Курс «Компьютерная геномика». Лекция 1

Операции с последовательностями, доступ к базам и форматы данных. Базы данных последовательностей ДНК EMBL и GenBank

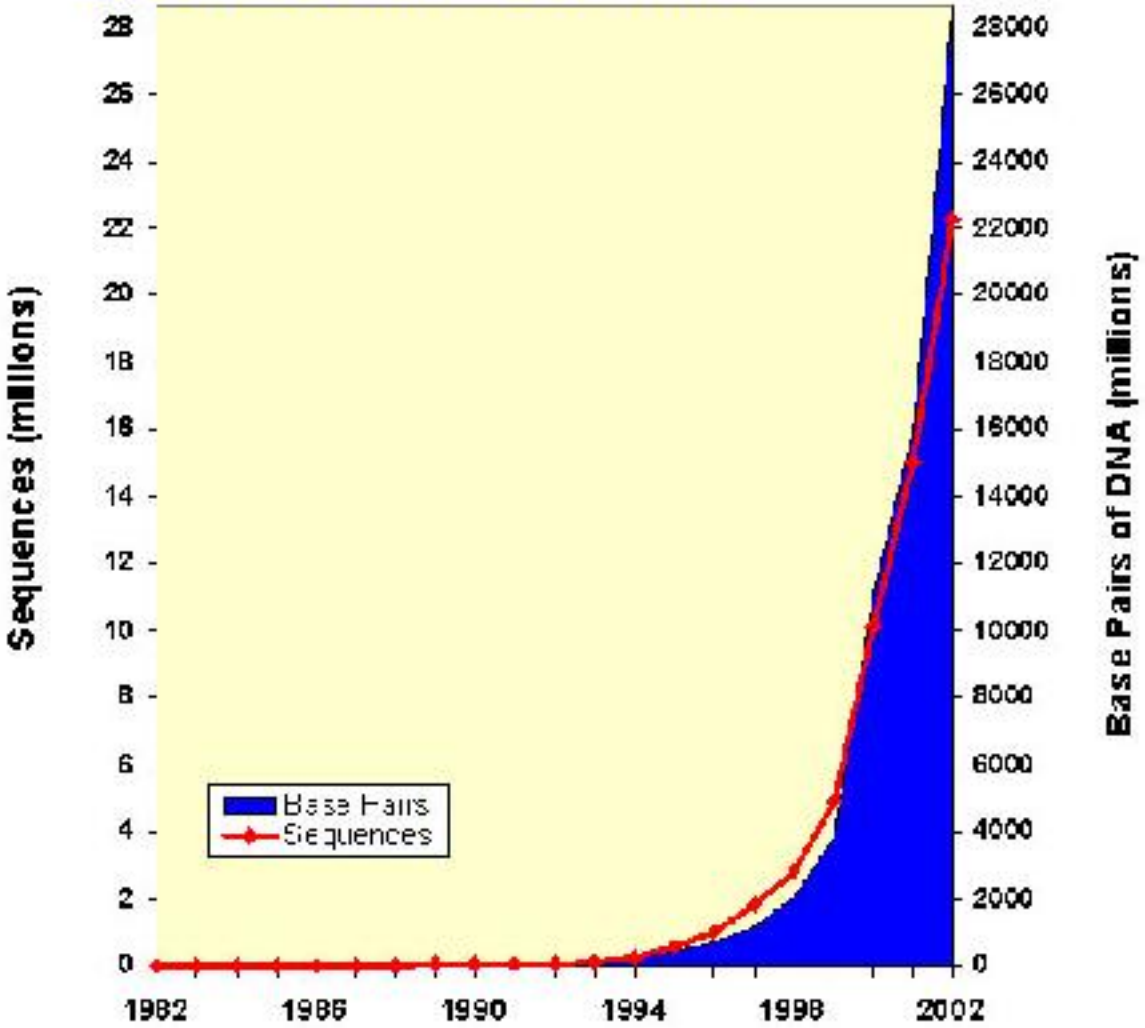
Орлов Юрий Львович

Кафедра информационной биологии ФЕН НГУ

Международные проекты геномных исследований

Стремительно растут темпы исследований по секвенированию геномной ДНК (Benson et al., 2000; Wheeler et al., 2000). На сентябрь 2003 г. доступны 139 полных геномов прокариот, включая 16 видов архебактерий и 123 вида бактерий.

Рост объема GenBank



http://www.ncbi.nih.gov/Genbank/genbankstats.html

GenBank за 2002 год:

28,507,990,166 п.о.

22,318,883 последовательностей

Полностью секвенированы эукариотические геномы мышевидного салата *Arabidopsis thaliana*, червя *Caenorhabditis elegans*, плодовой мушки *Drosophila melanogaster*, дрожжей *Saccharomyces cerevisiae* и *Schizosaccharomyces pombe*, некоторых внутриклеточных паразитических организмов (*Plasmodium falciparum*, *Encephalitozoon cuniculi*).

Информационный взрыв в молекулярной биологии и генетике

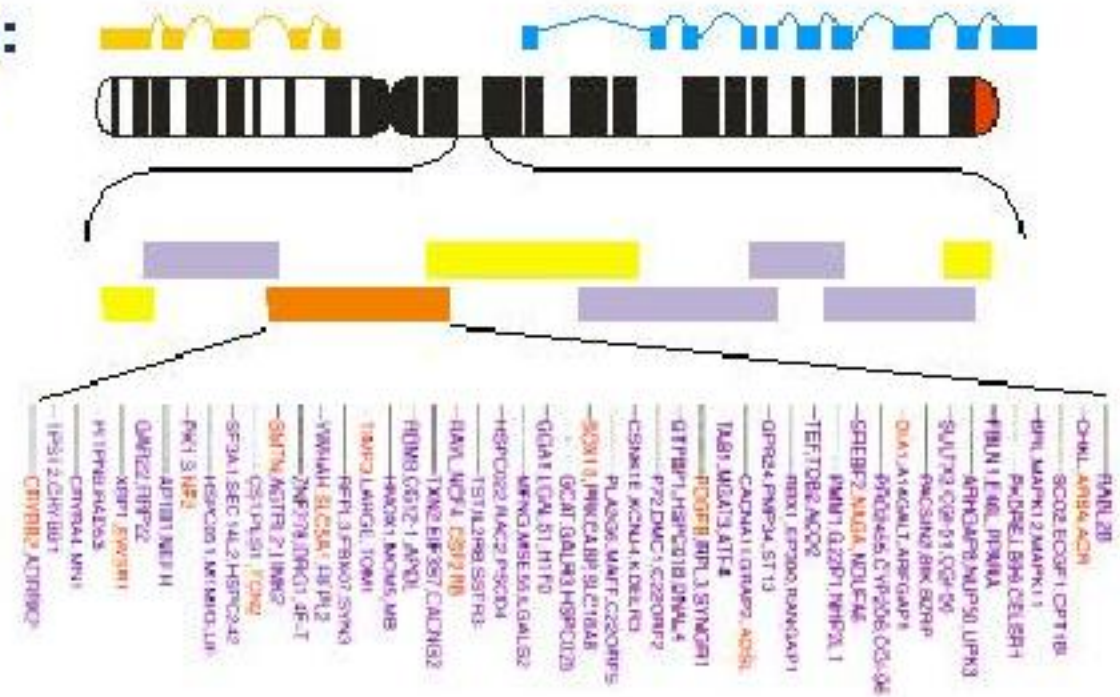
Характеристики генома человека:

размер - 3.3 миллиарда нуклеотидов: ATGCCCGGGTTTAATGCGTCAGTGACTGCACA..

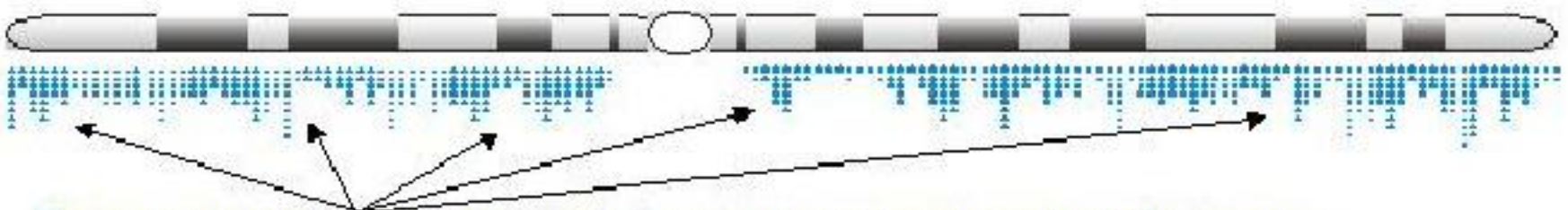
количество генов - 35000:

цитогенетическая карта:

физическая карта:



Локализация генов



Локализация мутаций: 1.5 миллиона полиморфизмов

Библиотеки геномных последовательностей

1000 томов

Человек
~ 3x 10^9 п.о.

1том

~ 10^6 п.о.

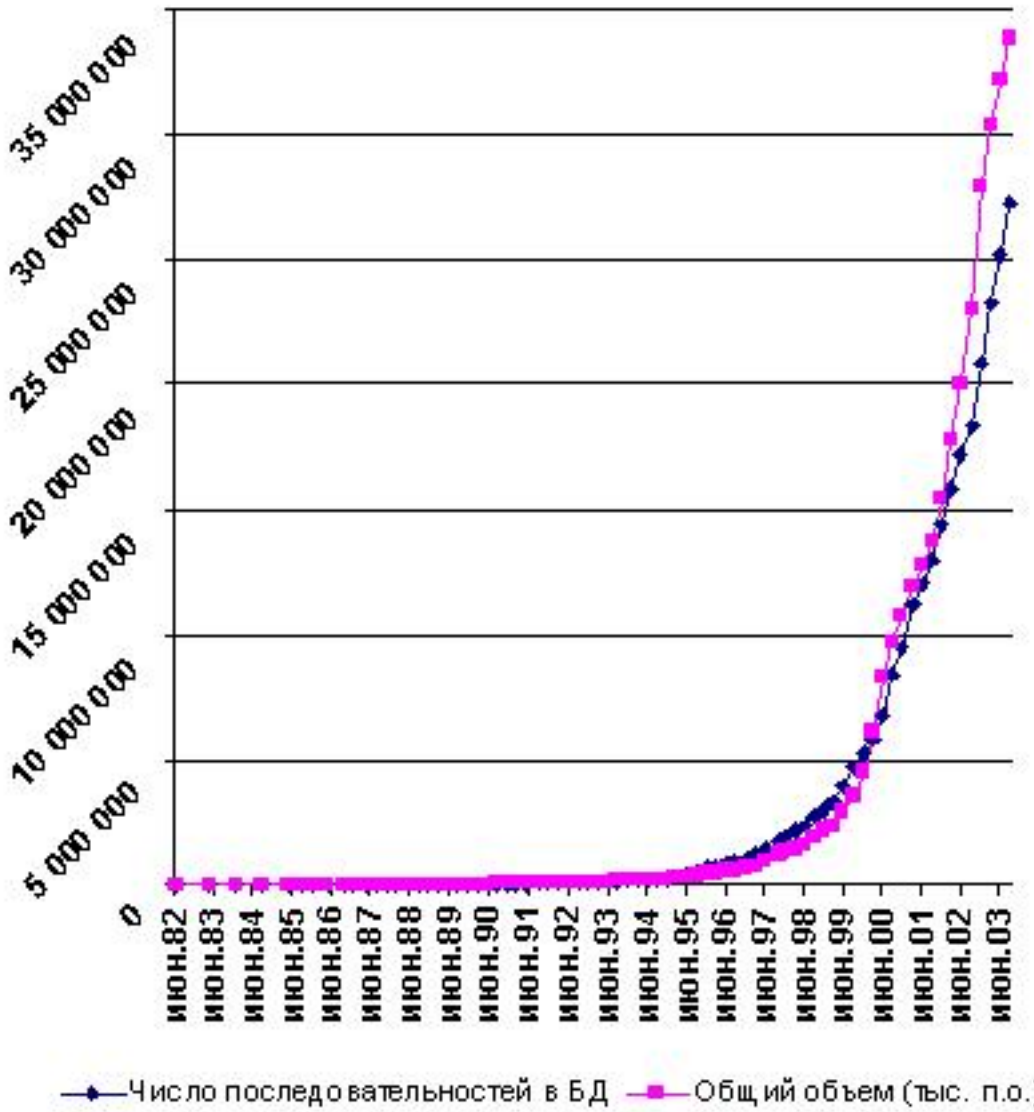
100 томов

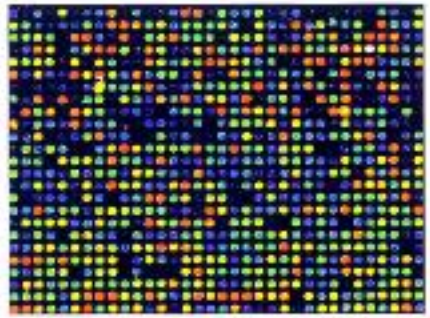
Дрозофила
~ 10^8 п.о.

10 томов

Дрожжи
~ 10^7 п.о.

Рост числа последовательностей в банке данных EMBL с 1982 по июнь 2003 г.



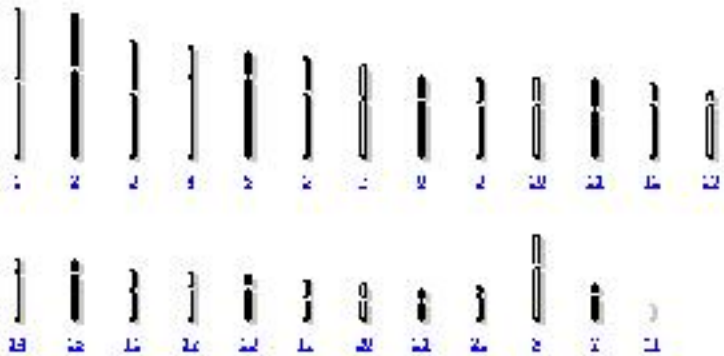


2000 г. – секвенирование генома человека

постгеномная эпоха

2003 г. – завершение чернового секвенирования

(6 государств, Россия с 1989 г.)



http://www.ebi.ac.uk

Дальнейшая структура лекции и курса

Лекция - Базы и банки данных.

Курс - Компьютерная геномика – задачи анализа генетических макромолекул

Структура лекции

- Интернет. Технические компьютерные средства поддержки БД.
- Классификация баз данных. История развития БД. Репозитарии и банки данных.
- Структура карточек GenBank/EMBL
- Поиск информации о нуклеотидных последовательностях. ENTREZ
- Запросы к БД. Перекрестные ссылки (линки).
- Поиск научных статей PubMed

Техническая основа работы с базами данных. Поиск информации в Интернет.

Пример:

http://www.bionet.nsc.ru/SRCG/index.html

http://тип документа (http = гипертекстовый документ)

www.bionet.nsc.ru адрес сервера в сети Internet

/SRCG/каталог сервера в котором находится HTML-документ **index.html** имя файла в котором находится HTML-документ

Сервис в Интернет: почта E-mail; FTP, Archie; Gopher; WWW, JAVA.

Просмотр гипертекстовых файлов (броузеры): программа "Netscape Communicator" “Internet Explorer”

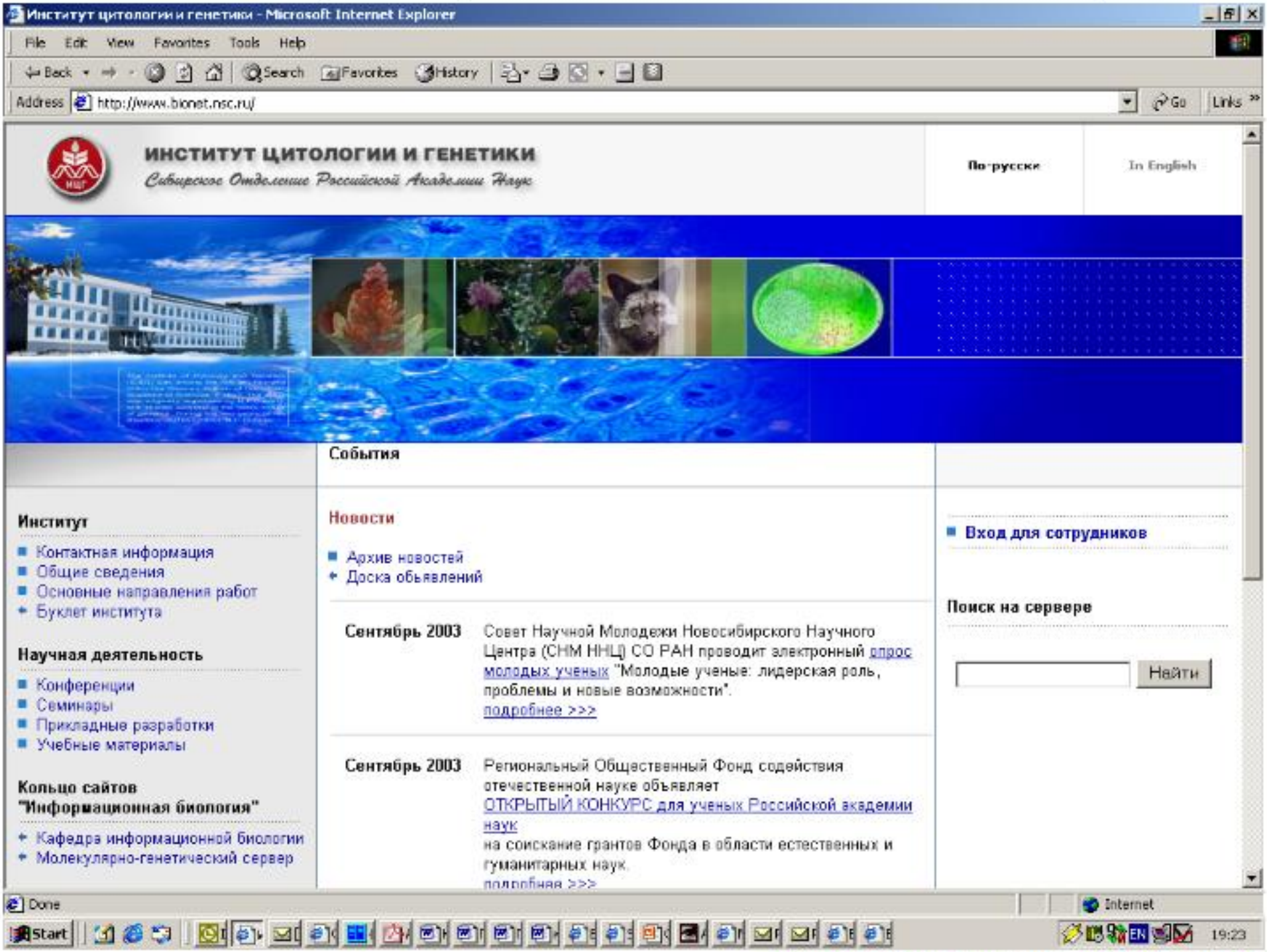
Поиск молекулярно-биологической информации проводится по тому же принципу, что поиск любой другой информации в Интернет с использованием поисковых систем.

Программы каталоги (Yahoo!, Excite). Программы поиска (Alta Vista

Сервер ИЦиГ СО РАН. Программа Netscape Navigator

«Компьютерная геномика» НГУ, Лекция 1, 2003

Тот же сайт просматривается с помощью Internet Explorer



Умение работать с Интернетом – необходимое практическое требование, без которого современная научная работа с информацией, с базами данных просто невозможна.

Список полезных сайтов:

Учебник по Интернет (Microsoft)	http://poms.microsoft.com/intl/ru/tutorial/default.htm
Средства поиска в Интернет (собраны вместе)	http://cui.unige.ch/meta-index.html
Список ссылок на поисковые системы	http://www.nsp.ru/search/
Средства поиска (Microsoft)	http://home.microsoft.com/intl/ru/access/allinone.asp
Поисковые сервера	
Yahoo!	http://www.yahoo.com/
AltaVista	http://www.altavista.com/
100hot	http://www.100hot.com
InfoSeek	http://www.infoseek.com/
Excite	http://www.excite.com/
Lycos	http://www.lycos.com/
WhoWhere	http://www.whowhere.com/
Российские поисковые системы	
Rambler	http://www.rambler.ru/
List.ru	http://www.list.ru/
Яndex	http://www.yandex.ru
@Rus	http://www.atrus.ru/
Апорт	http://www.aport.ru/
Weblist	http://weblist.ru/
Сервера бесплатной электронной почты	
Mail.Ru	http://mail.ru

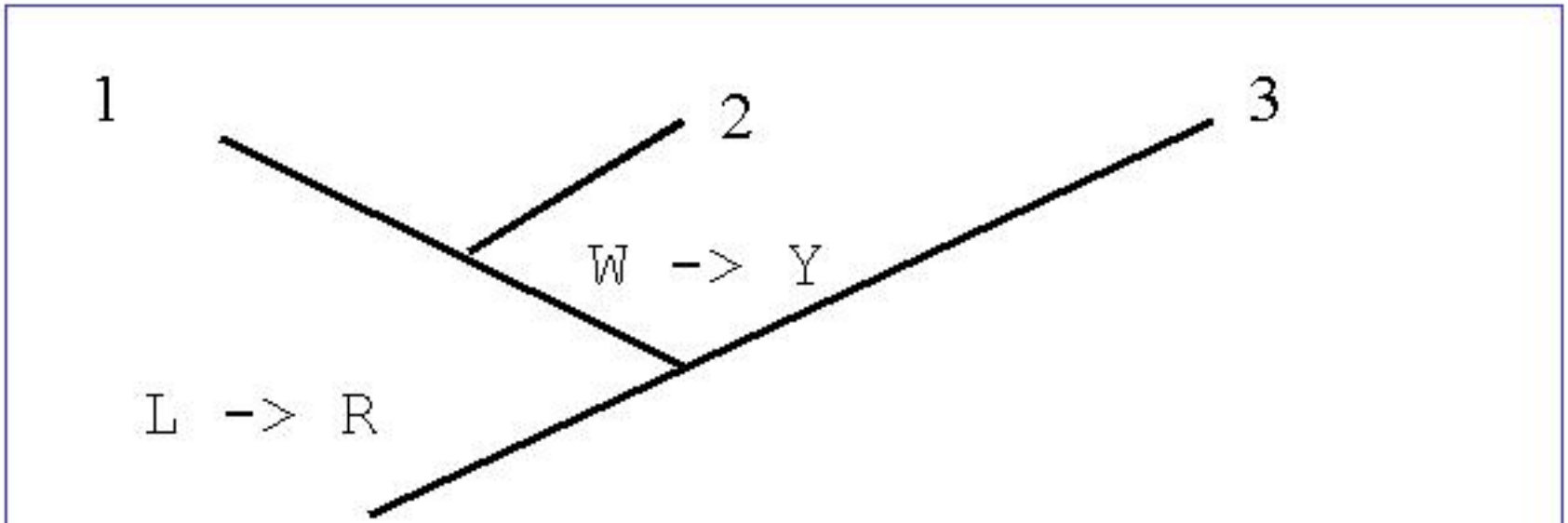
Базы данных последовательностей ДНК. Исторический обзор.

Первые последовательности, собранные в базах данных были аминокислотными последовательностями.

Методы секвенирования белков (Sanger and Tuppy, 1951)
Margaret Dayhoff (1925-1983)

National Biomedical Research Foundation (NBRF) (Вашингтон, США)
1960-е Protein Sequence Atlas
коллекция последовательностей, которая перешла в базу данных
Protein Informational Resource (PIR ранее Protein Identification Resorce)
http://watson.gmu.edu:8080/pirwww/index.html

1980-е PIR-International Protein Sequence Database
http://www-nbrf.georgetown.edu/pir
Сотрудничество международных центров NBRF, MIPS (Munich Center for Protein Sequences), JPID (Japan International Protein Information Database)
Данные были организованы в семейства и суперсемейства на основе схожести последовательностей.




```
Organism_1 AWTVASAVRTSI
Organism_2 AYTVAAAVRTSI
Organism_3 AWTVAAAVLTSI
```

Таблицы частот замен.

Были отобраны белки с не более 15% различий, чтобы наблюдаемые замены аминокислот отражали только одну замену, а не две.

Метод предсказания филогенетических отношений и вероятных аминокислотных замен в процессе эволюции родственных белковых последовательностей.

Матрицы аминокислотных замен PAM – матрицы Дайхоф (Dayhoff)
MDM (Mutation Data Matrix), или PAM (Percent Accepted Mutation)

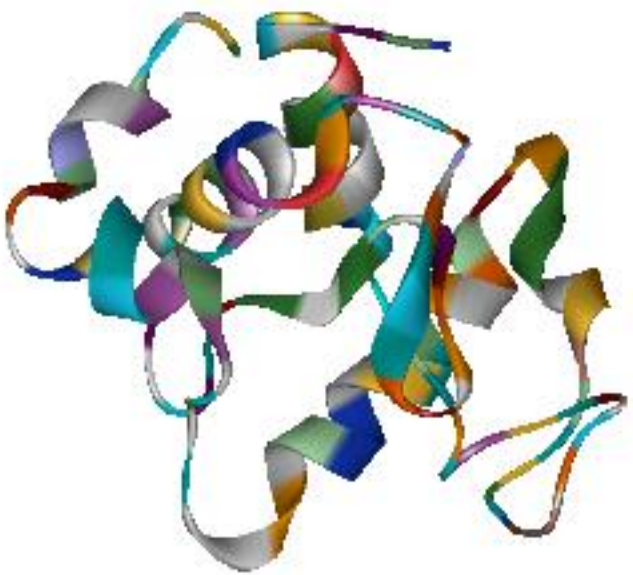
Научные задачи, решаемые с помощью анализа родственных белков

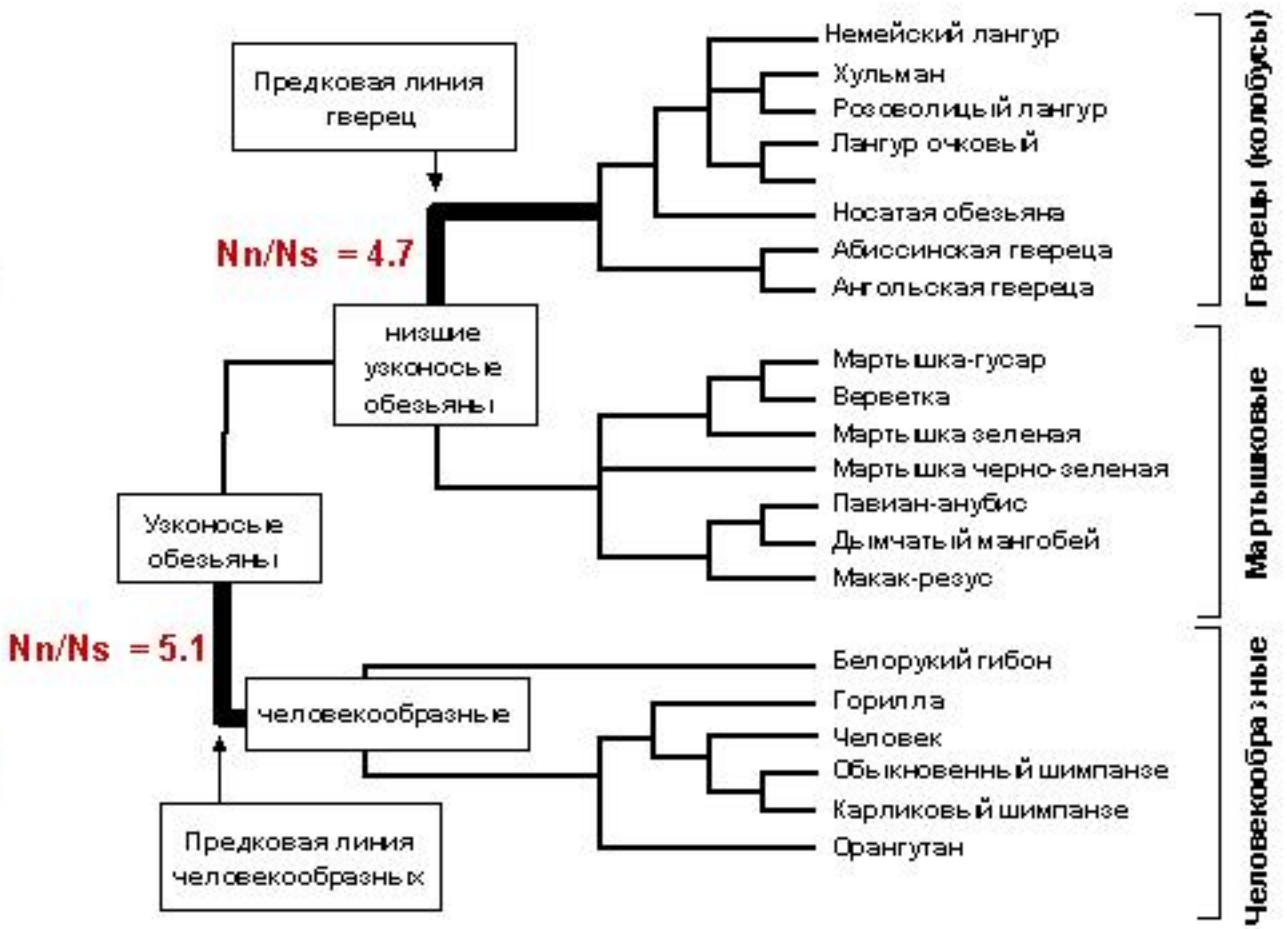
Молекулярный филогенетический анализ: адаптивная эволюция генов, кодирующих лизоцимы обезьян

Критерий адаптивной эволюции (Кимура, 1984):

Nn/Ns > 1

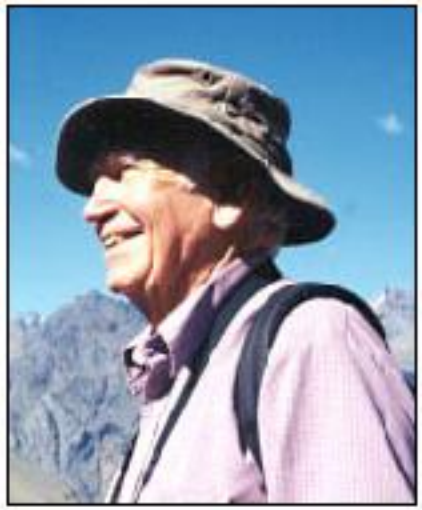
Здесь Nn и Ns - количество несинонимических замен, фиксировавшихся в исследуемом эволюционном маршруте





Recreating ancestral proteins, TREE, v15, N3, 109 – 114

Базы данных последовательностей ДНК



Walter Goad (1925-2000)

LANL (Los Alamos National Laboratory) Нью-Мексико, США

EMBL (European Molecular Biology Laboratory) Гейдельберг, Германия

1979 - прототип GenBank разрабатывался в LANL 1982-1992,

далее в NCBI (National Center for Biotechnology Information)

www.ncbi.nlm.nih.gov

1980 - EMBL Data Library (**www.ebi.ac.uk**)

1984 – DDBJ (**www.ddbj.nig.ac.jp**)

1983 - ГЕНЭКСПРЕСС и Программа "Геном человека" СССР (А.А.Баев)

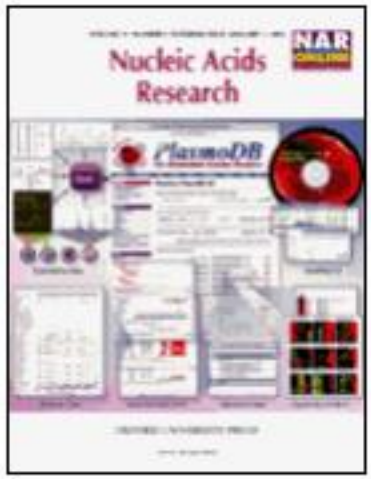
Сейчас эти центры образуют международный консорциум

International Nucleotide Sequence Database Collaboration

www.ncbi.nlm.nih.gov/collab

Описание баз данных (более 300 мол.биол.баз данных)

NAR - ежегодный первый выпуск (N1, январь) :



Nucleic Acids Research

OXFORD Journals online

http://nar.oupjournals.org/content/vol31/issue1/

Программы анализа последовательностей

Методы секвенирования ДНК

Maxam and Gilbert, 1977 - Sanger et al., 1977

Секвенирование

Программы анализа последовательностей ДНК, РНК и белков.

1982, 1984 - NAR - первые спец.

выпуски по программам

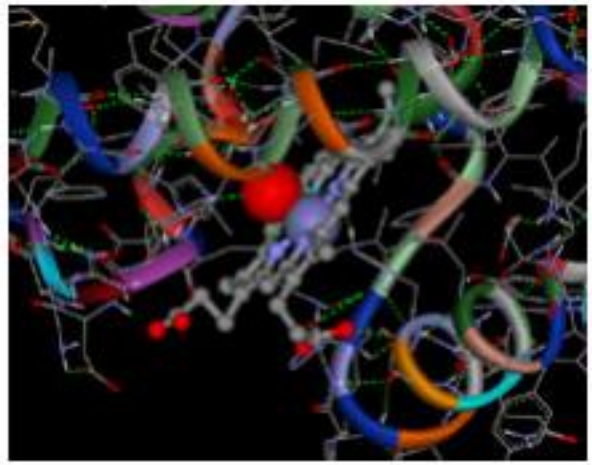
GCG на компьютерах VAX (Ун-т Висконсин)

www.gcg.com/

Intelligenetics, DNAStar

PHRED, PHRAP (www.codoncode.com)

коммерческие и некоммерческие пакеты



Методы, алгоритмы и современные программные реализации будут подробно рассмотрены в дальнейших разделах курса

Получение последовательностей из открытых баз данных

Важнейший шаг в обеспечении технического доступа к базам данных – разработка Web-страниц.

GENINFO – D.Benson, D.Lipman

Система ENTREZ (позднее включившая MedLine, NLM)

http://www.ncbi.nlm.nih.gov/entrez/

Новый релиз выходит каждые два месяца. Образован международный консорциум (International Nucleotide Sequence Database Collaboration), который состоит из Банка данных ДНК Японии (DNA DataBank of Japan, DDBJ), Европейской Молекулярно-биологической Лаборатории (the European Molecular Biology Laboratory, EMBL), и GenBank США (NCBI). Эти три организации обмениваются данными на ежедневной основе.

Коллекции молекулярно-биологических баз данных.

Основные репозитарии:

Andreas D. Baxevanis. The Molecular Biology Database Collection: 2003 update. Nucleic Acids Research, 2003, Vol. 31, No. 1 1-12.

(http://nar.oupjournals.org/cgi/content/full/31/1/1)

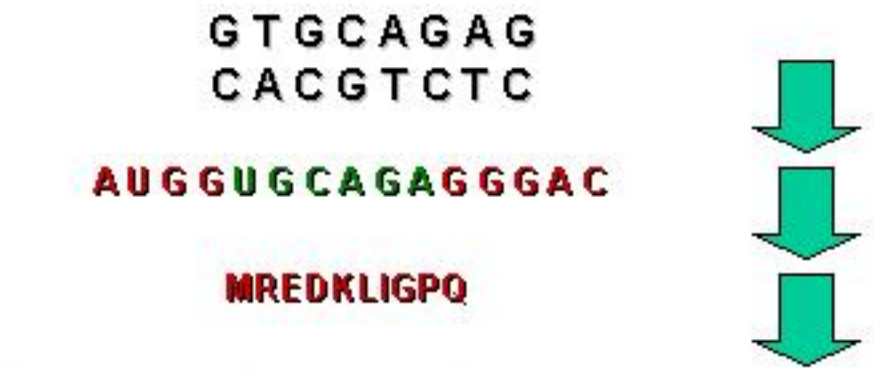
DNA Data Bank of Japan (DDBJ)	http://www.ddbj.nig.ac.jp	последовательности ДНК и белков, консорциум
EMBL Nucleotide Sequence Database	http://www.ebi.ac.uk/embl.html	-//-
GenBank	http://www.ncbi.nlm.nih.gov/	-//-
NCBI Reference Sequence Project	http://www.ncbi.nlm.nih.gov/RefSeq/	Природные биологическин молекулы
Ensembl	http://www.ensembl.org/	Аннотированная информация, Геномы эукариот
UCSC Genome Browser	http://genome.ucsc.edu/	Аннотация геномов
STACK	http://www.sanbi.ac.za/Dbases.html	Ген-ориентированные кластеры
TIGR Gene Indices	http://www.tigr.org/tdb/tgi.shtml	-//-
UniGene	http://www.ncbi.nlm.nih.gov/UniGene/	-//-

Ресурсы по молекулярной биологии могут быть классифицированы как **репозитарии, банки данных, базы данных, информационные сайты, электронные библиотеки и Интернет-доступное программное обеспечение**.

Каталог ресурсов: **http://wwwmgs.bionet.nsc.ru/mgs/links/links.html**

Специализированные базы данных можно подразделить в соответствии с иерархической организацией хранения и передачи наследственной информации:

- Уровень ДНК
- Уровень РНК
- Уровень белка
- Генные сети



SYSTEM FOR FORMALIZED DESCRIPTION, VISUALIZATION, AND MODELLING OF GENE NETWORKS

Эл. каталог ресурсов ИЦиГ СО РАН: **http://wwwmgs.bionet.nsc.ru/mgs/links/links.html**

[Links of IC&G SB RAS, Novosibirsk] [Priority links] [WWW-Biosciences] [GENOME] [NCBI National Center of Biotechnological Informatic] [Data bases] [Bio Tools] [EBI European Bioinformatics Institute] [BCM Baylor College of Medicine] [GDB Genome Database] [PDB Protein Data Bank Brookhaven National Lab.] [Patents] [PROGR] [France] [SilverPlatter]

Links of IC&G SB RAS, Novosibirsk:

INSTITUTE OF CYTOLOGY AND GENETICS, The Siberian Division of the Russian Academy of Sciences
The First International conference on Bioinformatics of Genome Regulation and Structure
Vogis Courier (Russian only)
Lab. of Experimental Modelling of Evolutionary Processes
Meiosis Laboratory Home Page
Sector of Molecular Evolution
X inactivation
Methods of Genetic Analysis Lab
The Laboratory of Animal Molecular Genetics

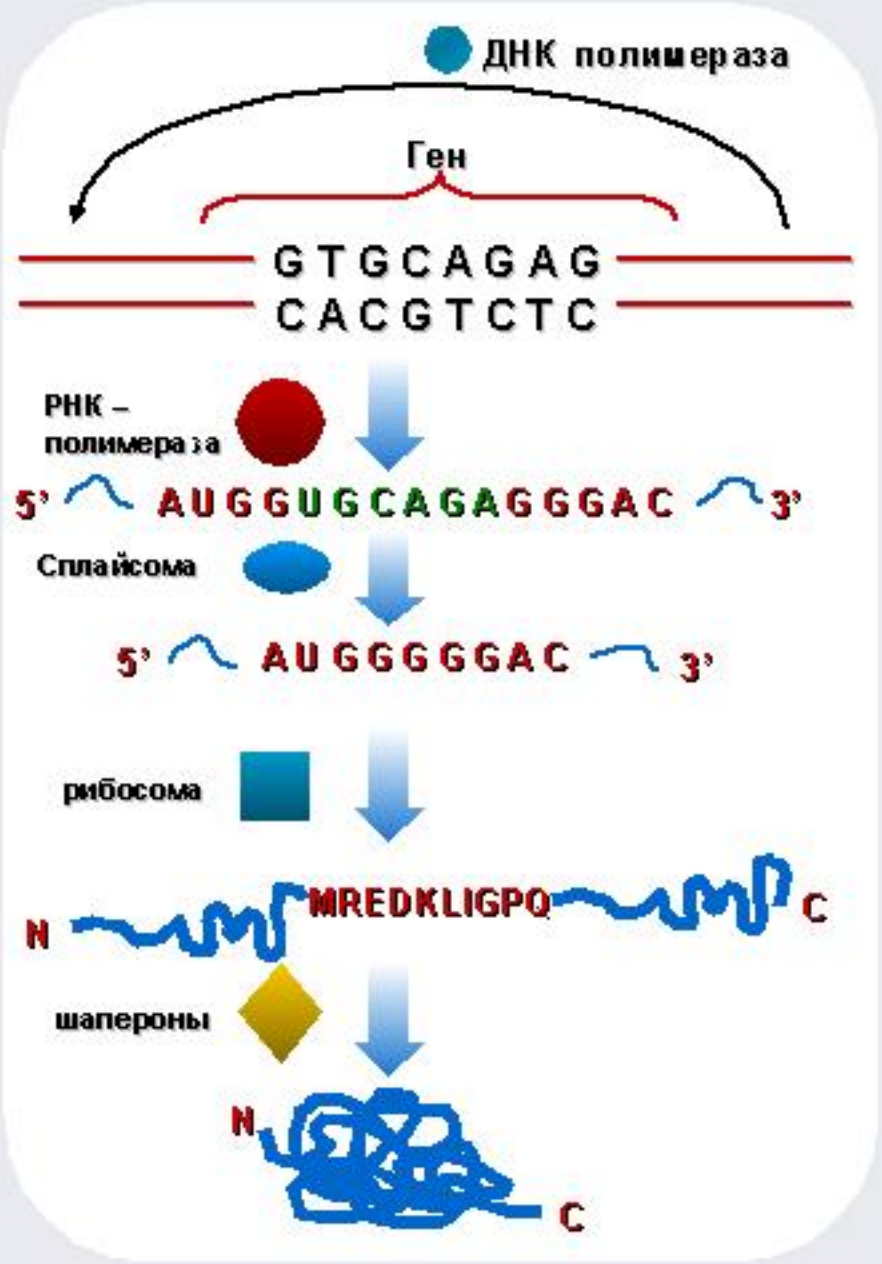
Полный каталог содержит более 1000 эл.ресурсов по молекулярной биологии

Priority links:

Novosibirsk Institute of Bioorganic Chemistry
Institute of Protein Research, RAS
Engelhardt Institute of Molecular Biology (EIMB)
A National Laboratory for Computational Science and Engineering (SDSC)
Molecular Bioinformatics of Gene Regulation , (Transfac DATA Base)
Institute of Advanced Biomedical Technologies (ITBA)

Базы данных и электронные ресурсы по биоинформатике

Фундаментальные генетические процессы



Геномная ДНК

репликация

Пре-мРНК

трансляция

мРНК

сплайсинг

Полипептидная цепь

транскрипция

белок

укладка полипептидной цепи

ОСНОВНЫЕ НАПРАВЛЕНИЯ (ГРУППЫ) БАЗ ДАННЫХ ПО МОЛЕКУЛЯРНОЙ БИОЛОГИИ (Более 300 баз данных)

- Сравнительная геномика (Comparative Genomics)
- Экспрессия генов с помощью микрочипов
- Структура и регуляция генов
- Генетическое и физическое картирование
- Геномные базы данных.
- Межмолекулярные взаимодействия
- Метаболические пути и клеточная регуляция
- Мутации
- Белки
- РНК
- Структуры (Пространственные)
- Трансгенные растения и животные
- Другие направления (Литературные базы данных, таксономия)

Andreas D. Baxevanis. The Molecular Biology Database Collection: 2003 update. Nucleic Acids Research, 2003, Vol. 31, No. 1 1-12.

(http://nar.oupjournals.org/cgi/content/full/31/1/1)

ОСНОВНЫЕ НАПРАВЛЕНИЯ БАЗ ДАННЫХ ПО МОЛЕКУЛЯРНОЙ БИОЛОГИИ (NAR, 2003)

Сравнительная геномика (Comparative Genomics)
Clusters of Orthologous Groups (COG) http://www.ncbi.nlm.nih.gov/COG
содержит филогенетическую классификацию белков из 43 полных геномов

Экспрессия генов с помощью микрочипов (Gene Expression) microarray

Структура и регуляция генов (Gene Identification and Structure)
SNP Consortium database http://snp.cshl.org SNP Consortium data
TRRD http://www.bionet.nsc.ru/trrd/ Transcription regulatory regions of eukaryotic genes

Генетическое и физическое картирование (Genetic and Physical Maps)

Геномные базы данных (Genomic Databases)
ACeDB information http://www.acedb.org/ *C. elegans*, *S. pombe*, and human sequences and genomic information
GeneCards http://bioinfo.weizmann.ac.il/cards/ Integrated database of human genes, maps, proteins and diseases
GOLD http://igweb.integratedgenomics.com/GOLD/ Information regarding complete and ongoing genome projects

Межмолекулярные взаимодействия (Intermolecular Interactions)
BIND http://bind.ca Molecular interactions, complexes and pathways

ГЕННЫЕ СЕТИ. Метаболические пути и клеточная регуляция (Metabolic Pathways and Cellular Regulation)
EcoCyc http://ecocyc.org/ *Escherichia coli* K-12 genome, metabolic pathways, transporters and gene regulation
EpoDB http://www.cbil.upenn.edu/EpoDB/ Genes expressed during human erythropoiesis
Kyoto Encyclopedia of Genes and Genomes (KEGG) http://www.genome.ad.jp/kegg Metabolic and regulatory pathways
RegulonDB http://www.cifn.unam.mx/Computational_Genomics/regulondb/ *Escherichia coli* transcriptional regulation and operon organization

Мутации
Mutation Databases
Human Gene Mutation Database (HGMD) http://www.hgmd.org Known (published) gene lesions underlying human inherited disease

ОСНОВНЫЕ НАПРАВЛЕНИЯ БАЗ ДАННЫХ ПО МОЛЕКУЛЯРНОЙ БИОЛОГИИ (NAR, 2003)

БЕЛКИ (Protein Databases)

Kabat Database http://immuno.bme.nwu.edu/ Sequences of proteins of immunological interest

PIR-NREF http://pir.georgetown.edu/pirwww/pirnref.shtml Non-redundant reference database with comprehensive protein sequences

SWISS-PROT/TrEMBL http://www.expasy.ch/sprot Curated protein sequences

TRANSFAC http://transfac.gbf.de/TRANSFAC/index.html Transcription factors and binding sites

Мотивы в белках (Protein Sequence Motifs)

ChuSTr http://www.ebi.ac.uk/clustr/ Automatic classification of SWISS-PROT+TrEMBL proteins

Pfam http://www.sanger.ac.uk/Software/Pfam/ Multiple sequence alignments and hidden Markov models of common protein domains

PROSITE http://www.expasy.org/prosite Biologically-significant protein patterns and profiles

Протеомные ресурсы (Proteome Resources)

Proteome Analysis Database http://www.ebi.ac.uk/proteome/ Online application of InterPro and chuSTr for the functional classification of proteins in whole genomes

Системы поиска информации (Retrieval Systems and Database Structure)

TESS http://www.cbil.upenn.edu/tess Transcription element search system

РНК (RNA Sequences)

ACTIVITY http://util.bionet.nsc.ru/databases/activity.html Functional DNA/RNA site activity

RNA Modification Database http://medlib.med.utah.edu/RNAmods/ Naturally modified nucleosides in RNA

UTRdb/UTRsite http://bighost.area.ba.cnr.it/srs6/ 5'- and 3'-UTRs of eukaryotic mRNAs and relevant functional patterns

Пространственные структуры (Structure)

HSSP http://www.sander.ebi.ac.uk/hssp/ Structural families and alignments; structurally-conserved regions and domain architecture

PDB http://www.pdb.org/ Structure data determined by X-ray crystallography and NMR

SCOP http://scop.mrc-lmb.cam.ac.uk/scop Familial and structural protein relationships

Трансгенные растения и животные (Transgenics)

Другие направления (Varied Biomedical Content)

PubMed http://www.ncbi.nlm.nih.gov/PubMed/ MEDLINE and Pre-MEDLINE citations

Tree of Life http://phylogeny.arizona.edu/tree/phylogeny.html Information on phylogeny and biodiversity

Поиск литературной информации.

База MEDLINE

Определение базы данных. Поле, вход, запись.

Адрес базы данных MEDLINE : *http://www.ncbi.nlm.nih.gov/PubMed/*

Простой поиск.

Формат запроса: word1 (AND word2 OR word3 ….)

Сложный поиск.

Формат запроса: word1[FieldCode] AND word2[FieldCode]

Сохранение найденной информации

(принцип Cut&Paste)

Ключевые слова структурированы в специальном словаре для поиска литературной (медицинской и др.) информации

NATIONAL LIBRARY OF MEDICINE | MEDICAL SUBJECT HEADINGS MeSH

Contact NLM | Site Index | Search Our Web Site | NLM Home

Health Information | Library Services | Research Programs | New & Noteworthy | General Information

Каталог доступен на сервере NCBI
http://www.nlm.nih.gov/mesh/

MeSH Browser

- Online searching of MeSH vocabulary
- About the MeSH Browser

All About MeSH

- MeSH Fact Sheet
- Presentations and papers by MeSH staff.
- Online introductory material to the *Annotated MeSH*
- Information from Previous Years
- Suggestions for authors' keywords

Obtaining MeSH

- Download electronic copies.
- Ordering information for Printed versions.

What's New

MeSH Staff

- Biographies and email.
- Publications and presentations.

Related Efforts

- Congenital Abnormalities associated with Mental Retardation (MCA/MR).
- Unified Medical Language System (UMLS®)
- NLM Classification

MeSH Suggestions

- Send MeSH vocabulary suggestions.

MEDLINE/PubMed

Ссылки и резюме статей из 4600 биомедицинских журналов

What's New | MeSH Fact Sheet | MeSH Browser | MeSH Files | MeSH Publications | MeSH Staff | MeSH Suggestions

Понятие карточки банка данных. Пример EMBL

```
     ID - identification             (begins each entry; 1 per
entry)
     AC - accession number           (>=1 per entry)
     SV - sequence version           (1 per entry)
     DT - date                       (2 per entry)
     DE - description                (>=1 per entry)
     KW - keyword                    (>=1 per entry)
     OS - organism species           (>=1 per entry)
     OC - organism classification    (>=1 per entry)
     OG - organelle                  (0 or 1 per entry)
     RN - reference number           (>=1 per entry)
     RC - reference comment          (>=0 per entry)
     RP - reference positions        (>=1 per entry)
     RX - reference cross-reference  (>=0 per entry)
     RG - reference group            (>=0 per entry)
     RA - reference author(s)        (>=0 per entry)
     RT - reference title            (>=0 per entry)
     RL - reference location         (>=1 per entry)
     DR - database cross-reference   (>=0 per entry)
     CC - comments or notes          (>=0 per entry)
     AH - assembly header            (0 or 1 per entry)
     AS - assembly information       (0 or >=1 per entry)
     FH - feature table header       (0 or 2 per entry)
     FT - feature table data         (>=0 per entry)
     XX - spacer line                (many per entry)
     SQ - sequence header            (1 per entry)
     CO - contig/construct line      (0 or >=1 per entry)
     bb - (blanks) sequence data     (>=1 per entry)
     // - termination line           (ends each entry; 1 per
entry)
```

Текстовый файл, длина строки ограничена,

Ключевые слова жестко заданы

http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html

Пример карточки EMBL, содержащий ген протеин-киназы дрозофилы

```
ID   DMCAPKCA   standard; DNA; INV; 3959 BP.
XX
AC   M18555; J03504;
XX
SV   M18555.1
XX
DT   16 JUL 1988 (Rel. 16, Created)
DT   06-JUL-1988 (Rel. 20, Last updated, Version 1)
XX
DE   D.melanogaster cAMP dependent protein kinase catalytic subunit,
DE   complete cds.
XX
KW   cAMP dependent protein kinase; kinase; protein kinase.
XX
OS   Drosophila melanogaster (fruit fly)
OC   Eukaryota; Metazoa; Arthropoda; Tracheata; Hexapoda; Insecta; Pterygota;
OC   Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Ephydroidea;
OC   Drosophilidae; Drosophila.
XX
RN   [1]
RP   1-3959
RX   MEDLINE; 88115281.
RA   Foster J.L., Higgins G.C., Jackson M.R.;
RT   "Cloning, sequence, and expression of the Drosophila cAMP-dependent protein
RT   kinase catalytic subunit gene";
RL   J. Biol. Chem. 263:1676-1681(1988).
XX
DR   FLYBASE; FBgn0002873; Pka C1.
DR   SWISS-PROT; P12370; KAPC_DROME.
XX
CC   Draft entry and computer readable copy of sequence [1] kindly
CC   provided by J.L.Foster (17-MAR-1988).
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..3959
FT                   /db_xref="taxon:7227"
```

Идентификатор

Код доступа

Описание последовательности

Литературная ссылка

Разметка

Пример карточки EMBL, продолжение

```
FT   source          1..3959
FT                   /db_xref="taxon:227"
FT                   /organism="Drosophila melanogaster"
FT   misc_feature    603..701
FT                   /note="cpa repetitive element"
FT   CDS             935..1966
FT                   /db_xref="FLYBASE:FBgn0002273"
FT                   /db_xref="SWISS-PROT:P12370"
FT                   /note="cAMP-dependent protein kinase catalytic subunit"
FT                   /protein_id="AAA28412.1"
FT                   /translation="MGNNAIAAGKVDAAETVKPPLEQVKEEVFDVKNDAKFIVCARDA
FT                   FERIKTLGTGSFGRVMLVKHKPTKQYYAMKILDKQKVVKLKQVEHTLNEKRILQAIDFP
FT                   FLVSLRYHFKDNSNLYMVLEYVTGGELFSHLERVGRFSEDRSRFYAAQIVLAFEYLHSL
FT                   DLIYRDLKPENLLIDSQGYLKVTDFGFAKRVKGRTWTLCGTPEYLAPEIILSKGYNKAV
FT                   DWWALGVLVYEMAAGYPPFFADQPIQIYEKIVSGKVRFPSHFSGDLKDLLRNLLQVDLT
FT                   KRYGNLKAGVNDIKNHQWFASTDWIAIYQRKIEAPFIPKFKGPGDTSNFDDYEEEALRI
FT                   NFEKCCAKEYWEE"
XX
SQ   Sequence 3959 BP; 1142 A; 943 C; 873 G; 957 T; 0 other;
     aagcttattt attcagtgat tatagccagc tttttattaa gcccaataaa agtgccagca        60
     aagcagctgag aacaaaaaca atatgaaatg aaaaatagca aagaagccaaa atacctaaga       120
     aacttaaaat ataatctata tgtgtgcata aagaactcgg aaaacgggac tctcatcttc       180
     agtcccttac tgaattaccc gttaatttgt gccgtttcct ctcccgttcc tcttcgtcagc      240
     ttctcaaatt agagtaaatc gcgatcataa gcaacctcaa agctctgaaac caaatgctag       300
     tctaagccta aagtgtataa ctagtttaag acctcgcaaa ttaaagaaac cgatacgtgc       360
     taaaatcata attcaaagtt gattcaaagt aagtcccttt aaacgcaatc taaatacaaa       420
     tttaaaagga tattcagcga ataaagagat gatagaagctc cagtgacaac ctcgctctag      480
     ccgataagga gcaactattt catcgcgacc gttgacccgt aactcggcaa aatagctaag       540
     tttattttgt agttcttttaa aaagaaaaag tttttatata gttttatatc ggttgaggtt      600
     aagctacagg cgcagtcgag ttgcgagtag gatagcagct gcagcgcccc tcgtcagcag       720
     cttgttgttt taagagtatg ttttttgtag ttcttgattt gttttggttt gggtggtatt       780
     cagagagagc cagcgagaga agagagagag gagcgacgag ccctagcact ccgagcgacga      840
     tcaggccgaa gtgtcggccg gcgcaacctt ccgttgtcca tcgctaacgg gcaccatctc       900
     taagatgggc aatgcatgtt ttgatgtgaa tatgaaggtt gatgttgttg agatggtgaa       960
     gagagttcct gagcagccca agaagagatt cagagacaag tggcgagcga atcccaccaa      1020
                                    . . .
     cacagccaca cgctgctggt cattaataaa ttataataag ttttatgtcc aaaaataatg      3360
     agcaaaacca aagcaagagc tgtgtgcgaa tgccgcataa agtttagaaa ttgtgggtaa      3420
     atgccaaaaa ctaaaaaacg aaacgcaatc aagagctgct aattctgtga ataaaagttc      3959
//
```

Разметка

Разметка кодирующих частей гена

Аминокислотная последова-тельность

Последова-тельность ДНК

Кроме форматов EMBL/GenBank, предназначенных для хранения информации существуют форматы для компьютерной обработки и анализа последовательностей :

NBRF-PIR

GCG

Plain/ASCII Staden

Genetic Data Environment (GDE)

Fasta/Pearson format

Intelligenetics

PIR/CODATA

ASN.1 Abstract Syntax Notation

NEXUS

Fasta

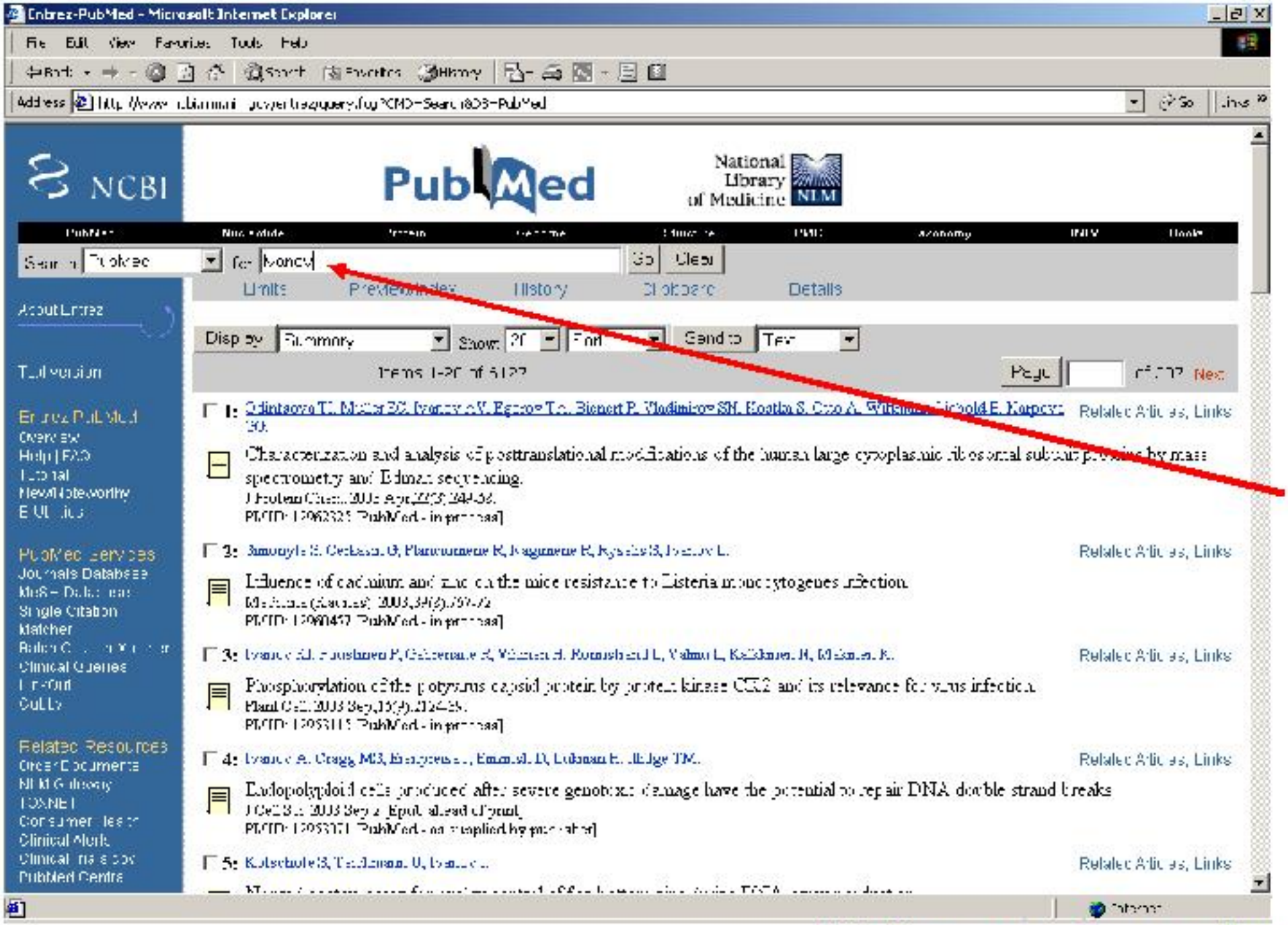
```
>seq1
aggctgct agct agct
>seq2
aactaact
```

NBRF

```
>DL;seq1
seq1., 16 bases, 268 checksum
aggctgctag ctagct*
>DL;seq2
aactaact
```

Entrez – система навигации по базе данных GenBank и электронной библиотеке PubMed

Entrez –электронная библиотека PubMed



Просмотр найденной записи – резюме статьи

Characterization and analysis of posttranslational modifications of the human large cytoplasmic ribosomal subunit proteins by mass spectrometry and Edman sequencing.

Odintsova TI, Muller EC, Ivanov AV, Egorov TA, Bienert R, Vladimirov SN, Kostka S, Otto A, Wittmann-Liebold B, Kiparisov GG.

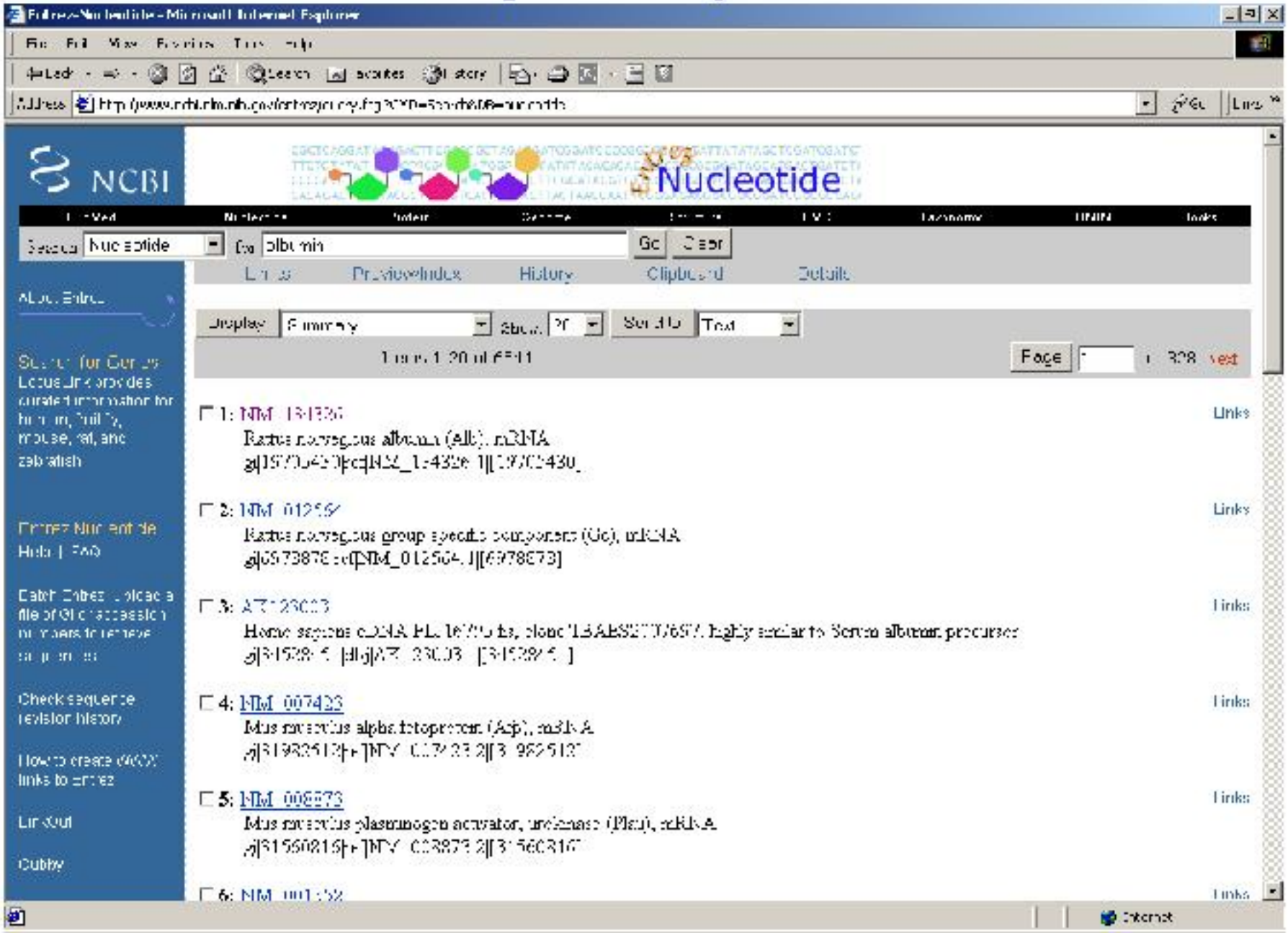
Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russian Federation.

The 60S ribosomal proteins were isolated from ribosomes of human placenta and separated by reversed phase HPLC. The fractions obtained were subjected to trypsin and Glu-C digestion and analyzed by mass fingerprinting (MALDI-TOF), MSMS (ESI), and Edman sequencing. Forty-six large subunit proteins were found, 32 of which showed masses in accordance with the SwissProt database (June 2002) masses (proteins L6, L7, L9, L10, L12, L17, L18, L21, L22, L24, L26, L27, L30, L32, L34, L35, L36, L37, L37A, L38, L39, L40). Eleven (proteins L7, L10A, L11, L12, L13A, L23, L23A, L27A, L28, L29, and P0) resulted in mass analysis data inconsistent with N-terminal loss of methionine, acetylation, trimethylation, or hydroxylation. A loss of methionine without acetylation was found for protein L8 and L17. For nine proteins (L3, L4, L5, L7A, L10, L14, L15, L31, and L40), the molecular masses could not be determined. Proteins P1 and protein L3-like were not identified by the methods applied.

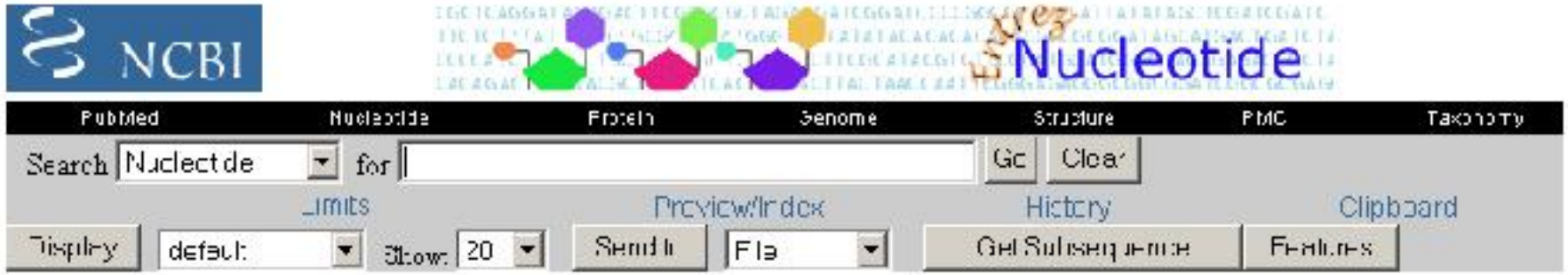
PMID: 12568225 [PubMed - in process]

Пример – статья найденная по ключевому слову Ivanov (автор)

Поиск по ключевым словам в базе данных GenBank. Запрос «альбумин».



Пример найденной карточки GenBank



1: NM_134326 Rattus norvegicus ...]gi:19705430]

```
LOCUS       NM_134326                1956 bp    mRNA    linear   ROD 16-SEP-2003
DEFINITION  Rattus norvegicus albumin (Alb), mRNA.
ACCESSION   NM_134326
VERSION     NM_134326.1  GI:19705430
KEYWORDS    .
SOURCE      Rattus norvegicus (Norway rat)
  ORGANISM  Rattus norvegicus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae;
            Rattus.
REFERENCE   1  (bases 1 to 1956)
  AUTHORS   Huot,J., Papet,I., Bechereau,F., Denis,P., Buffiere,C., Gimonet,J.,
            Glomot,F., Elyoutfi,M., Ertaille,D. and Obled,C.
  TITLE     Increased albumin plasma efflux contributes to hypoalbuminemia only
            during early phase of sepsis in rats
  JOURNAL   Am. J. Physiol. Regul. Integr. Comp. Physiol. 284 (3), R707-R713
            (2003)
  MEDLINE   22458063
   PUBMED   12571074
  REMARK    GeneRIF: an accelerated plasma efflux of albumin contributes to
            hypoalbuminemia only during the early period of sepsis
REFERENCE   2  (bases 1 to 1956)
  AUTHORS   Boulak,J., Dangers,M. and Thum,T.
  TITLE     Aroclor 1254 modulates gene expression of nuclear transcription
            factors: implications for albumin gene transcription and protein
            synthesis in rat hepatocyte cultures
```

Описание полей в GenBank:

Аналогичная структура

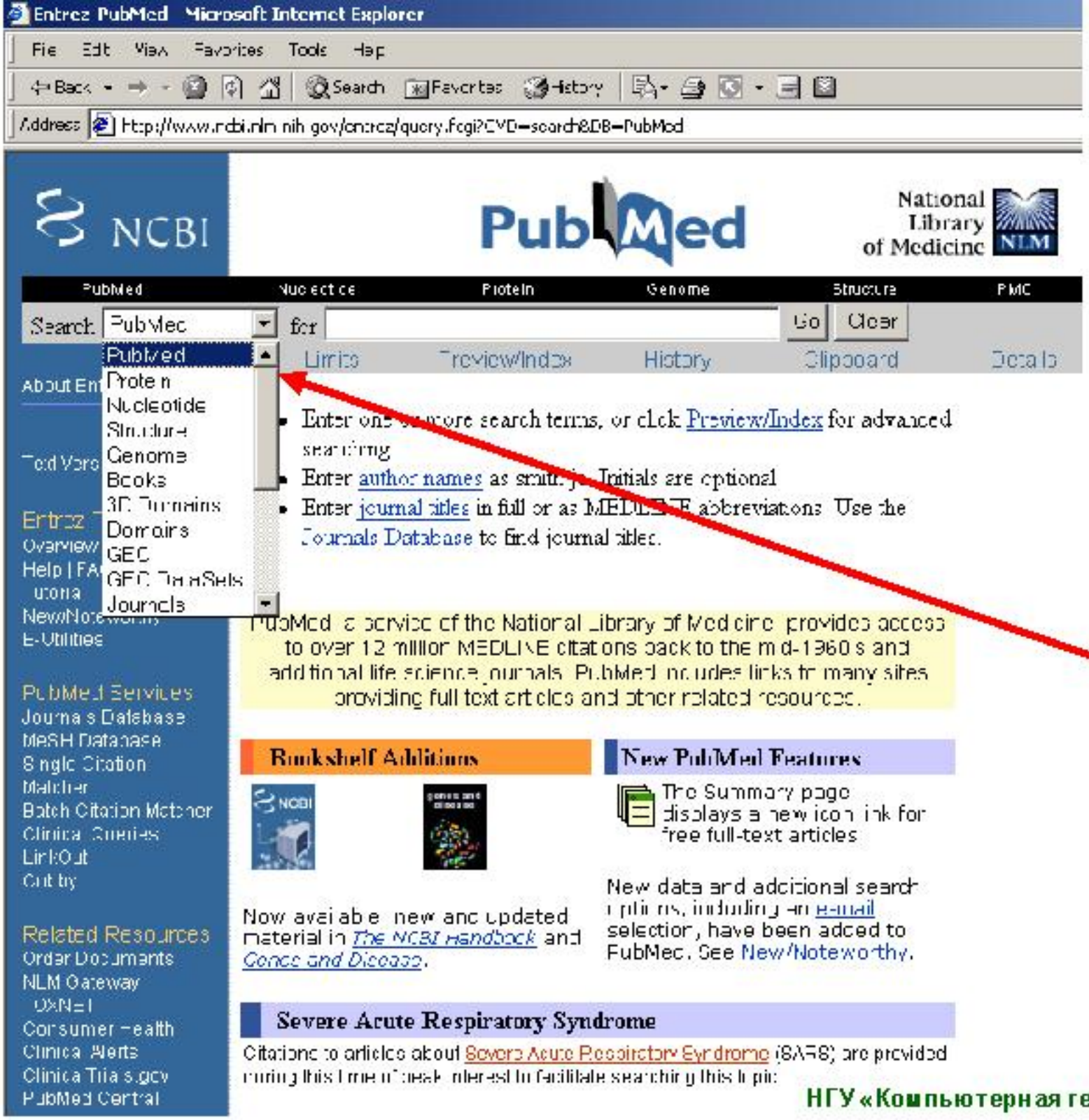
Более длинные ключевые слова (11)

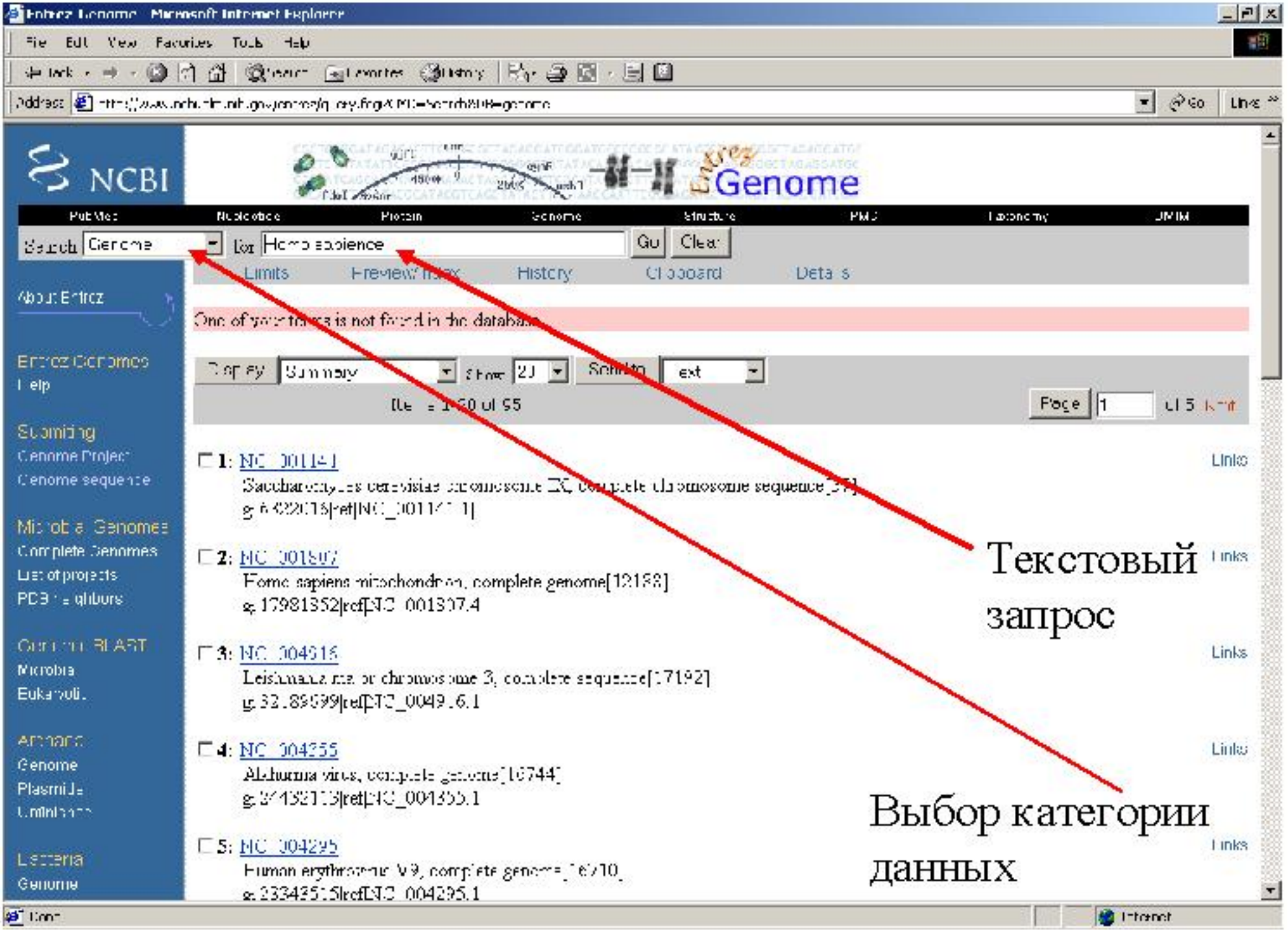
```
                     /note="reading frame (pro peptide)"
BASE COUNT      545 a    492 c    470 g    441 t
ORIGIN
        atgaagtggg taacctttat gttcctcctc ttatctgctt ttctaggggt
       61 ctatttccgc cagaaagcaca cagaagtgag atcgcccatc gatttcaagca ctcaggaggaa
      121 cagcatttca aggcctagt cctgattgcc ttttccgagt atctccagga atgcccatat
      181 gaagagtata tcaaattggt gtaggaagta acagacttta caaaaacatg tgttgttgat
      241 cagcaatccc aaaactgtga cagtccatt cacactctct tccgacacaa ctcatgcgcc
      301 atcccaaagc ctcgtgacaa ctacggtgaa ctggctgact gctgtgcaaa gcaagagccc
      361 gaaagaaatg agttttttt gtagtcatag gatgacaatt ttaatttgtc atttttttag
      421 aggccgcagc ctgacgccat gtgcacctcc ttccagtgaga accctaccag ctctccgga
      481 cactatttgc atgaagttgc caggagacat cattatttct atgccccaga acccctttac
      541 tatgttgaga aattacaatga ggttttgatc tagttgttta tagagtttga taaatgtagc
      601 tgcctgacac ccaaccttga tgccgtgaaa gaggaaagcat tgctcccacc tgccgtcac
      661 aggatgaagt gctccagtat gcagagattt ggagagagag ccttcaaagc ctgggcagca
      721 gttgttatga gttagagatt tttaaattgt gagttttgag aaattaattaa attggttaca
      781 cacctcacca aaatcaacaa ggagtactat cacagccaca tgttgcaatg cgccgatgac
      841 agggcagaac ctgccaagta catgtgtgag aacgaggcca ctatctccag caaacctgcag
      901 gcctgctgtg ataagccagt gctgcagaaa tcccagtgtc tcgctgagac agaacatgac
      961 cacctcccag ccgatctgcc ctcaatagct gctgactttg tggaggatca ggacgtgtgt
     1021 aagcactatg ctgaggccaa ggatgtcttc ctgggcacgt ttttgtatga atattcaaga
     1081 aggcaccccg attactcgtt gtcctgctcg ctgcgtactg ctcaagaaata tgaagccaca
     1141 ctggagcagt gctgtgctga aggcgatcct cctgcctgct acggccccagt gcctgccgac
     1201 tttcagcctc ttgtcgaaga acctaagaac ttggtcaaaa ctcactgtga gctttacgag
     1261 aagctcgtag agtatggatt ccaaaacgtc gttgtggttc gatacaccca gaaagtcacct
     1321 caggtgtcgc ccccactct cgtggaggca gcagacaacc tgggacgagt gggcaccaag
     1381 tgttgtaccc ttcctgaagc tcagagactg cctgtgtgg aagactatct gtctgccatc
     1441 ctgaaccgtc gtgtgttgct gcatgaaaag accccagtga gcgaagaggt caccaagtgc
     1501 tgcagtgggt cttcggtgga aagacggcca tgtttctctg ctctgccagt tgacgagacc
     1561 tatgtcccca aagagtttaa agctgagatc ttcaccttct actctgatat ctgcaccctc
     1621 ccagacaagg agaagcagat aaagaagtca acgggtttct cgtgggttga aaacacaag
     1681 ccccaggccc ccgacgatcc gctgaagagg gtgatgggtg acttcgccca ctcgtgggac
     1741 aagtgttgca aggctgcga caaggataac tgcttcgcca ctgagggcgc caaccttgtt
     1801 gctaggaagc aaggaggttt agttcaaaca tatgacaacc agttcaggct acctgaggaa
     1861 caacagccat gcagcctcag gatcatctc ttctgttggt gtcaaccccca cacccgcagg
     1921 aacccaaatt tctttgaaca tttgatttct ttttc
//
```

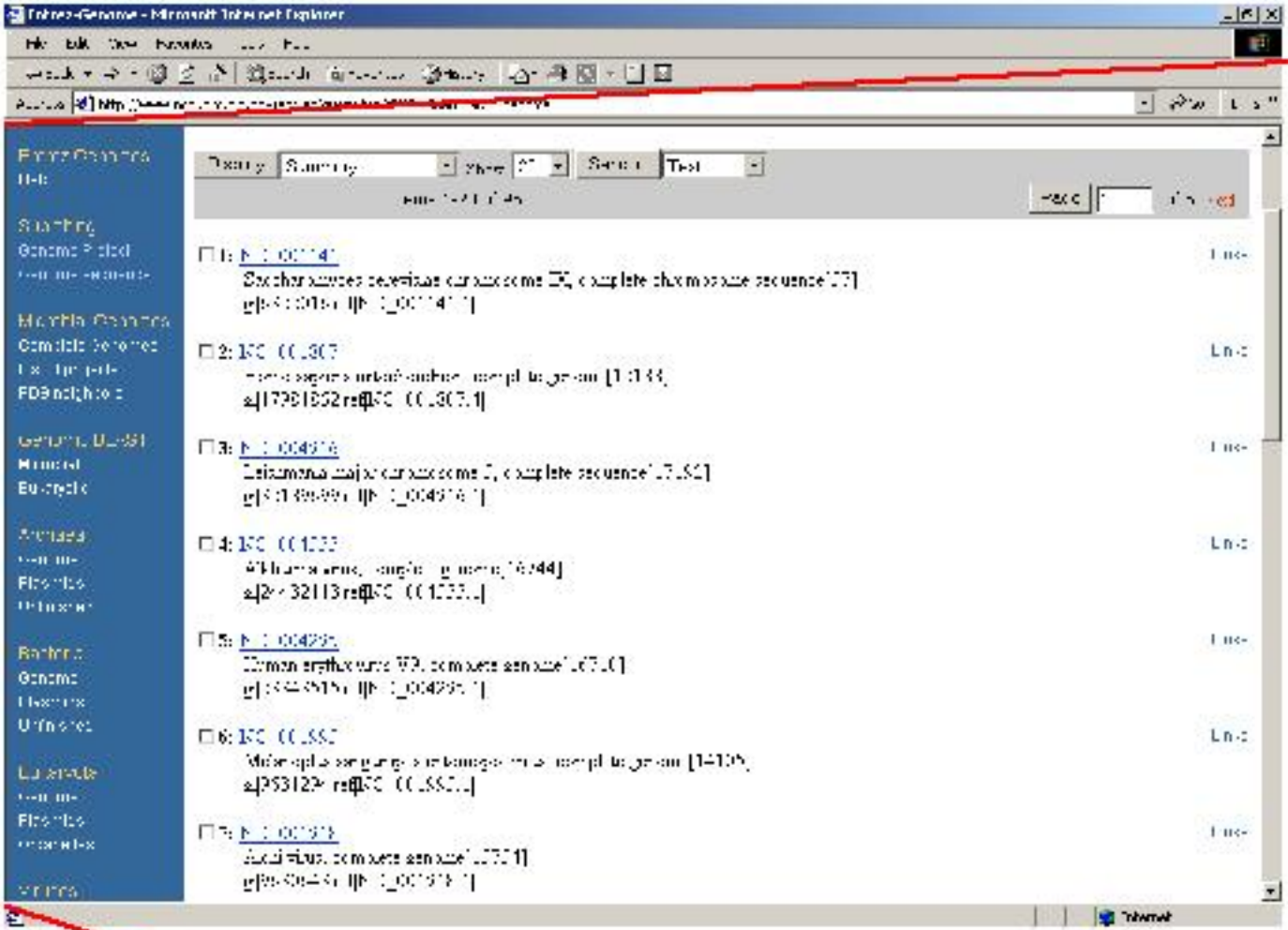
Продолжение карточки GenBank.

Описание последовательности с 11-й позиции, нумерация,

Ключевое слово ORIGIN


Entrez PubMed Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=PubMed
NCBI
PubMed
National Library of Medicine
NLM
PubMed Nucleotide Protein Genome Structure PMC
Search PubMed for Go Clear
Limits Preview/Index History Clipboard Details
PubMed
Protein
Nucleotide
Structure
Genome
Books
3D Domains
Domains
GEC
GEC Taxonomy
Journals
Enter one or more search terms, or click Preview/Index for advanced searching
Enter author names as smith j. Initials are optional
Enter journal titles in full or as MEDLINE abbreviations. Use the Journals Database to find journal titles.
PubMed, a service of the National Library of Medicine, provides access to over 12 million MEDLINE citations back to the mid-1960's and additional life science journals. PubMed includes links to many sites providing full text articles and other related resources.
Bookshelf Additions
Now available new and updated material in The NCBI Handbook and Genes and Disease.
New PubMed Features
The Summary page displays a new icon link for free full-text articles
New data and additional search options, including an e-mail selection, have been added to PubMed. See New/Noteworthy.
Severe Acute Respiratory Syndrome
Citations to articles about Severe Acute Respiratory Syndrome (SARS) are provided during this time of peak interest to facilitate searching this topic.
About Entrez
Text Version
Entrez
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby
Related Resources
Order Documents
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central


NCBI
Entrez Genome
Search Genome for Homo sapience
Go
Clear
Limits
Preview/Index
History
Clipboard
Details
About Entrez
Entrez Genomes
Help
Submitting
Genome Project
Genome sequence
Microbial Genomes
Complete Genomes
List of projects
BLAST
Microbial
Eukaryotic
Archaea
Genome
Plasmids
Bacteria
Genome
Links
Текстовый запрос
Выбор категории данных



Entrez Genomes
Help

Submitting
Genome Project
Genome sequence

Microbial Genomes
Complete Genomes
List of projects
PDB neighbors

Genomic BLAST
Microbial
Eukaryotic

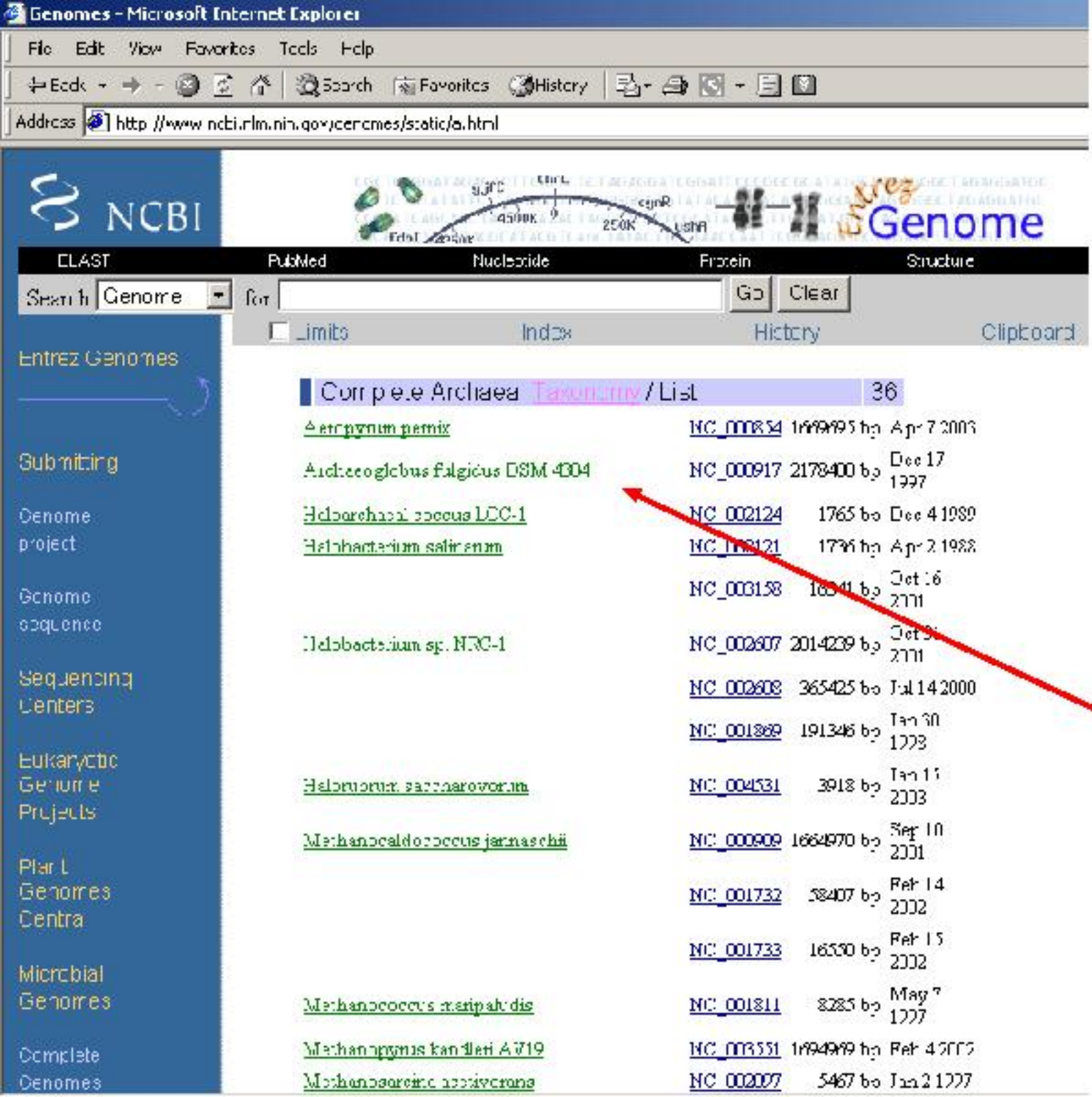
Archaea
Genome
Plasmids
Unfinished

Bacteria
Genome
Plasmids
Unfinished

Eukaryotic
Genome
Plasmids
Organelles

Viruses

Рассмотрим классификация геномов более подробно


Genomes - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Search Favorites History
Address http://www.ncbi.nlm.nih.gov/cgi-bin/Genomes/static/a.html
NCBI
Entrez Genome
BLAST PubMed Nucleotide Protein Structure
Search Genome for Go Clear
Limits Index History Clipboard
Entrez Genomes
Submitting
Genome project
Genome sequence
Sequencing Centers
Eukaryotic Genome Projects
Plant Genomes Centra
Microbial Genomes
Complete Genomes
Complete Archaea Taxonomy / List 36
Aeropyrum pernix NC_000854 1669695 bp Apr 7 2000
Archaeoglobus fulgidus DSM 4304 NC_000917 2178400 bp Dec 17 1997
Halobacterium sp. NRC-1 NC_002607 2014239 bp
Methanococcus jannaschii NC_000909 1664970 bp Sep 10 2001
Methanopyrus kandleri AV19 NC_003551 1694969 bp Feb 4 2002

NCBI

Taxonomy Browser

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

Search for [ ] As [complete name] [x] lock [Go] [Clear]

[Display] [3] lEvels using Filter: [none]

Archaeoglobus fulgidus DSM 4304

Entrez records	
Database name	Direct links
Nucleotide	174
Protein	4827
Genome	1
Taxonomy	1

Taxonomy ID: 224325
Rank: no rank
Genetic code: Translation table 11

Other names:

Archaeoglobus fulgidus str. DSM 4304[synonym]

Lineage (full)

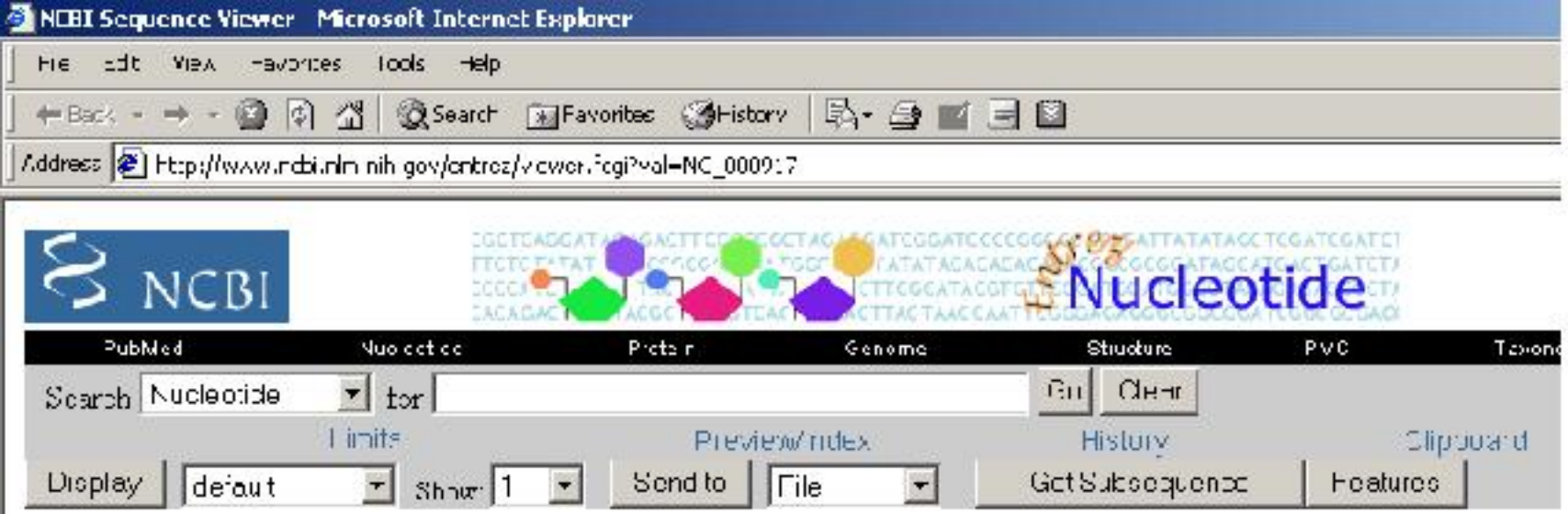
cellular organisms; Archaea; Euryarchaeota; Archaeoglobi; Archaeoglobales; Archaeoglobaceae; Archaeoglobus; Archaeoglobus fulgidus

Более подробная информация о полностью секвенированном геноме архебактерии *A.fulgidus*

Comments and References:

Klenk HP et al. (1997)

Klenk, H.P., Clayton, R.A., Tomb, J.-F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J.,



Карточка GenBank генома архебактерии *A.fulgidus*

Рассмотрим карточку GenBank NC_000917 для *A.fulgidus*

```
LOCUS       NC_000917              2178400 bp    DNA     circular BCT 13-JUN-2003
DEFINITION  Archaeoglobus fulgidus DSM 4304, complete genome.
ACCESSION   NC_000917
VERSION     NC_000917.1  GI:11497621
KEYWORDS    .
SOURCE      Archaeoglobus fulgidus DSM 4304
  ORGANISM  Archaeoglobus fulgidus DSM 4304
            Archaea; Euryarchaeota; Archaeoglobi; Archaeoglobales;
            Archaeoglobaceae; Archaeoglobus.
REFERENCE   1  (bases 1 to 2178400)
  AUTHORS   Klenk,H.P., Clayton,R.A., Tomb,J.-F., White,O., Nelson,K.E.,
            Ketchum,K.A., Dodson,R.J., Gwinn,M., Hickey,E.K., Peterson,J.D.,
            Richardson,D.L., Kerlavage,A.R., Graham,D.E., Kyrpides,N.C.,
            Fleischmann,R.D., Quackenbush,J., Lee,N.H., Sutton,G.G., Gill,S.,
            Kirkness,E.F., Dougherty,B.A., McKenney,K., Adams,M.D., Loftus,B.,
            Peterson,S., Reich,C.I., McNeil,L.K., Badger,J.H., Glodek,A.,
            Zhou,L., Overbeek,R., Gocayne,J.D., Weidman,J.F., McDonald,L.,
            Utterback,T., Cotton,M.D., Spriggs,T., Artiach,P., Kaine,B.P.,
            Sykes,S.M., Sadow,P.W., D'Andrea,K.P., Bowman,C., Fujii,C.,
            Garland,S.A., Mason,T.M., Olsen,G.J., Fraser,C.M., Smith,H.O.,
            Woese,C.R. and Venter,J.C.
  TITLE     The complete genome sequence of the hyperthermophilic,
            sulphate-reducing archaeon Archaeoglobus fulgidus
  JOURNAL   Nature 390 (6658), 364-370 (1997)
  MEDLINE   98049343
   PUBMED   9389475
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from AE000782.
FEATURES             Location/Qualifiers
     source          1..2178400
                     /organism="Archaeoglobus fulgidus DSM 4304"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:224325"
     gene            complement(436..786)
```

Рассмотрим далее поля карточки GenBank

```
FEATURES             Location/Qualifiers
     source          1..2178400
                     /organism="Archaeoglobus fulgidus DSM 4304"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:224325"
     gene            complement(406..786)
                     /locus_tag="AF0001"
     CDS             complement(406..786)
                     /locus_tag="AF0001"
                     /note="hypothetical protein; identified by GeneMark;
                     putative"
                     /codon_start=1
                     /transl_table=11
                     /product="hypothetical protein"
                     /protein_id="NP_068842.1"
                     /db_xref="GI:11497622"
                     /translation="MQLSIPFWSDFNSAFEEFVKLFLALSIPFWSDFNSISVSISLSM
                     PIPFQSHFGLISTEERGALFGSNPRFQSHFGLISIDTLFEDLRWNLADFQSHFGLIST
                     PRCLPCDPSRRLFQSHFGLISTIR"
     gene            complement(3222..3749)
                     /locus_tag="AF0002"
     CDS             complement(3222..3749)
                     /locus_tag="AF0002"
                     /note="hypothetical protein; identified by GeneMark;
                     putative"
                     /codon_start=1
                     /transl_table=11
                     /product="hypothetical protein"
                     /protein_id="NP_068843.1"
                     /db_xref="GI:11497623"
                     /translation="MKAASYGVPFQSHFGLISTIRGNDGLPRVVVFFQSHFGLISIVK
                     IVEVLEKDEALSIPFWSDFNCHLQMDQRIRLAFFQSHFGLISTGTLQTSGTNANFLSI
                     PFWSDFMLELLERIEKIEMLFQSHFGLISTRNMVWGRSELQRAIFQSHFGLISTQSSQ
                     LIMPVKMGTFNPILV"
     gene            1200..1784
```

Продолжение карточки GenBank

```
     gene            complement(2177643..2178146)
                     /locus_tag="AF2436"
     CDS             complement(2177643..2178146)
                     /locus_tag="AF2436"
                     /note="identified by sequence similarity; putative"
                     /codon_start=1
                     /transl_table=11
                     /product="conserved hypothetical protein"
                     /protein_id="NP_071258.1"
                     /db_xref="GI:11500312"
                     /db_xref="COG:COG1468"
                     /translation="MVVESELFVRGTEVSYFVCKTKLVLFSRNIAKEHESDSVKLGK
                     LVHRQHFSRDDKEVRICRVALDIVRDCSELSVVSVKKSDRMEKADVYQLAYYLYYLSK
                     HGVPARGRISYPKSRKNVSVELIENLLVKLKSLSEIKAIKSSSEMPKPEKKSYCTKCA
                     YYELCFS"
BASE COUNT   562096 a 527300 c 531003 g 558001 t
ORIGIN
        1 gaaaatgggt caatcggaaa ttgagtagga gaatagaagt cgcatgctac tataaatgag
       61 atgcaatttc gacacctcgc ggaagtatat caatgacagc agcctccgga aaccgttaca
      121 ttggaaatag agggaaaatt actgatggtt caaatccgac ccaaatgcga ttgaaagagc
      181 ctttcagccc tagtgtgagt gtcagggttta ctctttgttga aatccgacca aaatgcgatt
      241 gaaaggtttg ttaagggctt tgatttgctc ctcggcggct ctggttgcaa tccgaccaca
      301 atgggattga aagtaaagca gttcaccgct gttcctcggt tcactgcctt gttgaaatca
      361 gaccaaaatg gaattgaaag gtattgaaag gtatttgaat caatgaaaag aaagctgacc tcgttcgtga
      421 aatcagaccg aaatgggatt gaaagagtct tctcggatgg gcaccaggga gacatcgagg
      481 cgttgaaatc agaccaaaat gggattgaaa gtcagccagc ttacgtcgga gatcctcgaa
      541 gaggtatcca gttgaaatca gaccaaaatg gaattgaaag cgaggagctg tgccaaagag
      601 agcgcctcgt tcttcggttg aaatcagacc caactgcgat tgaacgacag tgcacctgct
      661 taaagaaatg ctgacagaaa ttgagttgaa atccgaccaa aatgcgattg aaagagccga
      721 gaagagcttg acgaattctt caaaagccga gttgaaatca gaccaaaatg ggattgaaag
      781 ttgcatttac atcggcagaa ttggtctcgt cggcaggcat gttgcaatca gaccacaatg
      841 ggattgaaag cggttctctt acgtactcat ccgcgaagcg agactcgcgt tggttgaaat
      901 cagaccaaaa tgggattgaa agaagcaagc gtgaaattga gcagtcaaaa cagaccgtta
      961 gttgaaatca gaccaaaatg ggattgaaag ttttcccatc ccattacgac ttcgccggca
     1021 aaaaagttga aatcagacca aaatgcgatt gaaagagccg gttcgaccac gtcgtcggtc
     1081 tgctgtccgc aagttgaaat cagaccaaaa tgggattgaa agtgttcgaa gtcagttgaa
```

Окончание карточки GenBank

. . .

```
2176561 ccctcgaaaa gcgagctctg aaccccattc aaatagcgcc tcaaaaactt cttgaccctg
2176621 ttgaccctct ccacgctccc gtcgtcggca actatgacgt acacaatcac cacccacaca
2176681 cgaacgggtga gtattttcca acggccgtaa actgctttac aaagtttctag catttgagtc
2176741 ttatcaaggcg ctggtaagac acttcctgct ttagcttcct gtgcttgacg gtctttctaa
2176801 gcctctcgct gcagtgcctg acgaacttct ttttccccta ctcgctgagc aacatcccgc
2176861 cgagcctgct gtcgaagtct gattccgtaa ccactcccgtt gttcacccaag tagcaaaatca
2176921 gcgtctcaac aattacgacc ttgaaaaagct ccccaatgtc gaagggcaga gtcaatcgtc
2176981 tctcagacgg ctcgtggacg tagcttattg ccggatctag ctgggtgtgg tacatttcag
2177041 aaagcacggc cgagtaaagc agggaatttc cgaagccgat catcgcattc acttcatttt
2177101 caggaggcat tctgccgctc ttccccgaacc caaaattttt gagtatttca tcgaatccag
2177161 tgtactaggc gtttccgcga gcagccttcg ccccctatac ctccaacaatg cgtttagcag
2177221 agttcagatg ctgcagagat gcgatgacgt ctgagtcatc ctgttccgcc tttttcagaa
2177281 ccgttgccat attgagaata gccgcctcca cgaaagctct ggcaaggtgc agcctccgct
2177341 cgctgtcaag gtgatgctca gcctgcctca aaccaaccgc accgaaacg aggctccccc
2177401 tcggatagaa tcagccgata tagtagcccgc aacggttgaa gaagtggatg caacactccct
2177461 cctttgcaag aaggcttatt gctttcgaag ttatgcccaa agagcccaga gccgtagattg
2177521 cgtagattga gttgatagga attggccttc tgccatcctc attctcaaag taaatggcgt
2177581 tctccgtgcc tctcagcttc ccatcagaca cgaggtagta attttcttt ctcatggccc
2177641 gatcacgaaa agcacagctc gtagtaggcg cacttcgtgc agtagctctt cttctcaggc
2177701 tttggcattg atgagctctt gattgcctta atctctccaa gaatgctttt caacttcacc
2177761 agtacattct cgtccagctc cacacttaca ttttttctgc tcttcggata gcttatcctc
2177821 ccgctcgccc taacacgctg ctgcgagagg tactaaaagt agtaggcaag gcgataagag
2177881 tcggccttct ccatcctgtc tgacttcttg acctcaacaa cccgcagctc ctcccccgc
2177941 ctcacgatgc ccagagccac cctgccaatc cgcaccccct tgccgtctct gctgaaatgc
2178001 tgcctgtgaa ccagccttcc cagcttaaca gcgtcgctct catgctccat ggcaatgctg
2178061 cgggagaaaa gccacagctt tgcctttgcaa accaagcagt agctccaccc cgttcccctt
2178121 acaaatagcc ctcccctcaac aacctatgat caaatatcag tttgttatca tatataacaa
2178181 tcgcgattcc tgcggaaata gtctggattc ggcaaaactcg gtcctttgag tgtttgaatg
2178241 tgttcgaaaa agctatcaacg ggctgtcaca gtcattcatc ttgagttta attgggcaga
2178301 actatgtccc atatccagct gtcctgcaga tccgttcatca tccgagcagt ttgatcgtct
2178361 aaagcttgag gaaaagattt taaagcaaaa atgtattcag
//
```

http://www.ncbi.nlm.nih.gov/genomes/framik.cgi?db=Genome&gi=131

Accession: NC_000917
Total Bases: 2178400 bp
Completed: Dec 17, 1997.

Feature table:
Protein coding genes
Structural RNAs

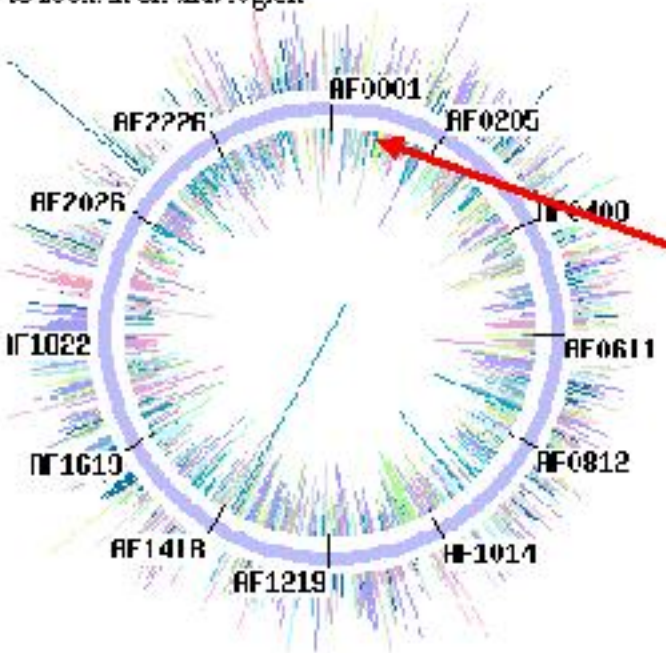
BLAST protein homologs:
COGs (Clusters of Orthologous Groups)
3D Structure (Sequences with known structure)
TaxMap (Sequences grouped by superkingdom)
TaxPlot (3-way genome comparison)
CDD(Conserved Domain Database)

Contributor: TIGR
Download chromosome sequence data from NCBI FTP site

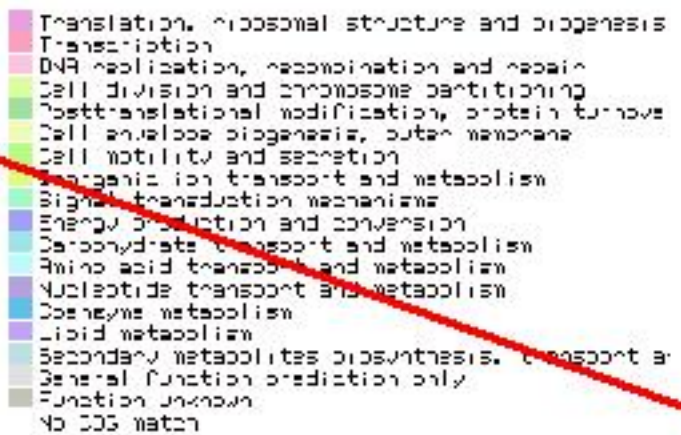
BLAST your query sequence against the genome

Start from : Go Search for gene Find

Protein coding genes distribution map
To see map locations of genes, click on a region in the map, to zoom in on that region

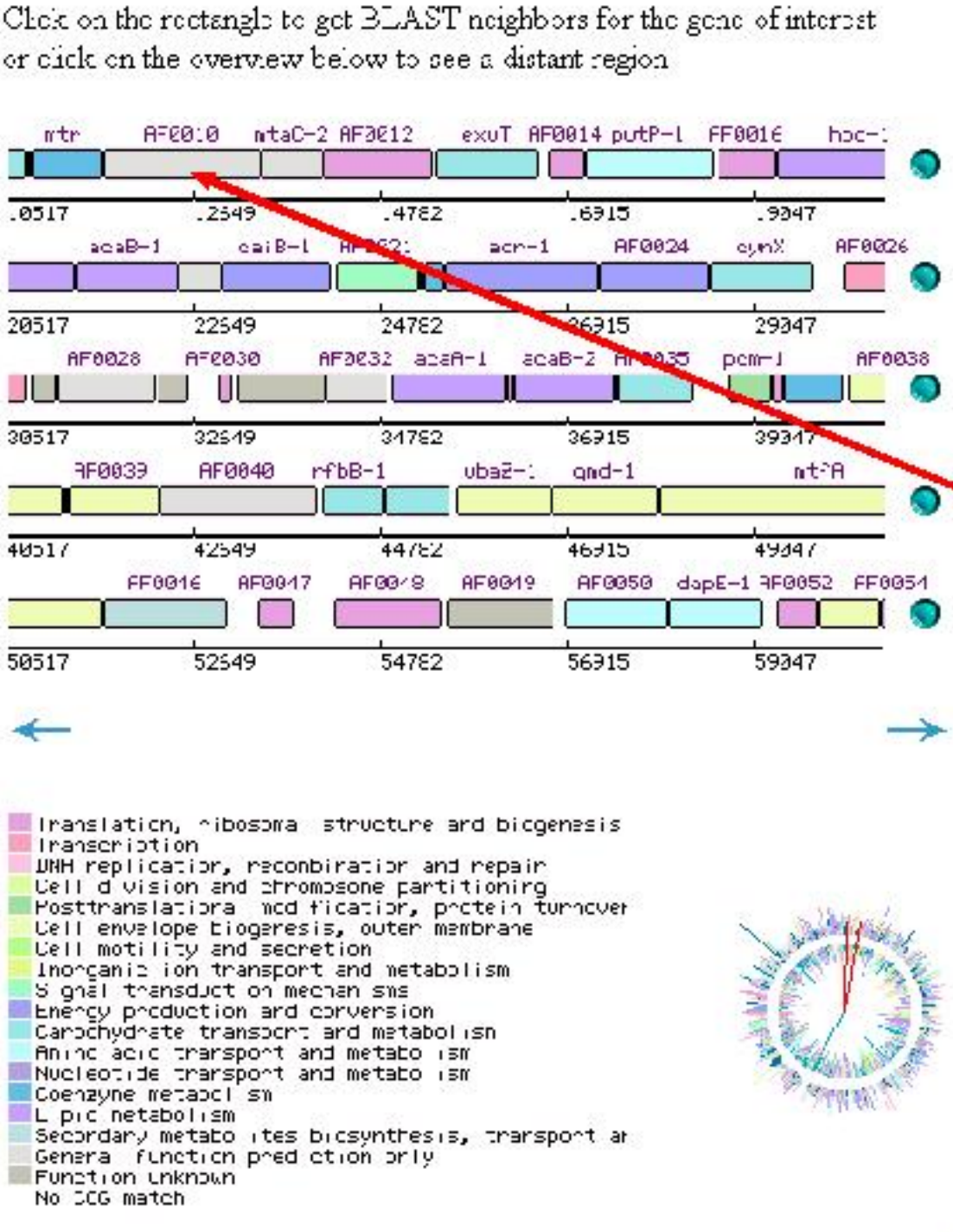


Gene Classification based on: COG functional categories



Рассмотрим графическое представление генов в геноме *A.fulgidus*

Рассмотрим район более подробно



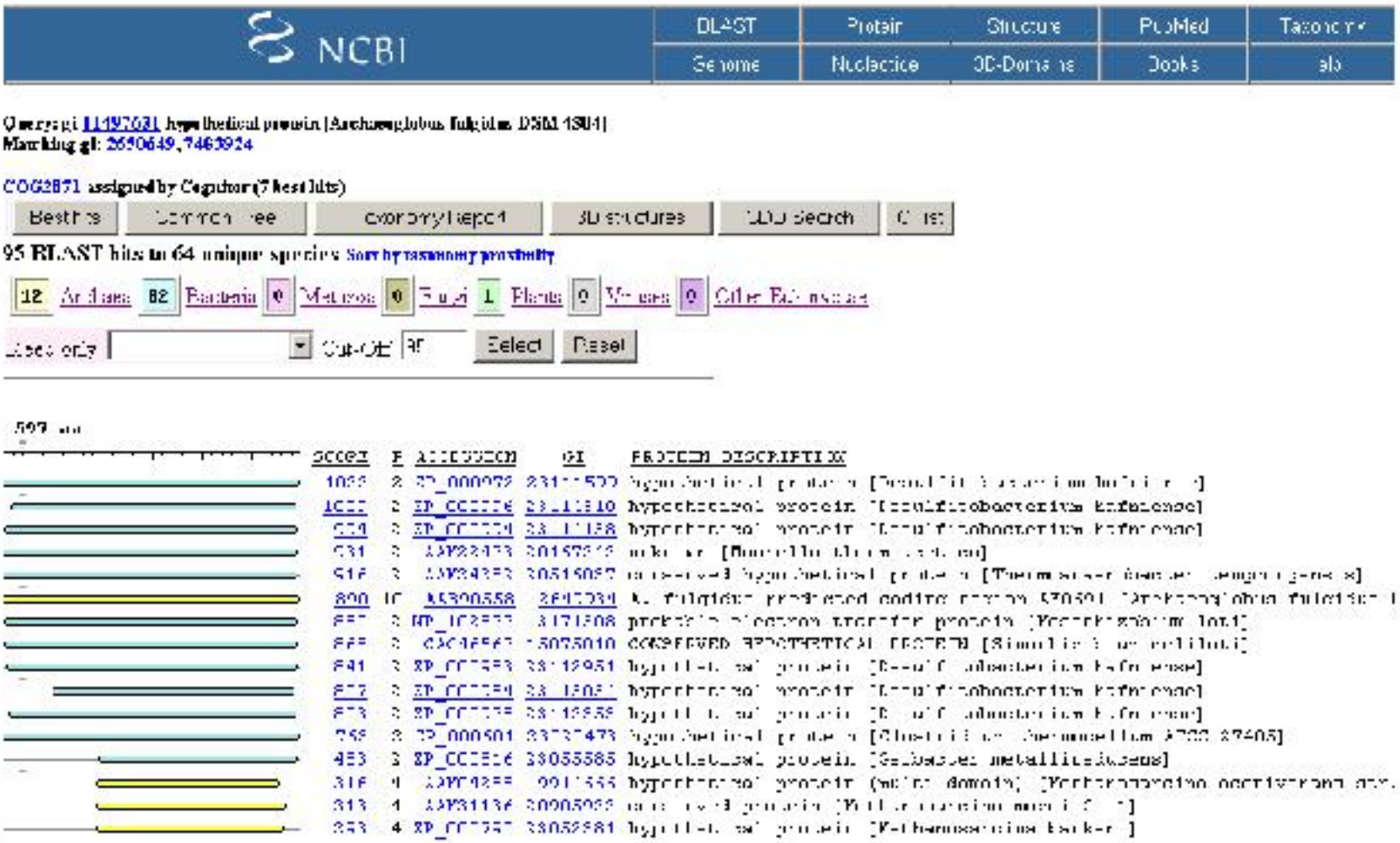
Указано расположение генов на последовательности

Рассмотрим ген AF0010 более подробно

Рассматриваемый район на карте генома

Поиск гомологии интегрирован в банк данных GenBank NCBI

Результаты поиска гомологичных генов для AF0010, *A.fulgidus*



Геномы эукариот в GenBank

Eukaryotic Genomes Taxonomy | List

Complete genome

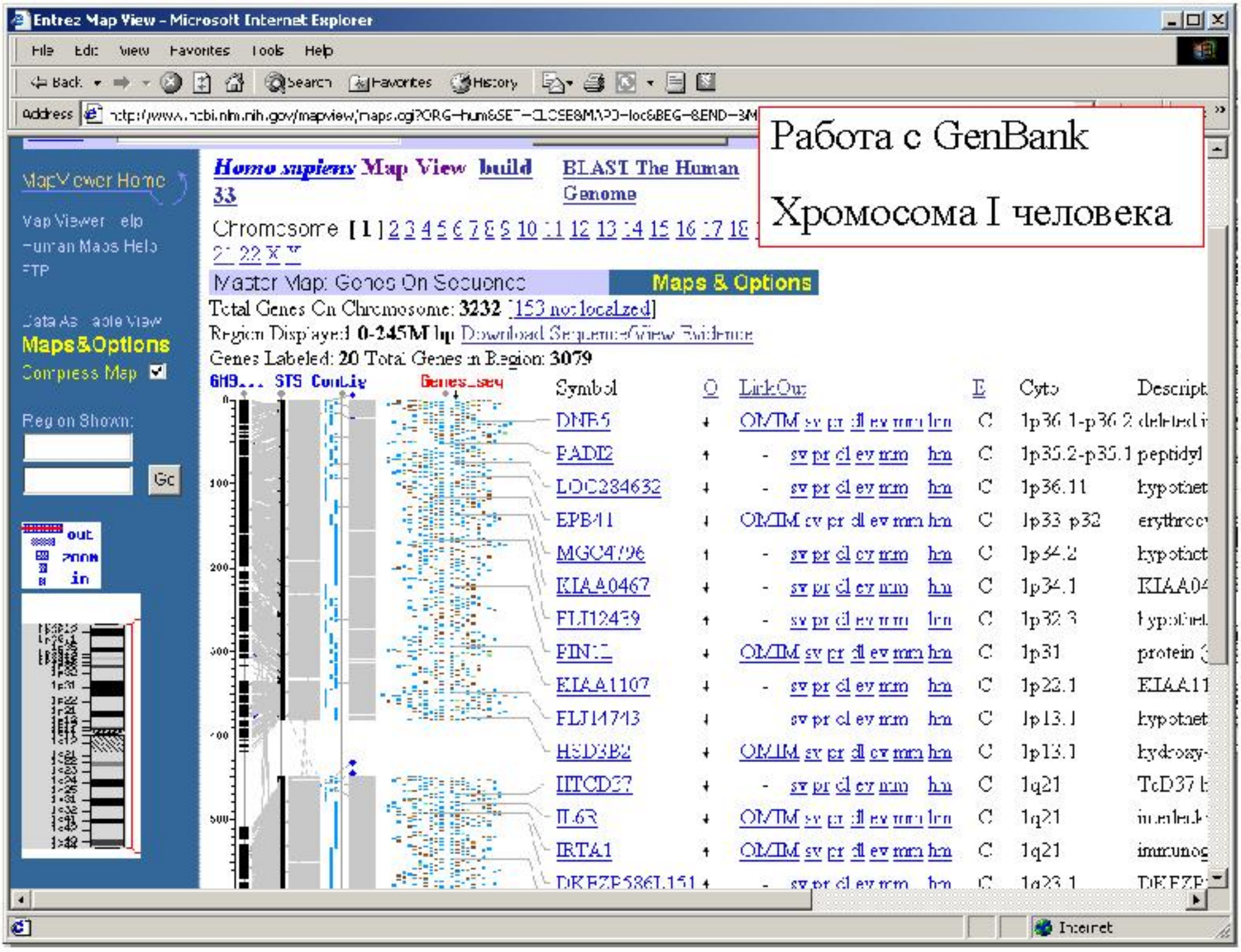
- [5] *Anopheles gambiae*
  chromosomes X, 2, 3,
- [5] *Arabidopsis thaliana*
  chromosomes I, II, III, IV, V,
- [6] *Caenorhabditis elegans*
  chromosomes I, II, III, IV, V, X,
- [5] *Drosophila melanogaster*
  chromosomes 1, 2, 3, 4, Y,
- [11] *Encephalitozoon cuniculi genome*
  chromosomes I, II, III, IV, V, VI, VII, VIII, IX, X, XI,
- [3] *Guillardia theta nucleomorph genome*
  chromosomes 1, 2, 3,
- [16] *Saccharomyces cerevisiae*
  chromosomes I, II, III, IV, V, VI, VII, VIII, IX, X, XI, XII, XIII, XIV, XV, XVI,
- [14] *Plasmodium falciparum*
  chromosomes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,
- [3] *Schizosaccharomyces pombe*
  chromosomes I, II, III,

Maps -> *See genomes in Map Viewer*

Vertebrates

- [24] *Homo sapiens*
  chromosomes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, Y.
- [21] *Mus musculus*
  chromosomes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, X, Y.
- [12] *Rattus norvegicus (rat)*
  chromosomes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,

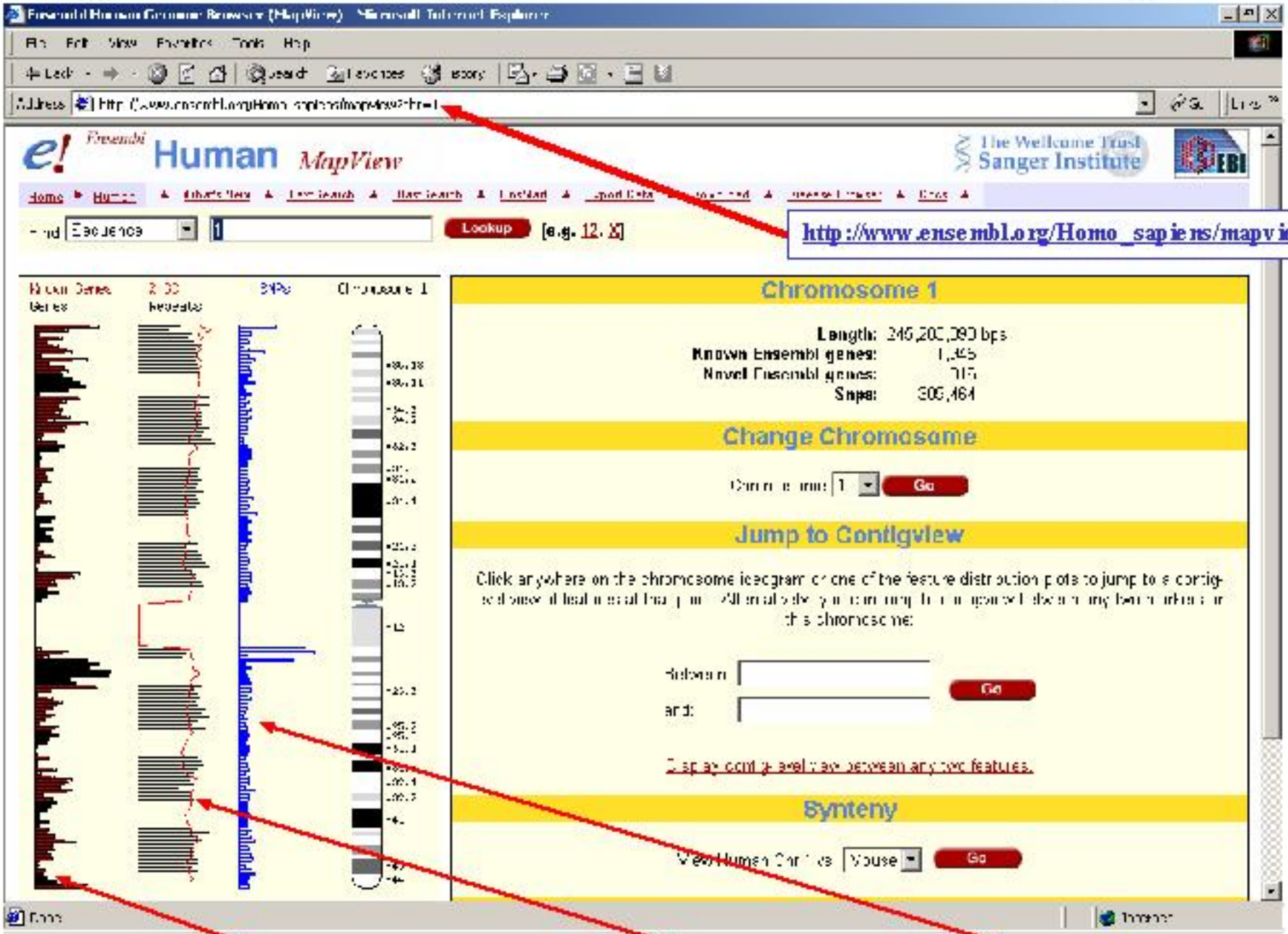
Интегрированная система представления данных


Entrez Map View - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address
Работа с GenBank
Хромосома I человека
MapViewer Home
Maps&Options
Compress Map
Region Shown:
Go
Homo sapiens Map View build 33
BLAST The Human Genome
Chromosome [1] 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
21 22 X Y
Master Map: Genes On Sequence
Maps & Options
Total Genes On Chromosome: 3232 [153 not localized]
Region Displayed: 0-245M bp Download Sequence/View Evidence
Genes Labeled: 20 Total Genes in Region: 3079
Symbol
LinkOut
Cyto
Descript
DNE5
EADI2
LOC284632
EPB41
MGC4796
KIAA0467
FLJ12439
PIN1
KIAA1107
FLJ14713
HSD3B2
IITCD37
IL6R
IRTA1
Internet

Европейский институт биоинформатики (EBI) http://www.ebi.ac.uk

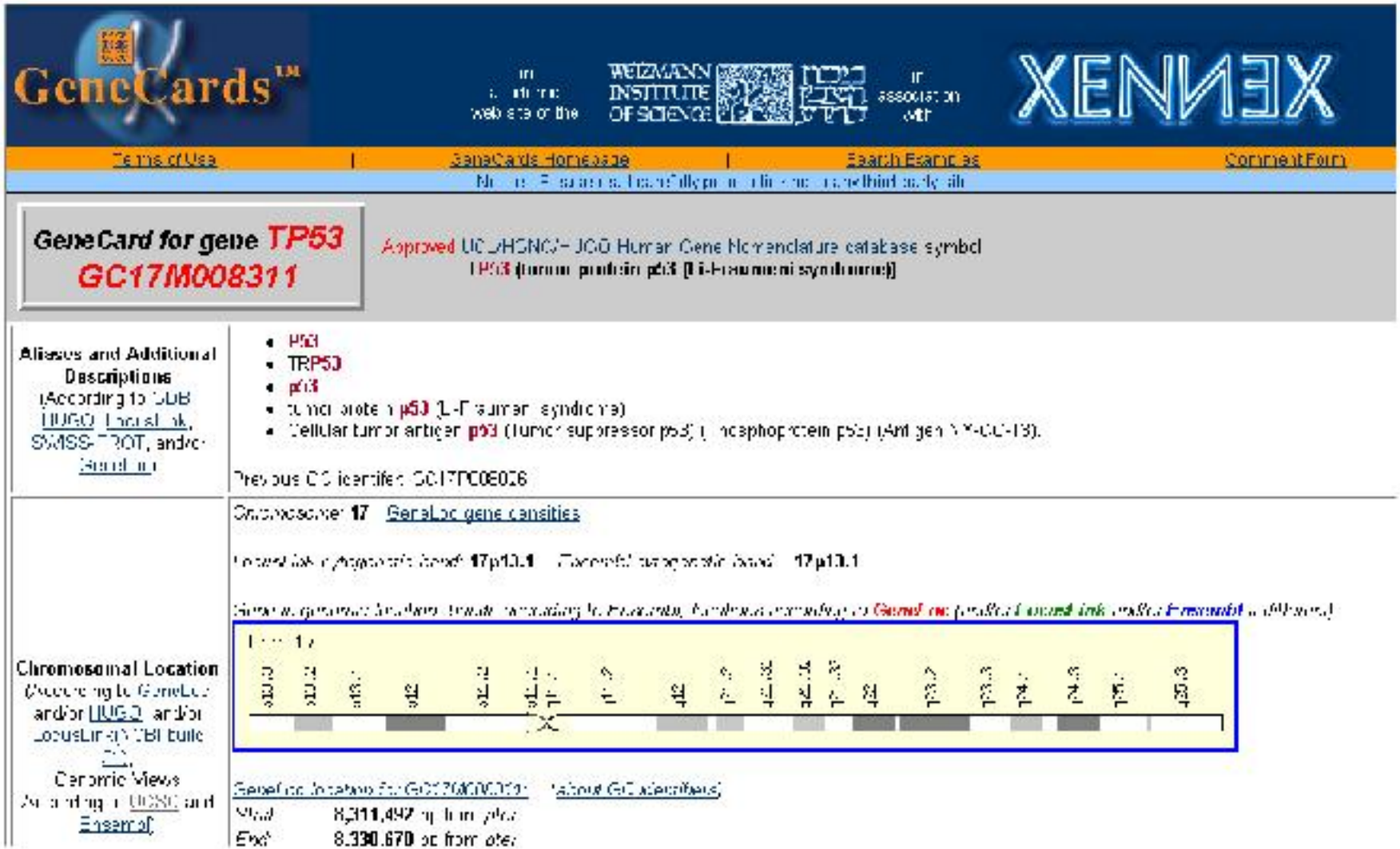
БД EMBL http://www.ebi.ac.uk/embl/

Представление хромосомы I человека на сервере EBI



Информация о плотности генов на хромосоме, GC составе и нуклеотидном полиморфизме

База данных GeneCards
(Вейцмановский институт, Израиль)



Основная идея базы GeneCards: анализ и биомедицинских знаний в Интернете и представление их для пользователя.

База данных GeneCard:

Пример карточки гена p53

Homologues:

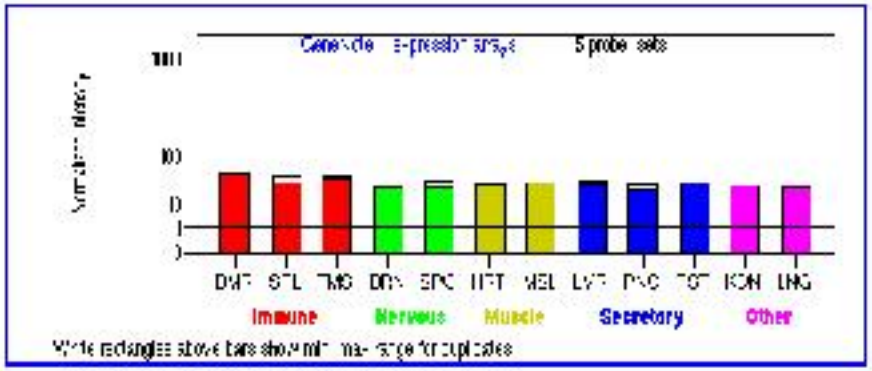
	gene	locus	description	% similarity to human	NCBI accessions
pig (HomoloGene)	[illegible]		[illegible] tumor suppressor p53 (p53), mRNA, complete cds	94.06	[illegible]
cow (HomoloGene)	TP53		p53 tumor suppressor phosphoprotein	93.43	[illegible]
rat (HomoloGene)	Trp53		tumor protein p53 ([illegible] transforming protein)	92.66	[illegible]
mouse (MGD)	Trp53	11 [illegible] cM	transformation related protein 53	--	[illegible]
mouse (HomoloGene)	Trp53		transformation related protein 53	82.66	[illegible]
zebrafish (HomoloGene)				74.13	[illegible]
C. elegans (Sanger)	B0373.7B			38.43	

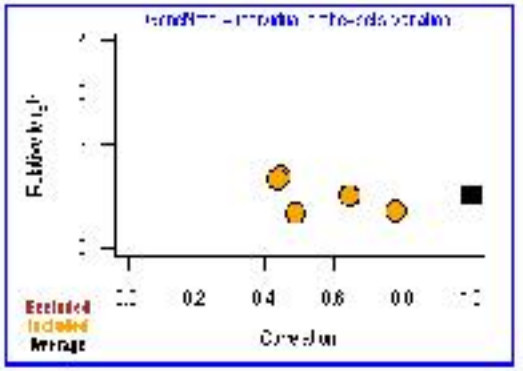
Similar Genes in Other Organisms (According to NCBI Ortholog Data) HomoloGene, Sanger Trunk Collection, Ensembl Comparative Genomics, Database and/or publication. It is possible. Further links to Ensembl and/or Mouse Base.

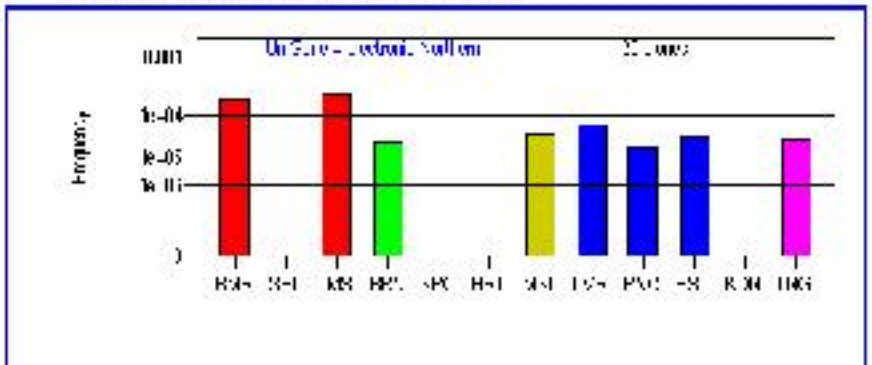
Affymetrix probe-set	array			
[illegible]	A			
[illegible]	A			
78468_i	D	0.978	0.385	1
31516_at	A	0.642	0.381	2
32151_at	D	0.312	0.153	2

No similarity/ortholog data found for p53 in HomoloGene for organisms (worm, tomato, bee, pig, wheat, coral, fish, cress, medaka, fruit

Expression in Human Tissues (Experimental results according to SeraMole Annotation connecting genes to Affymetrix probe sets according to SeraMap. Electronic Northern calculations according to data from UniGene (Build 150 Homo sapiens), plus additions. Links to SOURCE and/or SAGEmap)







Tissue		Clones per gene	Total clones
BMR	Bone marrow	4	19,806
SPL	Spleen	0	9,440
THM	Thymus	1	3,941
BRN	Brain	4	180,338
SPC	Spinal cord	0	472
HRT	Heart	0	3,287
MSL	Skeletal muscle	1	25,046
LVR	Liver	6	78,110
PNC	Pancreas	1	55,159
PST	Prostate	2	58,485
KDN	Kidney	0	67,819
LNG	Lung	2	72,824

Детальная информация о генах, включая данные по SNP, microarray, наследственным заболеваниям

«Компьютерная геномика» НГУ, Лекция 1, 2003

База данных GeneCard:

Пример карточки гена p53

Информация по мутациям и наследственным заболеваниям

OMIM ID: 191170

search databases for MIM named disorders.

- Colorectal cancer
- Li-Fraumeni syndrome

SWISS-PROT: P53_HUMAN

- **Disease:** TP53 is found in increased amounts in a wide variety of transformed cells. TP53 is frequently mutated or inactivated in about 50% of cancers.
- **Disease:** defects in tp53 are **also the cause of** germline cancers **such as li-fraumeni syndrome** (lfs) [mim:151623]. LFS is an autosomal dominant familial cancer syndrome that in its classic form is defined by the existence of both a proband with a sarcoma and two other first-degree relatives with a cancer by age 45 years. In these families the affected relatives develop a diverse set of malignancies including breast carcinomas, sarcomas and brain tumors at unusually early ages.
- **Disease:** Variant Ala-143 is temperature sensitive. At 32.5 degrees celsius it possesses strong DNA binding ability, but at 37.5 degrees celsius its transcriptional activities are greatly reduced.
- **Disease:** Defects in TP53 are also the cause of Barrett's adenocarcinomas (BA). BA is a condition in which the normally stratified squamous epithelium of the lower esophagus is replaced by a metaplastic columnar epithelium. The condition develops as a complication in approximately 10% of patients with chronic gastroesophageal reflux disease and predisposes to the development of esophageal adenocarcinoma.
- **Disease.** Defects in TP53 are the cause of head and neck squamous cell carcinomas (HNSCC) and oral squamous cell carcinomas (OSCC). Cigarette smoke is a primary mutagenic agent in cancer of the upper digestive tract.

Disorders & Mutations (in which this Gene is involved, According to OMIM, SWISS-PROT, GeneAtlas, GeneTests, HGMD, BCGD and/or TIGDE)

GeneTests: TP53

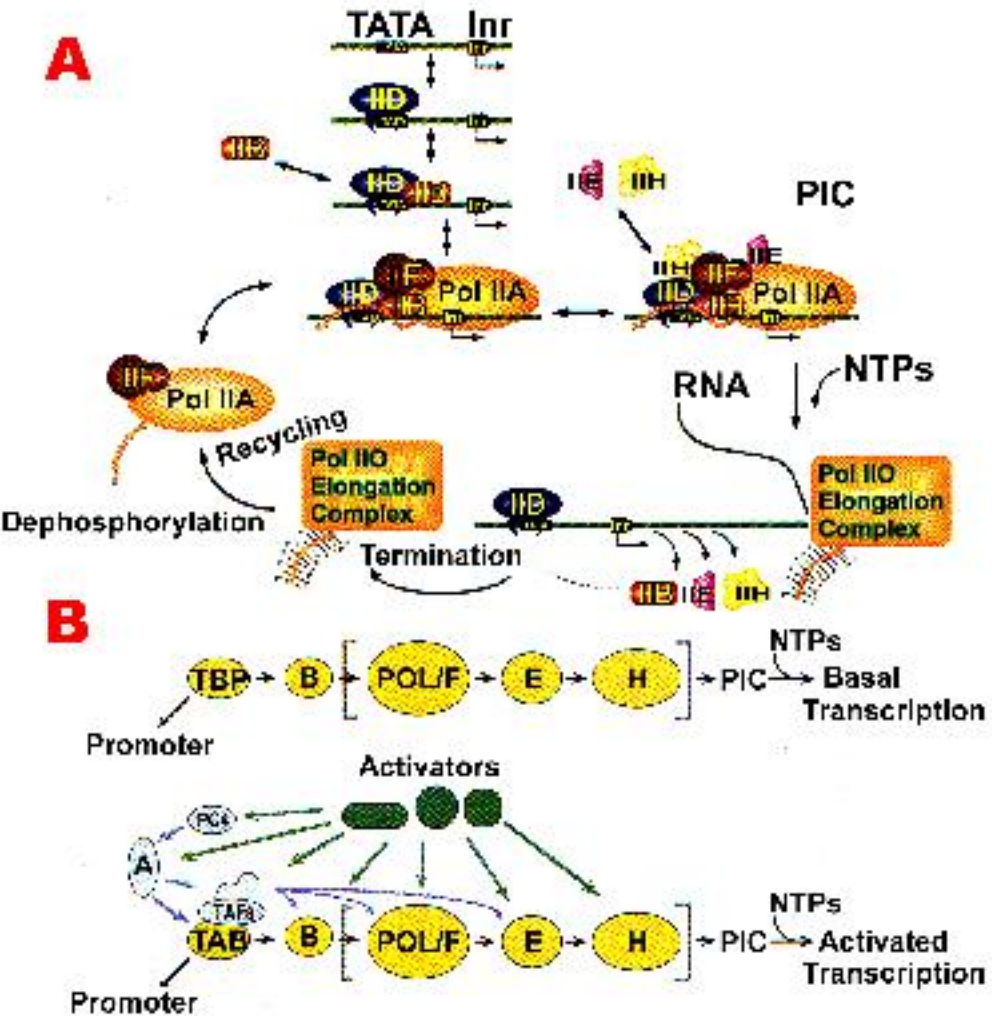
- Li-Fraumeni Syndrome

Human Gene Mutation Database entry for TP53

Breast Cancer Gene Database entry for TP53

Задачи исследования регуляции генной экспрессии

Следует отметить новые подходы к анализу текстов, связанные с моделированием генных сетей - комплексов взаимодействующих макромолекул, включающих гены и их продукты - РНК, белки и метаболиты. Это новое направление биоинформатики оперирует в качественно иных терминах, связанных с биофизическим и биохимическим моделированием и не затрагивает напрямую анализ самих последовательностей.



Эти задачи будут освещены в отдельном курсе

Регуляторные геномные последовательности (РГП)

Список рекомендуемой литературы

Математические методы для анализа последовательностей ДНК. (Под ред. М.С.Уотермена, Перевод с англ. под ред. П.А.Певзнера), Москва, «Мир», 1999.

Франк-Каменецкий М.Д. под ред. (1990) Компьютерный анализ генетических текстов Москва, Наука, 1990, 267 с.

Кель А.Э., Колчанов Н.А., Соловьев В.В. Математическое моделирование в молекулярной биологии и генетике. Теория мутационного процесса: делеции и дупликации. Новосибирск, НГУ, учебное пособие, 1989, 86 с.

David W. Mount «Bioinformatics. Sequence and genome analysis» CSHL Press, New York, 2001.

BLAST By Joseph Bedell, Ian Korf, Mark Yandell. Publisher : O'Reilly, July 2003 (ISBN : 0-596-00299-8) 360 pp.

Durbin R., Eddy S.R., Krogh A., Mitchson G. Biological sequence analysis. 1998, Cambridge: Cambridge University Press, 356 p.

Sequence-Evolution-Function. Computational Approaches in comparative Genomics. (Eds Eugene V. Koonin, Michael Y. Galperin) Kluwer, 2002, 488 pp.

Ресурсы:

ДНК

NCBI (Natl Center Biotech Information) - GenBank http://www.ncbi.nlm.nih.gov/
EBI (European Bioinformatics Institute) - EMBL http://www.ebi.ac.uk/
ENTREZ http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
PUBMED http://www.ncbi.nlm.nih.gov/pubmed
DNA Search http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide
GenBank http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html

Выравнивание

BLAST (Basic Local Alignment Sequence Tool) http://www.ncbi.nlm.nih.gov/blast
BLAST 2 SEQUENCES http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html

Белки

Expasy (Expert Protein Analysis System, SwissProt, TrEMBL) http://www.expasy.ch/
PDB - Protein 3D Structure database http://www.rcsb.org/pdb/

Номенклатура IUPAC

http://www.chem.qmul.ac.uk/iupac/misc/naabb.html

Рекомендуемый ресурс на русском языке