} - Canis familiaris (Dog)

Advances search



UniProt Knowledgebase (Swiss-Prot and TrEMBL) Advanced Search

This search program uses SRS to perform queries. Simpler forms are available to search by description or by full text. Available connectors within a field are "&" (and), "|" (or) a "!" (but not). You can prefix your search terms by ! to specify "not" (this is not possible in SRS). Example queries:

- To retrieve all AP1 complex proteins from mouse (AP1S1, AP1G1, etc. but not MIAP1, IQGAP1, ...), specify *Gene Name: ap1**, *Organism: Mus*, and deselect *"Append a prefix * to query terms"*.
- To retrieve the three human beta-adrenergic receptor proteins in UniProtKB/Swiss-Prot, but not the beta-adrenergic receptor kinases, specify *Description: beta&adrenergic&receptor!kinase*, *Organism: Homo sapiens*, and select *"Append and prefix * to query terms"*.

Search ☑ UniProtKB/Swiss-Prot □ UniProtKB/TrEMBL

	Description	ATPase	
AND	Gene name		
AND	Organism		or choose from the list:

☑ Append and prefix * to query terms

detailed view of 100 results

This tool can be used to create links to UniProtKB by using the URL of the results page.

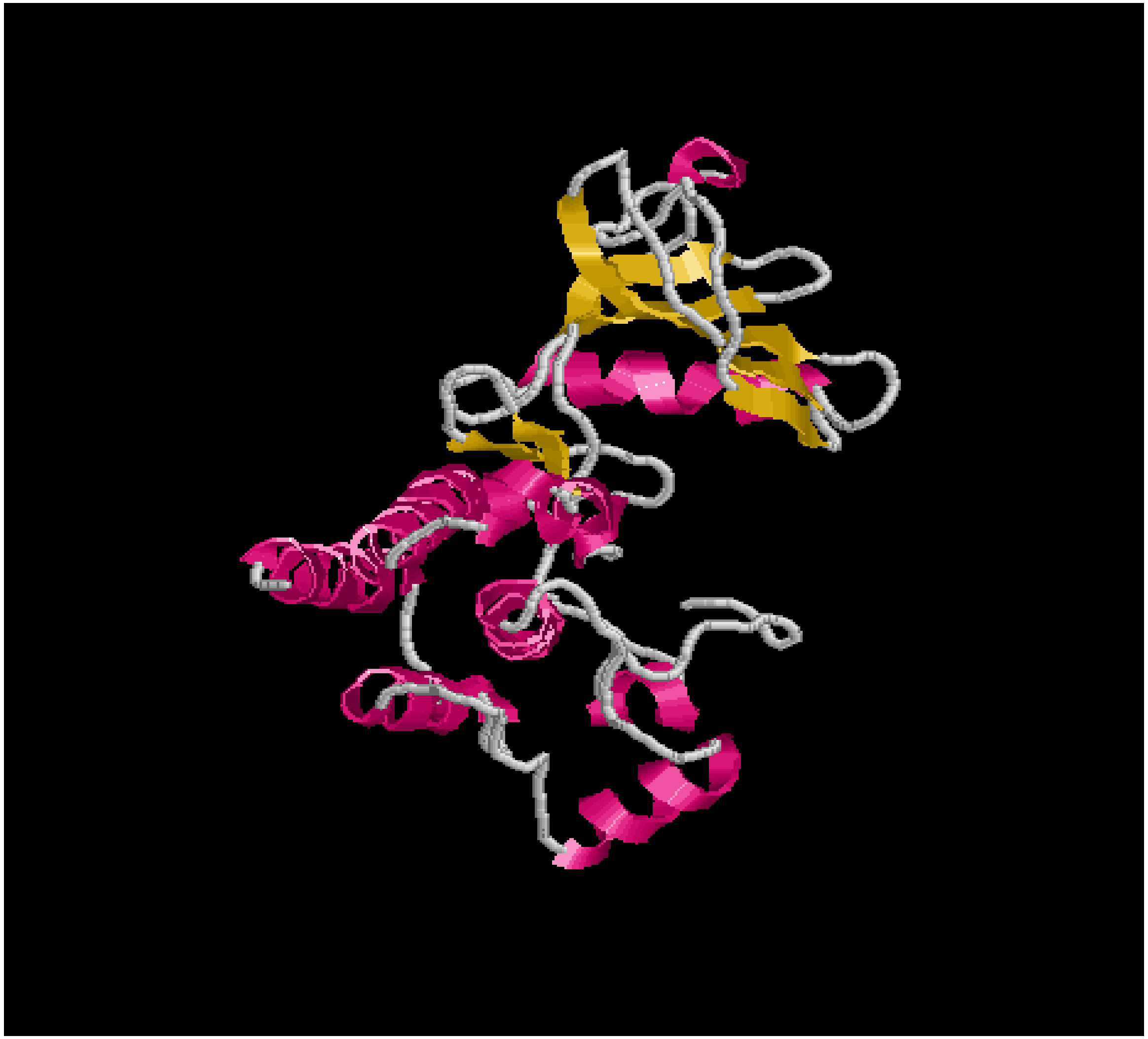
The gory details:

- **The description line** is indexed as a series of words. If no wildcard (*) is present at the start of the query, it will only match entries where the query is the start of the

Biomolecule Structure Database

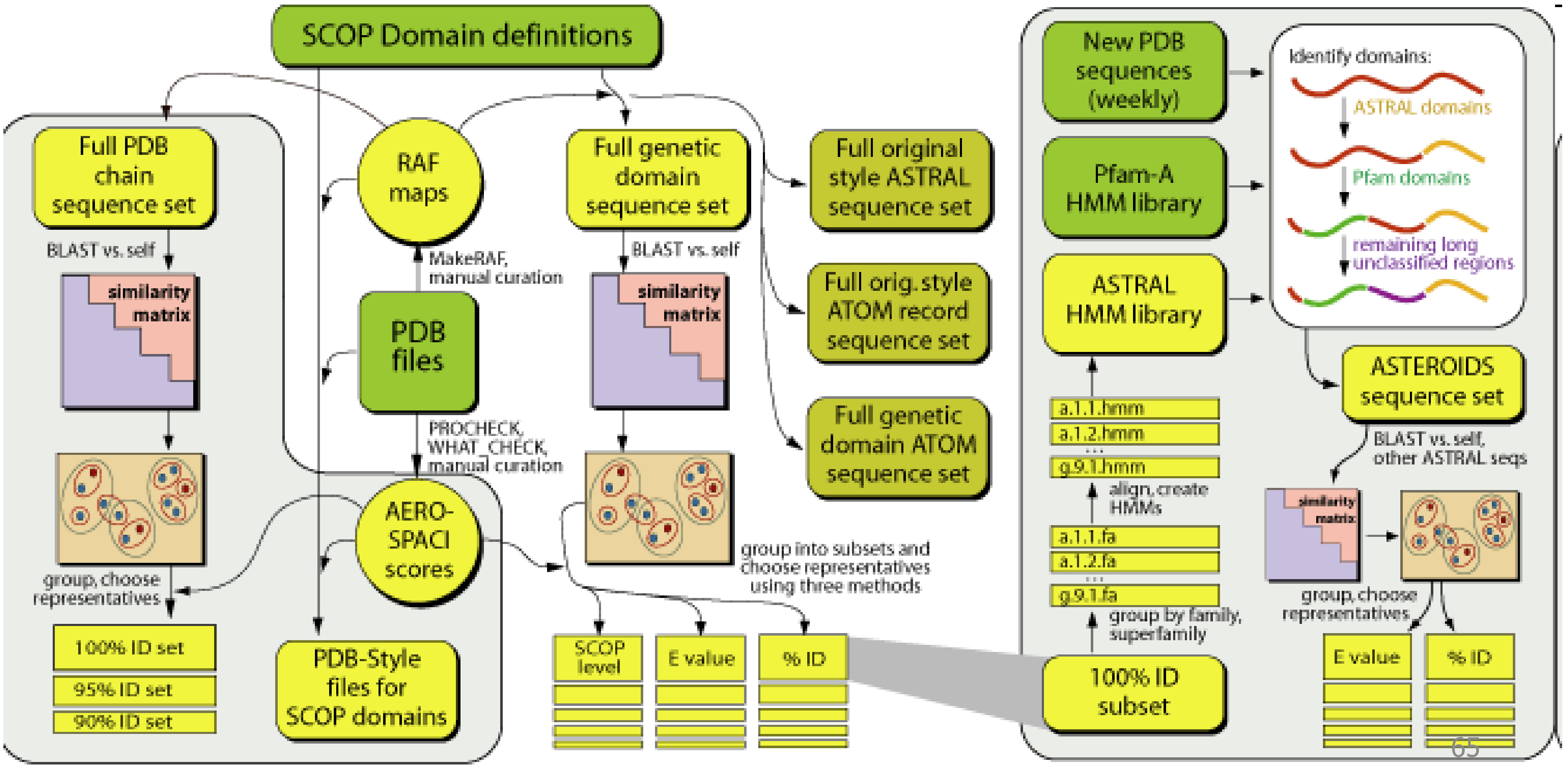
- PDB: http://www.rcsb.org
- SCOP: http://scop.berkeley.edu
- CATH: http://biochem.ucl.ac.uk/bsm/CATH
- ASTRAL: http://astral.berkeley.edu
- Interfaces to PDB:
  - **PDB at a glance** http://cmm.info.nih.gov/modeling/pdb_at_a_glance.html
  - **Molecules to go** http://molbio.info.nih.gov/cgi-bin/pdb/
  - **EBI interface**: http://www.ebi.ac.uk/msd/
  - **PDBSum:** http://www.ebi.ac.uk/thornton-srv/databases/pdbsum

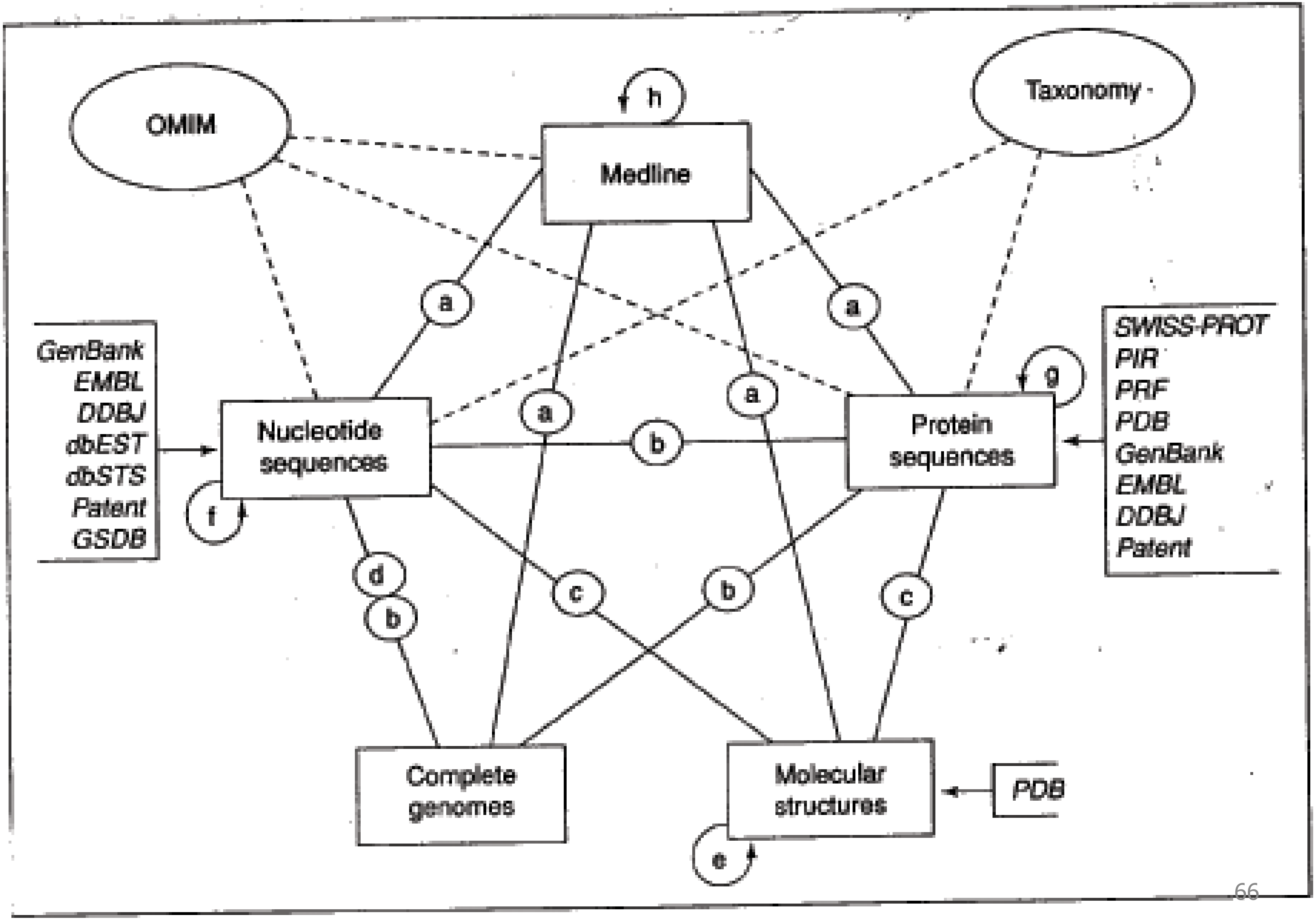
Serine-threonine and tyrosine protein kinases



Data flow in ASTRAL

The **ASTRAL** compendium provides databases and tools useful for analyzing protein structures and their sequences


SCOP Domain definitions
Full PDB chain sequence set
BLAST vs. self
similarity matrix
group, choose representatives
100% ID set
95% ID set
90% ID set
RAF maps
MakeRAF, manual curation
PDB files
PROCHECK, WHAT_CHECK, manual curation
AERO-SPACI scores
PDB-Style files for SCOP domains
Full genetic domain sequence set
BLAST vs. self
similarity matrix
group into subsets and choose representatives using three methods
SCOP level
E value
% ID
Full original style ASTRAL sequence set
Full orig. style ATOM record sequence set
Full genetic domain ATOM sequence set
New PDB sequences (weekly)
Pfam-A HMM library
ASTRAL HMM library
Identify domains:
ASTRAL domains
Pfam domains
remaining long unclassified regions
ASTEROIDS sequence set
BLAST vs. self, other ASTRAL seqs
similarity matrix
group, choose representatives
E value
% ID
a.1.1.hmm
a.1.2.hmm
...
g.9.1.hmm
align, create HMMs
a.1.1.fa
a.1.2.fa
...
g.9.1.fa
group by family, superfamily
100% ID subset


OMIM
h
Medline
Taxonomy
a
a
GenBank
EMBL
DDBJ
dbEST
dbSTS
Patent
GSDB
a
a
SWISS-PROT
PIR
PRF
PDB
GenBank
EMBL
DDBJ
Patent
g
Nucleotide sequences
b
Protein sequences
f
d
b
c
b
c
Complete genomes
Molecular structures
PDB
e

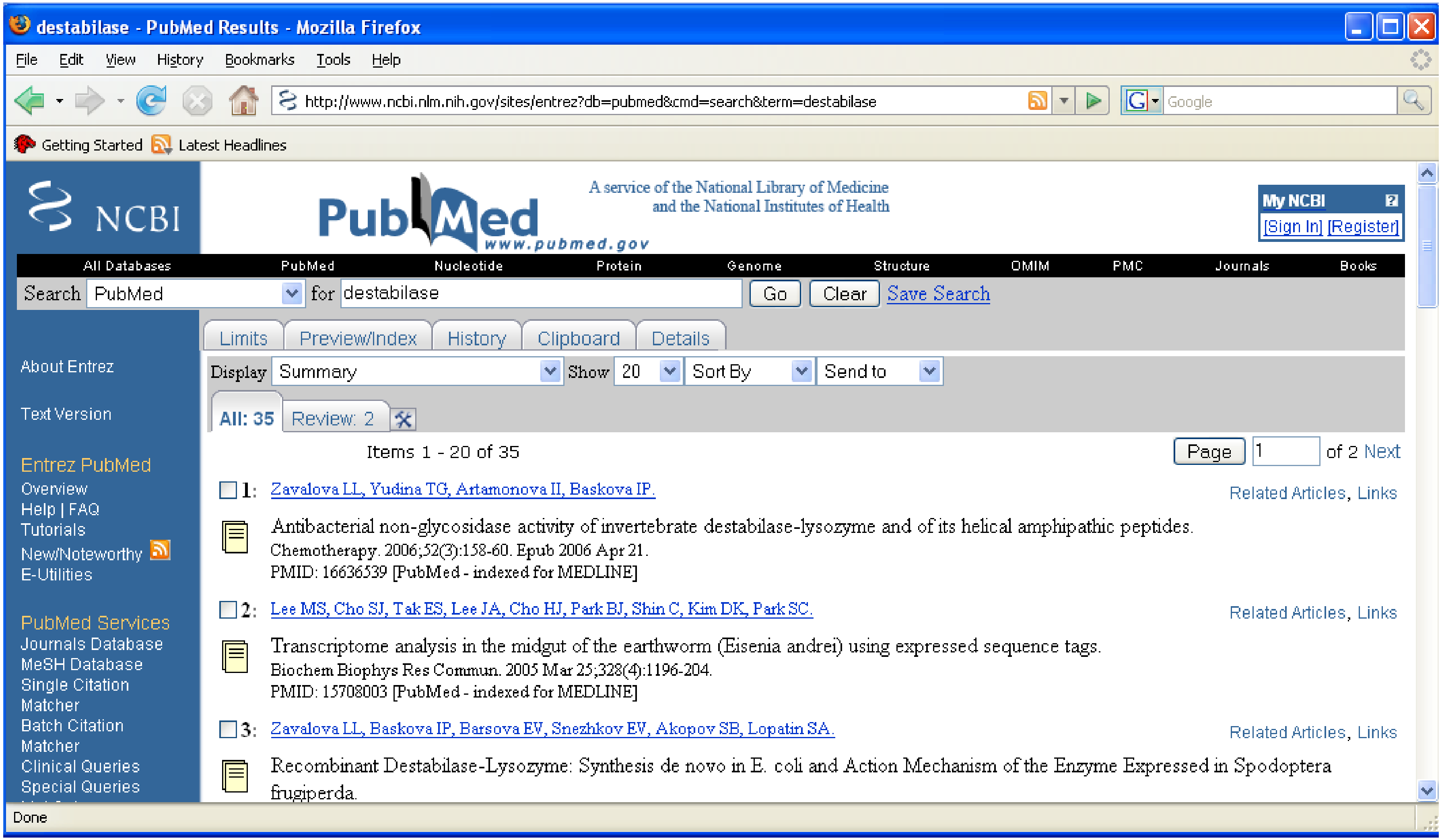
Поиск литературы: PubMed

PubMed is a service of the **U.S. National Library of Medicine** that includes over 18 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s. PubMed includes links to full text articles and other related resources.

URLs: **www.pubmed.gov**

www.ncbi.nlm.nih.gov

Поиск по названию белка



Search PubMed for destabilase

Display Summary; Show 20

All: 35 | Review: 2

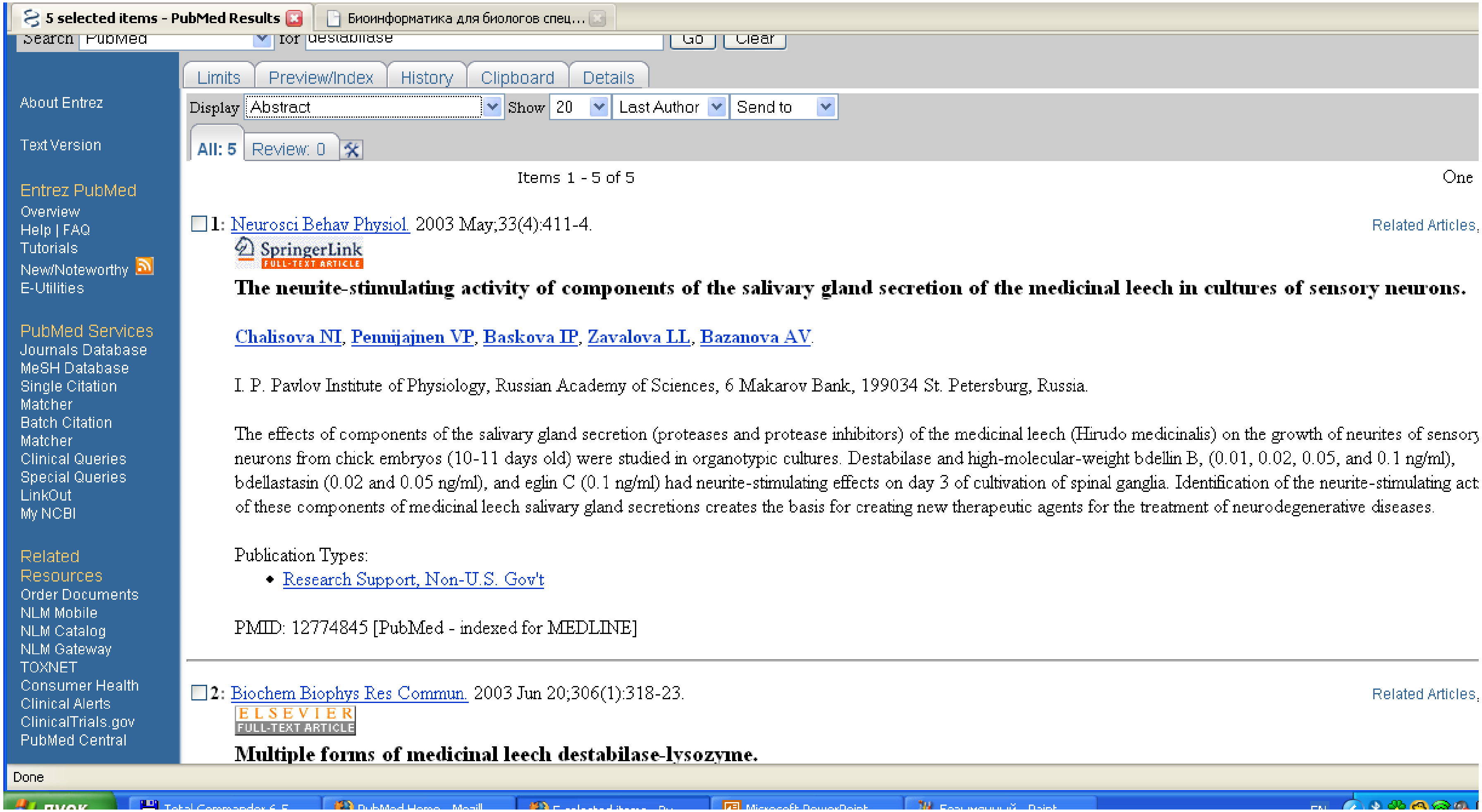
Items 1 - 20 of 35 — Page 1 of 2

□ 1: Zavalova LL, Yudina TG, Artamonova II, Baskova IP. — Related Articles, Links
Antibacterial non-glycosidase activity of invertebrate destabilase-lysozyme and of its helical amphipathic peptides.
Chemotherapy. 2006;52(3):158-60. Epub 2006 Apr 21.
PMID: 16636539 [PubMed - indexed for MEDLINE]

□ 2: Lee MS, Cho SJ, Tak ES, Lee JA, Cho HJ, Park BJ, Shin C, Kim DK, Park SC. — Related Articles, Links
Transcriptome analysis in the midgut of the earthworm (Eisenia andrei) using expressed sequence tags.
Biochem Biophys Res Commun. 2005 Mar 25;328(4):1196-204.
PMID: 15708003 [PubMed - indexed for MEDLINE]

□ 3: Zavalova LL, Baskova IP, Barsova EV, Snezhkov EV, Akopov SB, Lopatin SA. — Related Articles, Links
Recombinant Destabilase-Lysozyme: Synthesis de novo in E. coli and Action Mechanism of the Enzyme Expressed in Spodoptera frugiperda.

Как это выглядит



Items 1 - 5 of 5

One

□ 1: Neurosci Behav Physiol. 2003 May;33(4):411-4.

The neurite-stimulating activity of components of the salivary gland secretion of the medicinal leech in cultures of sensory neurons.

Chalisova NI, Pennijajnen VP, Baskova IP, Zavalova LL, Bazanova AV.

I. P. Pavlov Institute of Physiology, Russian Academy of Sciences, 6 Makarov Bank, 199034 St. Petersburg, Russia.

The effects of components of the salivary gland secretion (proteases and protease inhibitors) of the medicinal leech (Hirudo medicinalis) on the growth of neurites of sensory neurons from chick embryos (10-11 days old) were studied in organotypic cultures. Destabilase and high-molecular-weight bdellin B, (0.01, 0.02, 0.05, and 0.1 ng/ml), bdellastasin (0.02 and 0.05 ng/ml), and eglin C (0.1 ng/ml) had neurite-stimulating effects on day 3 of cultivation of spinal ganglia. Identification of the neurite-stimulating act of these components of medicinal leech salivary gland secretions creates the basis for creating new therapeutic agents for the treatment of neurodegenerative diseases.

Publication Types:

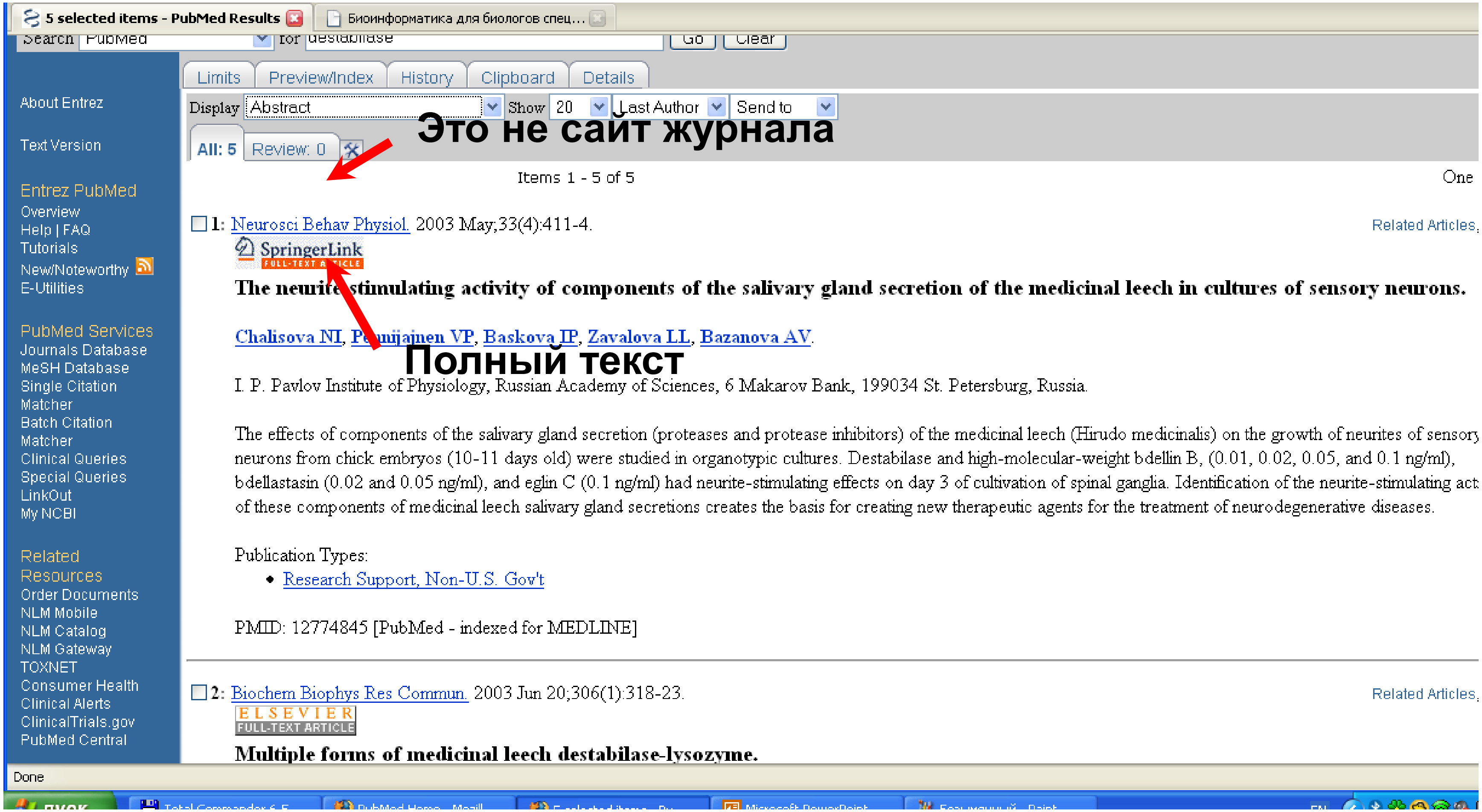
- Research Support, Non-U.S. Gov't

PMID: 12774845 [PubMed - indexed for MEDLINE]

□ 2: Biochem Biophys Res Commun. 2003 Jun 20;306(1):318-23.

Multiple forms of medicinal leech destabilase-lysozyme.

Как получить статью



Это не сайт журнала

Полный текст

Items 1 - 5 of 5

One

□ 1: Neurosci Behav Physiol. 2003 May;33(4):411-4.

The neurite-stimulating activity of components of the salivary gland secretion of the medicinal leech in cultures of sensory neurons.

Chalisova NI, Pennijajnen VP, Baskova IP, Zavalova LL, Bazanova AV.

I. P. Pavlov Institute of Physiology, Russian Academy of Sciences, 6 Makarov Bank, 199034 St. Petersburg, Russia.

The effects of components of the salivary gland secretion (proteases and protease inhibitors) of the medicinal leech (Hirudo medicinalis) on the growth of neurites of sensory neurons from chick embryos (10-11 days old) were studied in organotypic cultures. Destabilase and high-molecular-weight bdellin B, (0.01, 0.02, 0.05, and 0.1 ng/ml), bdellastasin (0.02 and 0.05 ng/ml), and eglin C (0.1 ng/ml) had neurite-stimulating effects on day 3 of cultivation of spinal ganglia. Identification of the neurite-stimulating act of these components of medicinal leech salivary gland secretions creates the basis for creating new therapeutic agents for the treatment of neurodegenerative diseases.

Publication Types:

- Research Support, Non-U.S. Gov't

PMID: 12774845 [PubMed - indexed for MEDLINE]

□ 2: Biochem Biophys Res Commun. 2003 Jun 20;306(1):318-23.

Multiple forms of medicinal leech destabilase-lysozyme.

Другие виды поиска

По любым ключевым словам или их сочетаниям (AND – необязательно)

По автору (лучше с инициалами!)

По названию статьи

По журналу

По аффилиации авторов

Только в аннотациях

По PMID

По дате – год, либо год/месяц

По словосочетанию – взять в кавычки

Ген-ориентированные базы данных и геномные браузеры

Что такое ген-ориентированные базы данных?

Самые простые примеры таких БД

Примеры геном-ориентированных баз данных и геномные браузеры

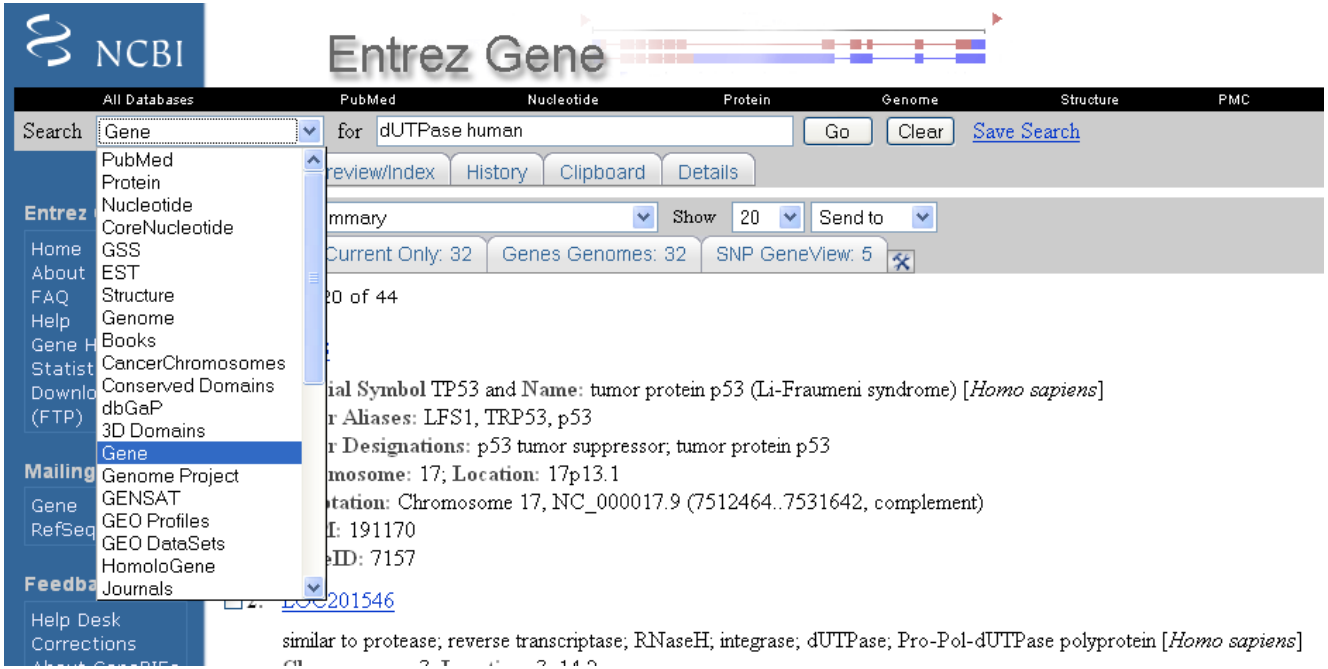
Human Genome Browser

Что такое ген-ориентированные базы данных?

- Единица исследования – ген (а не экспериментальная последовательность)
- Призваны снабжать информацией по конкретному гену, а не “последовательностям, относящимся ко данному конкретному гену” – интегрируют все такие части в единое целое за Вас

Первый пример – Gene Entrez (бывший LocusLink) в NCBI

- Единица – генетический локус – конкретное место на хромосоме, кодирующее данный белок и/или соответствующее данному гену



Search Gene for dUTPase human

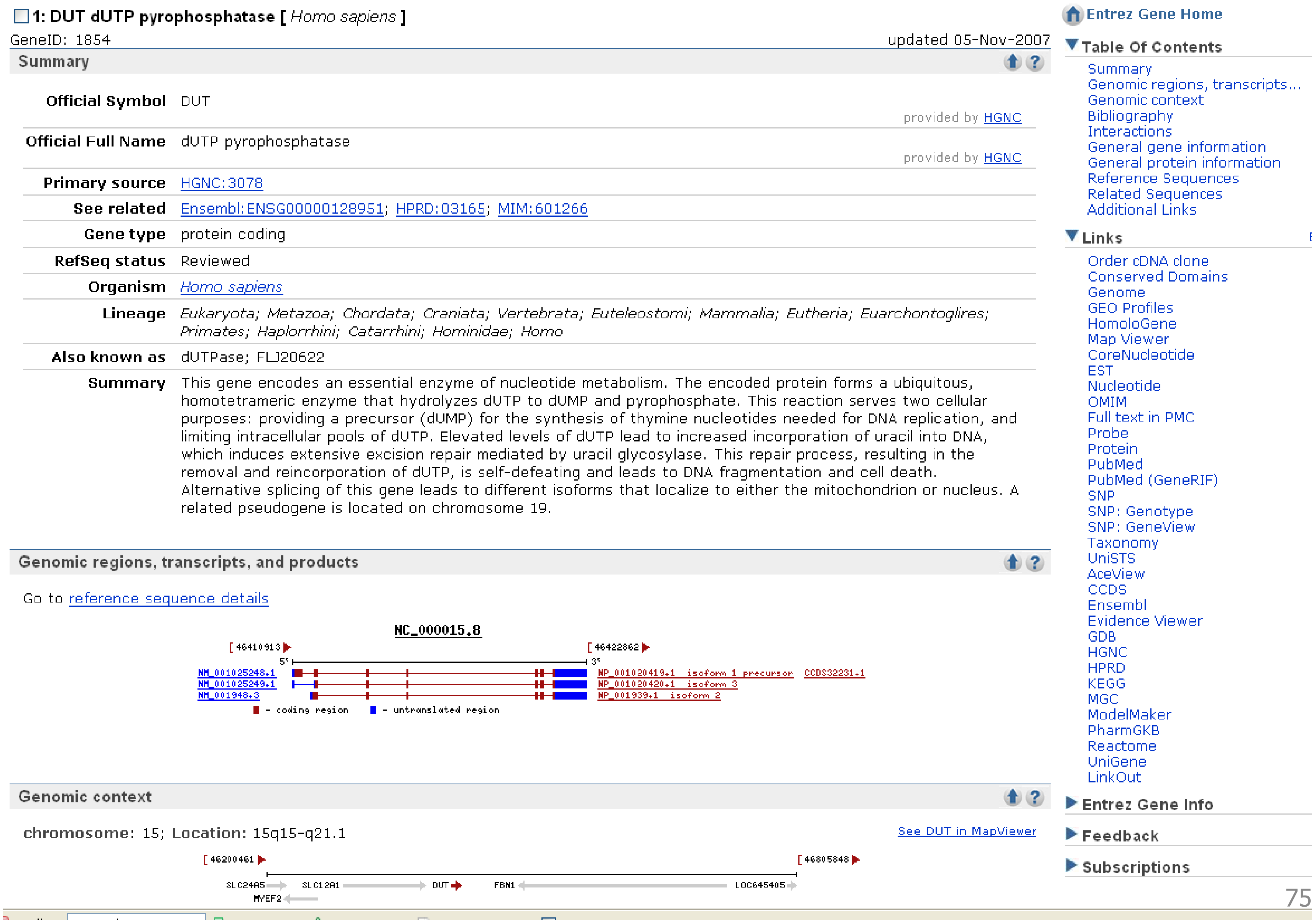
20 of 44

ial Symbol TP53 and Name: tumor protein p53 (Li-Fraumeni syndrome) [*Homo sapiens*]
r Aliases: LFS1, TRP53, p53
r Designations: p53 tumor suppressor; tumor protein p53
mosome: 17; Location: 17p13.1
tation: Chromosome 17, NC_000017.9 (7512464..7531642, complement)
I: 191170
ID: 7157

□ 2: LOC201546

similar to protease; reverse transcriptase; RNaseH; integrase; dUTPase; Pro-Pol-dUTPase polyprotein [*Homo sapiens*]

DUT ген человека



□ **1: DUT dUTP pyrophosphatase [** *Homo sapiens* **]**

GeneID: 1854 updated 05-Nov-2007

Summary

Official Symbol	DUT provided by HGNC
Official Full Name	dUTP pyrophosphatase provided by HGNC
Primary source	HGNC:3078
See related	Ensembl:ENSG00000128951; HPRD:03165; MIM:601266
Gene type	protein coding
RefSeq status	Reviewed
Organism	*Homo sapiens*
Lineage	*Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo*
Also known as	dUTPase; FLJ20622
Summary	This gene encodes an essential enzyme of nucleotide metabolism. The encoded protein forms a ubiquitous, homotetrameric enzyme that hydrolyzes dUTP to dUMP and pyrophosphate. This reaction serves two cellular purposes: providing a precursor (dUMP) for the synthesis of thymine nucleotides needed for DNA replication, and limiting intracellular pools of dUTP. Elevated levels of dUTP lead to increased incorporation of uracil into DNA, which induces extensive excision repair mediated by uracil glycosylase. This repair process, resulting in the removal and reincorporation of dUTP, is self-defeating and leads to DNA fragmentation and cell death. Alternative splicing of this gene leads to different isoforms that localize to either the mitochondrion or nucleus. A related pseudogene is located on chromosome 19.

Genomic regions, transcripts, and products

Go to reference sequence details

Genomic context

chromosome: 15; **Location:** 15q15-q21.1

See DUT in MapViewer

Entrez Gene Home

▼ **Table Of Contents**

▼ **Links**

- Order cDNA clone
- Conserved Domains
- Genome
- GEO Profiles
- HomoloGene
- Map Viewer
- CoreNucleotide
- EST
- Nucleotide
- OMIM
- Full text in PMC
- Probe
- Protein
- PubMed
- PubMed (GeneRIF)
- SNP
- SNP: Genotype
- SNP: GeneView
- Taxonomy
- UniSTS
- AceView
- CCDS
- Ensembl
- Evidence Viewer
- GDB
- HGNC
- HPRD
- KEGG
- MGC
- ModelMaker
- PharmGKB
- Reactome
- UniGene
- LinkOut

▶ **Entrez Gene Info**

▶ **Feedback**

▶ **Subscriptions**

Продолжение записи:

Bibliography

- **Related Articles in PubMed**
- **GeneRIFs: Gene References Into Function**

Interactions

General gene information

- **Markers**
- **Genotypes**
- **Pathways**
- **Homology**

GeneOntology

General protein information (Names, ECs, ACs)

NCBI Reference Sequences (RefSeq)

- **mRNAs and proteins**
- **Reference assembly + Alternate assembly: Genomic**

Related Sequences (links between ACs of different types)

Additional Links (OMIM, PharmGKB, HRDP, UniGene)

Геномные базы данных

Объект – полный геном

Возможность одновременно изучать все гены одного генома

Сравнение друг с другом целых геномов – сравнительная геномика (comparative genomics)

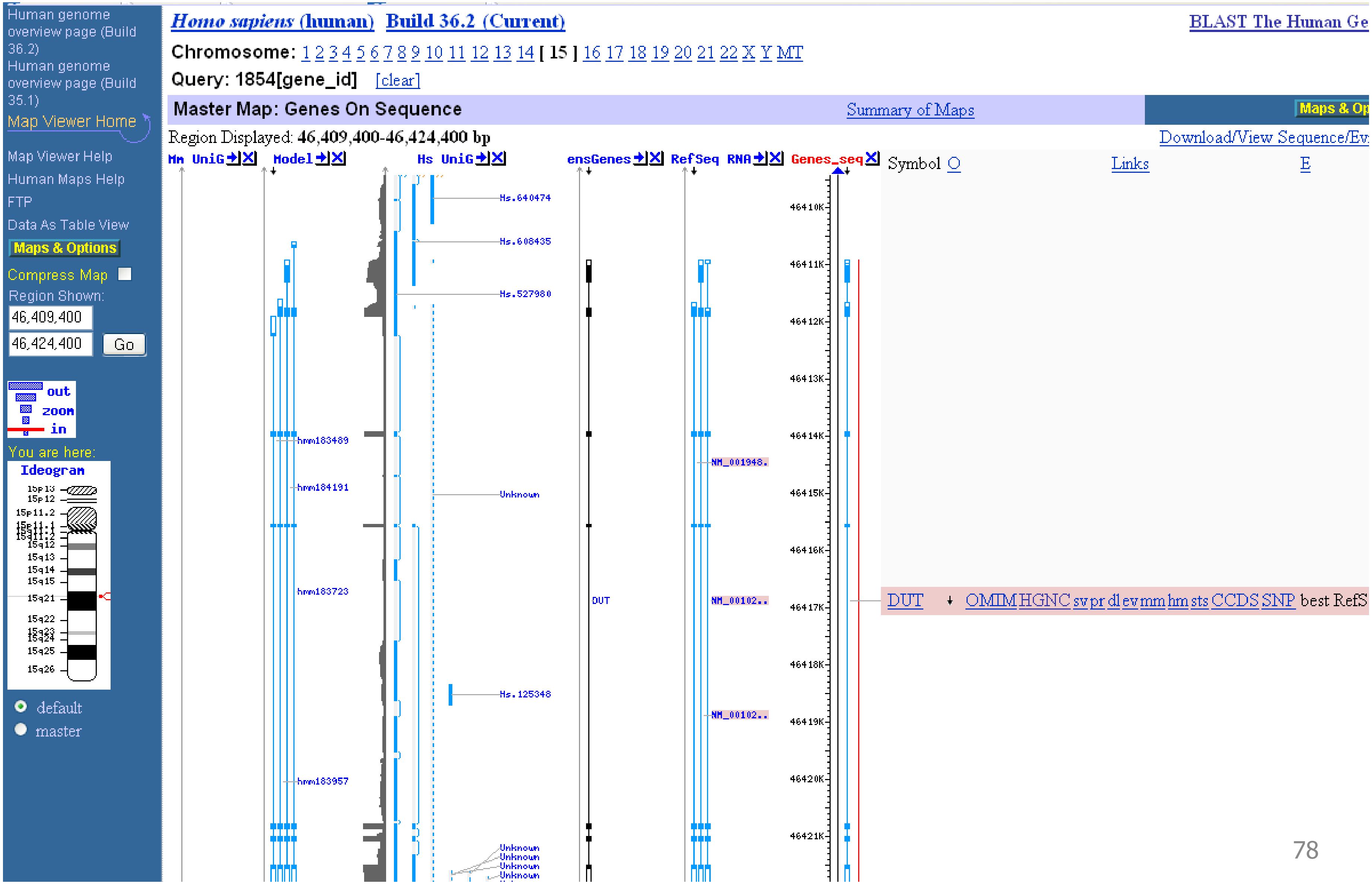
Интеграция всей доступной информации о данном геноме

Основная информация о генах, но в геномном контексте

Геномные браузеры – графическое представление всей интегрированной информации

NCBI -> Genomic Biology (http://www.ncbi.nlm.nih.gov/Genomes/)

MapViewer



***Homo sapiens* (human)** **Build 36.2 (Current)**

Chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 [15] 16 17 18 19 20 21 22 X Y MT

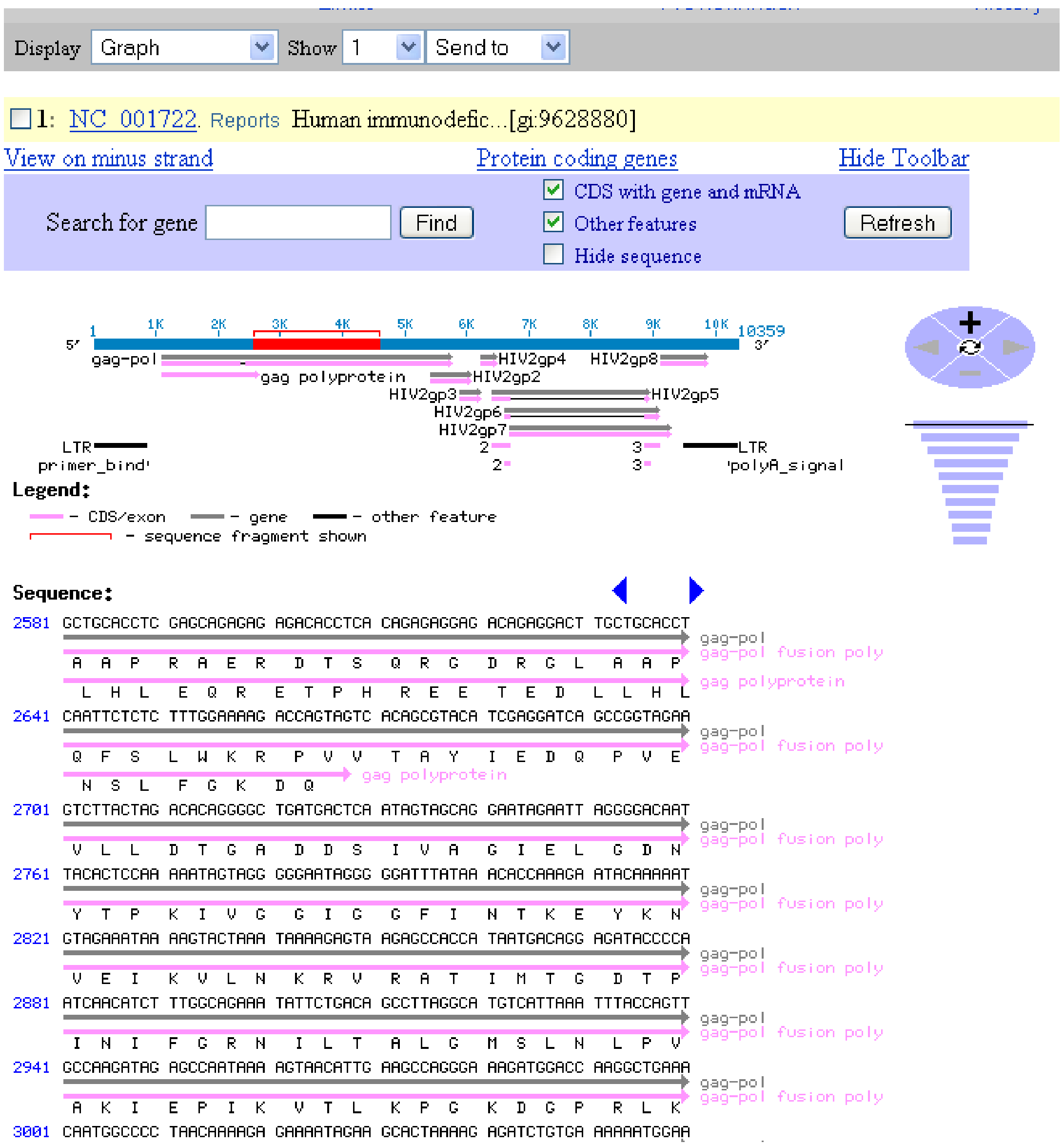
Query: 1854[gene_id] [clear]

Master Map: Genes On Sequence

Region Displayed: 46,409,400-46,424,400 bp

Region Shown: 46,409,400 46,424,400

Sequence Viewer



Display Graph Show 1 Send to

□ 1: NC_001722. Reports Human immunodefic...[gi:9628880]

View on minus strand   Protein coding genes   Hide Toolbar

Search for gene [ ] Find

☑ CDS with gene and mRNA
☑ Other features
□ Hide sequence

Refresh

Legend:

- CDS/exon - gene - other feature
- sequence fragment shown

Sequence:

```
2581 GCTGCACCTC GAGCAGAGAG AGACACCTCA CAGAGAGGAG ACAGAGGACT TGCTGCACCT
                                                                   gag-pol
                                                                   gag-pol fusion poly
      A  A  P   R  A  E  R   D  T  S   Q  R  G   D  R  G  L   A  A  P
                                                                   gag polyprotein
       L  H  L   E  Q  R   E  T  P  H   R  E  E   T  E  D   L  L  H  L
2641 CAATTCTCTC TTTGGAAAAG ACCAGTAGTC ACAGCGTACA TCGAGGATCA GCCGGTAGAA
                                                                   gag-pol
                                                                   gag-pol fusion poly
      Q  F  S   L  W  K  R   P  V  V   T  A  Y   I  E  D  Q   P  V  E
                             gag polyprotein
       N  S  L   F  G  K   D  Q
2701 GTCTTACTAG ACACAGGGGC TGATGACTCA ATAGTAGCAG GAATAGAATT AGGGGACAAT
                                                                   gag-pol
                                                                   gag-pol fusion poly
      V  L  L   D  T  G  A   D  D  S   I  V  A   G  I  E  L   G  D  N
2761 TACACTCCAA AAATAGTAGG GGGAATAGGG GGATTTATAA ACACCAAAGA ATACAAAAAT
                                                                   gag-pol
                                                                   gag-pol fusion poly
      Y  T  P   K  I  V  G   G  I  G   G  F  I   N  T  K  E   Y  K  N
2821 GTAGAAATAA AAGTACTAAA TAAAAGAGTA AGAGCCACCA TAATGACAGG AGATACCCCA
                                                                   gag-pol
                                                                   gag-pol fusion poly
      V  E  I   K  V  L  N   K  R  V   R  A  T   I  M  T  G   D  T  P
2881 ATCAACATCT TTGGCAGAAA TATTCTGACA GCCTTAGGCA TGTCATTAAA TTTACCAGTT
                                                                   gag-pol
                                                                   gag-pol fusion poly
      I  N  I   F  G  R  N   I  L  T   A  L  G   M  S  L  N   L  P  V
2941 GCCAAGATAG AGCCAATAAA AGTAACATTG AAGCCAGGGA AAGATGGACC AAGGCTGAAA
                                                                   gag-pol
                                                                   gag-pol fusion poly
      A  K  I   E  P  I  K   V  T  L   K  P  G   K  D  G  P   R  L  K
3001 CAATGGCCCC TAACAAAAGA GAAAATAGAA GCACTAAAAG AGATCTGTGA AAAAATGGAA
```

Human

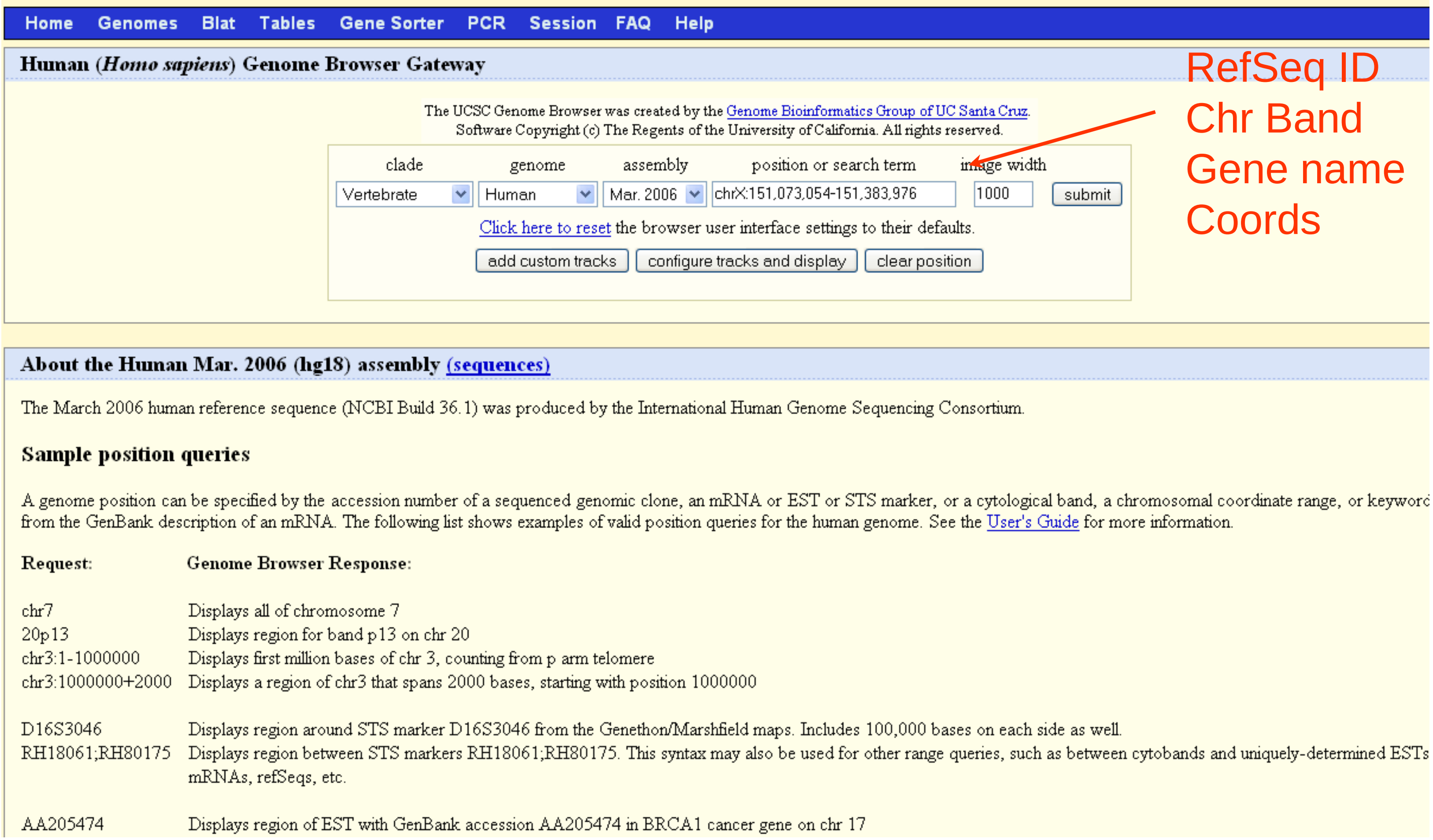
Два основных браузера:

Ensembl (http://www.ensembl.org) – EBI & Sanger Institute, использует свои IDs, 35 эукариотических видов

Human Genome Browser (http://genome.ucsc.edu/) – UCSC, USA

использует GenBank IDs, 41 эукариотический вид

Human Genome Browser



Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade	genome	assembly	position or search term	image width	
Vertebrate	Human	Mar. 2006	chrX:151,073,054-151,383,976	1000	submit

Click here to reset the browser user interface settings to their defaults.

RefSeq ID
Chr Band
Gene name
Coords

About the Human Mar. 2006 (hg18) assembly (sequences)

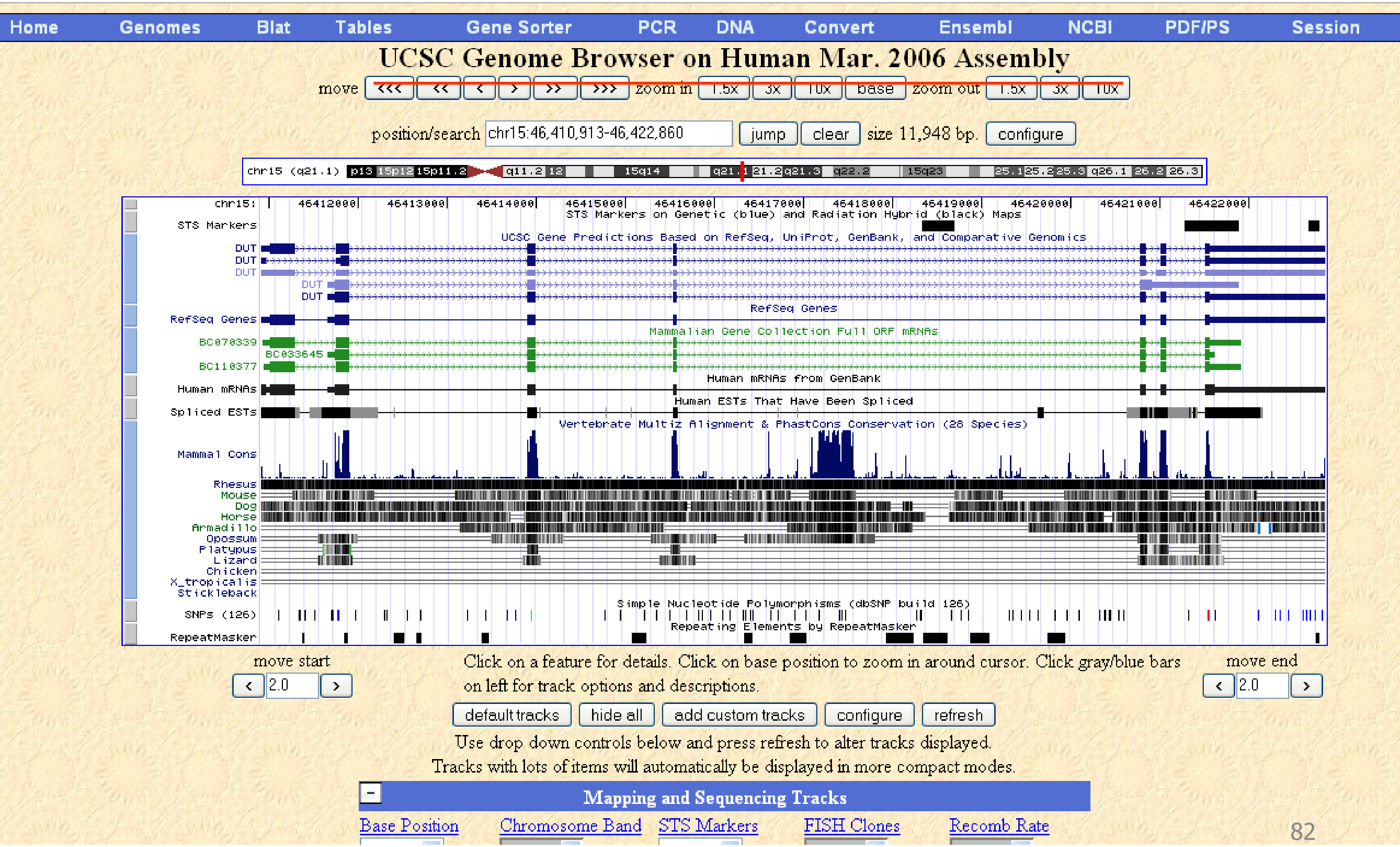
The March 2006 human reference sequence (NCBI Build 36.1) was produced by the International Human Genome Sequencing Consortium.

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, or a cytological band, a chromosomal coordinate range, or keyword from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's Guide for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
D16S3046	Displays region around STS marker D16S3046 from the Genethon/Marshfield maps. Includes 100,000 bases on each side as well.
RH18061;RH80175	Displays region between STS markers RH18061;RH80175. This syntax may also be used for other range queries, such as between cytobands and uniquely-determined ESTs mRNAs, refSeqs, etc.
AA205474	Displays region of EST with GenBank accession AA205474 in BRCA1 cancer gene on chr 17

DUT gene (dUTPAse)



UCSC Genome Browser on Human Mar. 2006 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr15:46,410,913-46,422,860 size 11,948 bp.

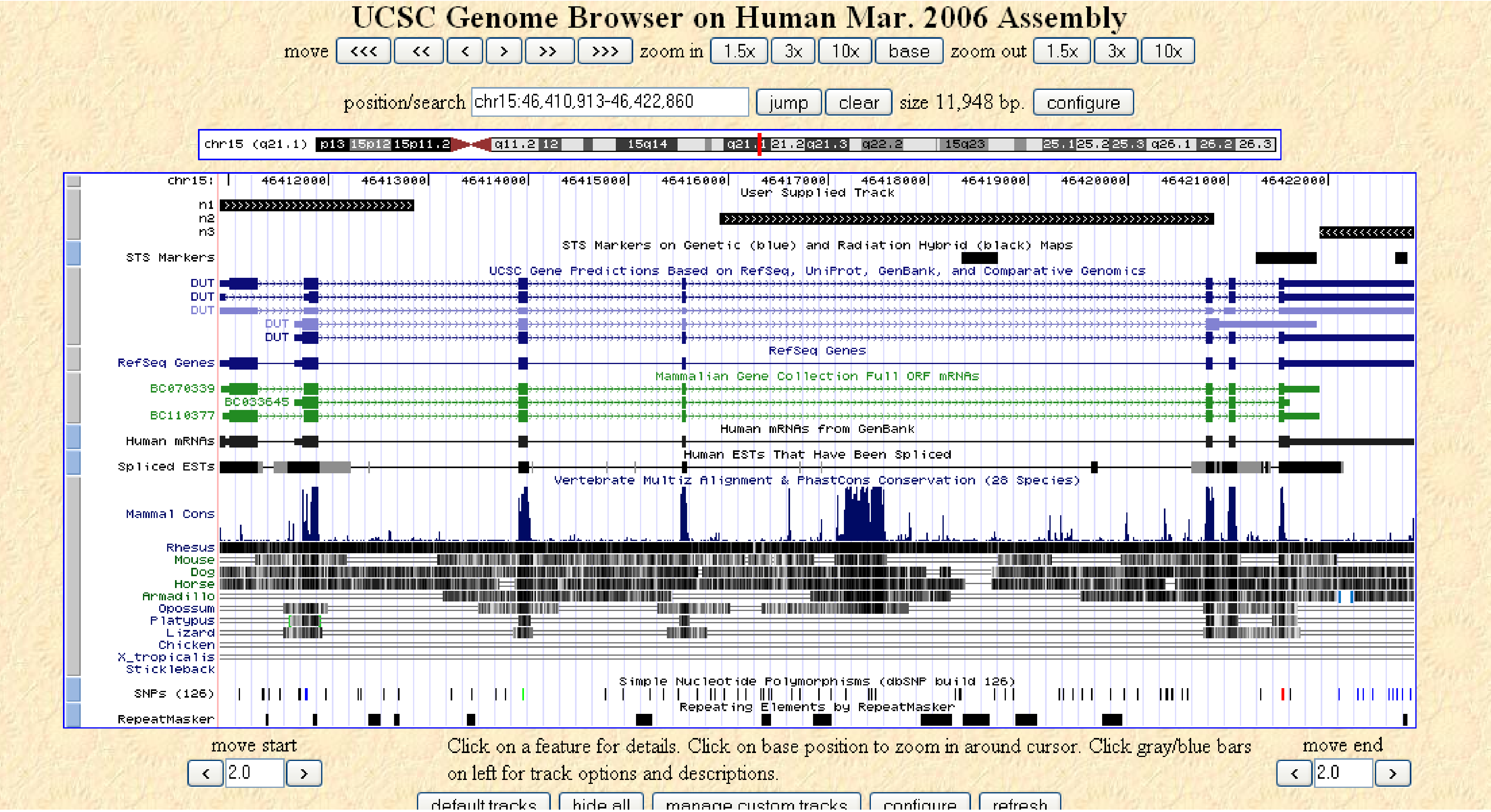
Click on a feature for details. Click on base position to zoom in around cursor. Click gray/blue bars on left for track options and descriptions.

Use drop down controls below and press refresh to alter tracks displayed.

Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing Tracks

Как это выглядит?



UCSC Genome Browser on Human Mar. 2006 Assembly

position/search chr15:46,410,913-46,422,860 size 11,948 bp.

Click on a feature for details. Click on base position to zoom in around cursor. Click gray/blue bars on left for track options and descriptions.

Different perspectives on Bioinformatics

- Bioinformatics is a tool
  - Biologists, biochemists, medical professionals, etc.
  - Obtain meaningful and understandable results
- Bioinformatics is a discipline
  - Informaticians, mathematicians, statisticians, etc.
  - Generate meaningful and understandable results

Summary

- Bioinformatics is truly interdisciplinary
  - Biology (natural sciences), informatics, mathematics & statistics
- Databases
  - Large, semistructured, incomplete, inaccurate
- Wide-range of problems
  - Solutions employ knowledge from sciences with algorithms and models from informatics, mathematics, and statistics

- Веб-страница для курса
- http://intbio.org/bioinf2018

- Связь с лекторами
- bioinf2018@googlegroups.com
-

  Онлайн опрос
- https://goo.gl/forms/0RDO3xnIqiotvFYz1