Л е к ц и я   № 1

Биоинформатика
(введение в дисциплину)

Разработал профессор П. М. Васильев
Кафедра фармакологии и биоинформатики

Для студентов, обучающихся по направлению 06.03.01 «Биология»
профили Биохимия, Генетика
при изучении дисциплины «Биоинформатика»

План лекции

- **Введение.**

- **Предмет, цели и задачи биоинформатики.**

- **Поиск научной информации.**

- **База данных PubMed.**

- **Биоинформатика последовательностей.**

- **Поиск последовательностей.**

- **База знаний по белкам UniProtKB.**

- **Банк данных по нуклеотидным последовательностям GenBank.**

Николай Павлович Кравков
русский фармаколог, академик Императорской Военно-медицинской академии (1914), член-корреспондент Российской академии наук (1920)
Основоположник советской фармакологии

Василий Васильевич Закусов
фармаколог, академик АМН СССР
Основатель и первый директор
Института фармакологии и химиотерапии
(1952 – 1979)

Артур Викторович Вальдман
фармаколог, академик АМН СССР
Директор НИИ фармакологии (1979 – 1990)

Геннадий Васильевич Ковалев
фармаколог, доктор медицинских наук,
профессор
Заведующий кафедрой фармакологии
ВГМИ (1970 – 1990)

Владимир Иванович Петров
фармаколог, академик РАН
Президент ВолгГМУ, директор НЦИЛС, заведующий кафедрой клинической фармакологии и интенсивной терапии

Александр Алексеевич Спасов
фармаколог, академик РАН
Советник ректора по науке ВолгГМУ, заведующий кафедрой фармакологии и биоинформатики

Лев Арамович Пирузян
биофизик, академик РАН
Основатель и директор НИИ по биологическим испытаниям химических соединений (1972 – 1984)
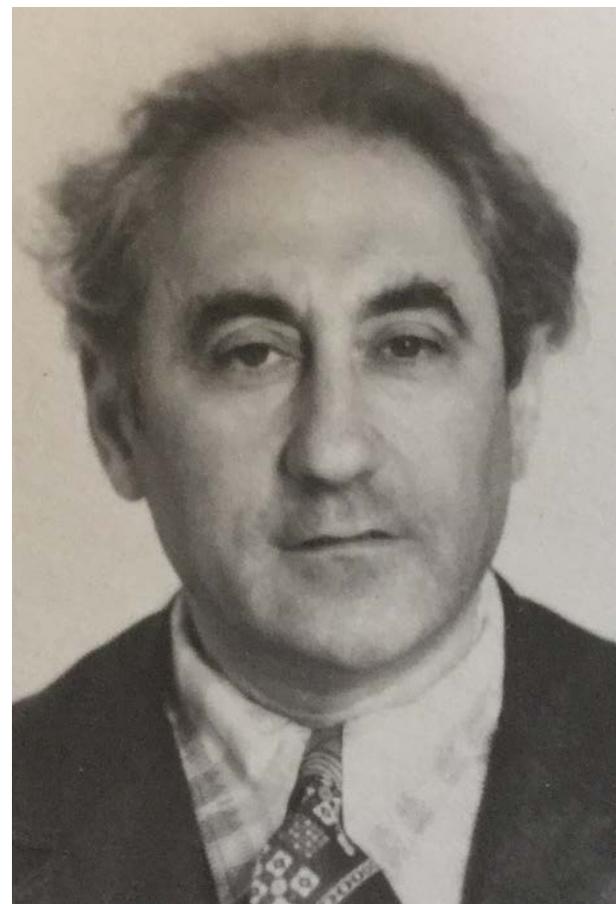
Ландау Михаил Александрович
химик, доктор химических наук, профессор
Заведующий лабораторией структурной химии НИИ по БИХС (1972 – 1984)

Что такое биоинформатика

Область науки, в которой решаются биологические задачи с помощью вычислительных методов математики и информационных технологий

Этапы развития биоинформатики

Год	Технология	Биоинформатика
1965	Секвенирование tRNA	База данных PIR
1970	Обратная транскрипция	Алгоритм выравнивания NW
1972	Клонирование	
1977	Секвенирование	База данных PDB
1980		База данных NAR
1981		Алгоритм выравнивания SW
1982	Секвенирование ДНК фага лямбда	
1983	PCR	Алгоритм поиска по БД WL
1985	Секвенирование ДНК вирусов	FASTA-поиск по базе данных
1987		GeneBank
1989	Программа "Геном человека"	Swiss-Prot, NCBI
1992	Первая хромосома дрожжей	BLOSSUM
1993	Автоматическое секвенирование	
1995	Первый геном бактерии	База данных SCOP
1996	Первый геном архейный	
1997		PSI-BLAST
1998	Геном червя	
2001	Геном человека	

Рост объема информации
по нуклеотидным последовательностям

Sequences
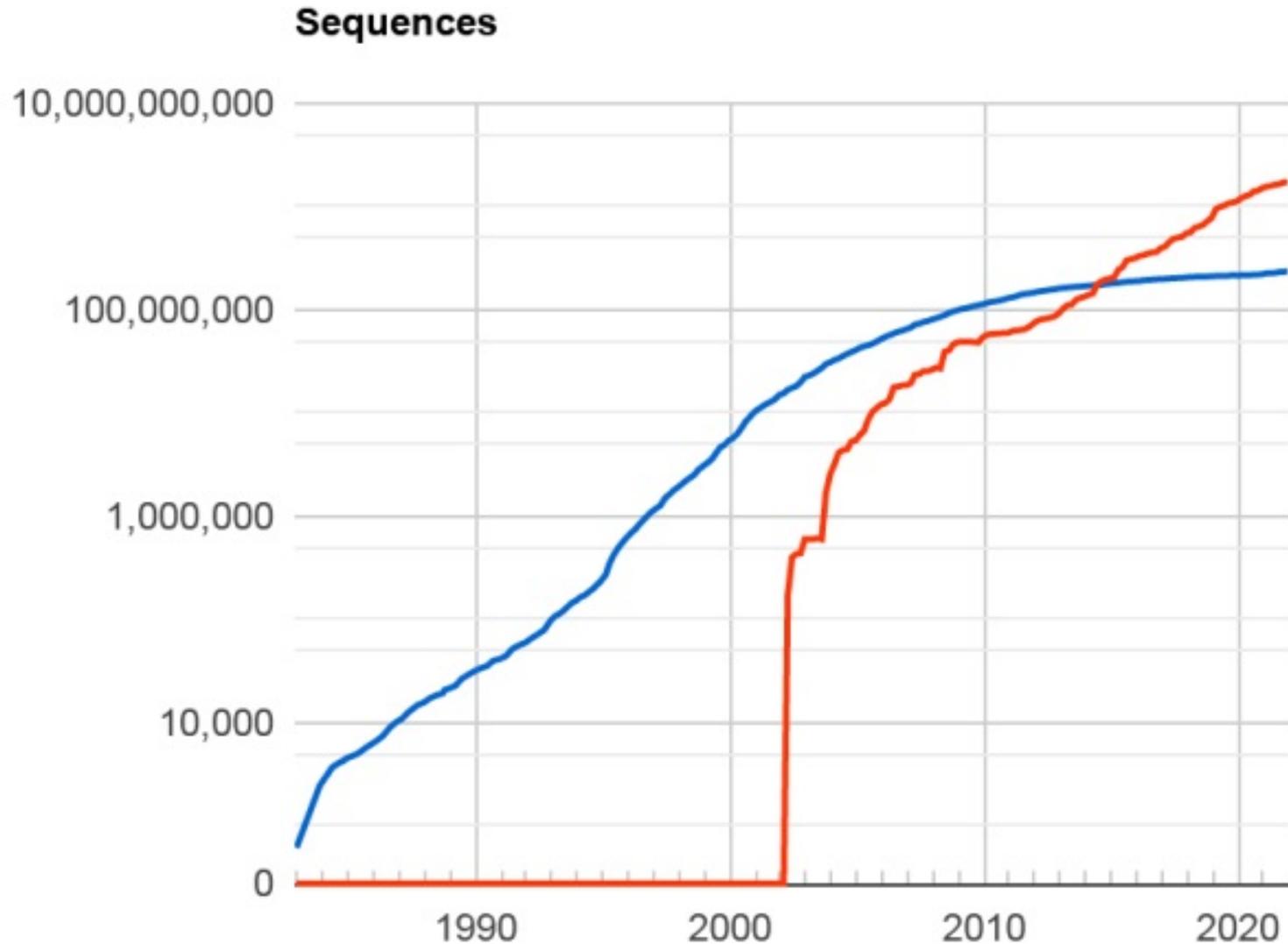
Разделы биоинформатики
(по объектам исследования)

Биоинформатика последовательностей

Структурная биоинформатика

Компьютерная геномика

Разделы биоинформатики
(по способам исследования)

Применение известных методов анализа
для получения новых биологических знаний

Разработка новых методов анализа
биологических данных

Разработка новых баз данных

Задачи биоинформатики

- Разработка алгоритмов для анализа биологических данных большого объема
  - Алгоритм поиска генов в геноме

- Анализ и интерпретация различных типов биологических данных таких, как нуклеотидные и аминокислотные последовательности, домены белков, структура белков и т.д.
  - Изучение структуры активного центра белка

- Разработка программного обеспечения для управления и быстрого доступа к биологическим данным
  - Создание банка данных аминокислотных последовательностей

Биоинформатика последовательностей

- Выравнивание и определение сходства двух последовательностей
- Построение множественных выравниваний
- Распознавание генов
- Предсказание вторичной структуры белков и РНК
- Предсказание сайтов связывания белков

Структурная биоинформатика

- Предсказание функциональных участков белковой молекулы

- Сравнительный анализ структур родственных белков

- Классификация белков на основе их 3D-структуры

- Анализ структур комплексов двух или нескольких молекул белка

- Анализ комплексов молекул белка с другими молекулами

- Предсказание воздействия молекул химических веществ (например, потенциальных лекарств) на молекулы белков

- Предсказание структуры белка по структуре белка с похожей последовательностью

Компьютерная геномика

- Предсказание генов в последовательностях
- Предварительная аннотация генов по сходству белковых последовательностей
- Поиск «пропущенных» генов
- Сравнительный анализ геномов
- Исследование регуляции работы генов (экспрессия, коэкспрессия)
- Полногеномный анализ ассоциаций

ПОИСК И БАЗЫ ДАННЫХ

PubMed – поиск биомедицинской информации

https://www.ncbi.nlm.nih.gov/pubmed/

GenBank – нуклеотидные последовательности

https://www.ncbi.nlm.nih.gov/genbank/

UniProt – информация о белках

https://www.uniprot.org/

PDB – 3D-структуры белков

https://www.rcsb.org/

DrugBank – лекарства, действующие на биомишени

https://www.drugbank.ca/

Объем баз данных 2022

PubMed – **34 млн.** статей

GenBank – **1 трл. 492,8 млрд.** оснований

239,9 млн. последовательностей

UniProt – **227,3 млн.** белков

PDB – **194,8 тыс.** 3D-структур биомолекул

DrugBank – **>500 тыс.** лекарств и продуктов

29,1 тыс. белков-мишеней

Типы баз данных

- **Архивные базы данных**

  Эти базы данных не проверяются

    GeneBank, EMBL – первичные последовательности

    PDB – пространственные структуры белков

- **Курируемые базы данных**

  Информацию отбирают эксперты

    Swiss-Prot – аминокислотные последовательности

    белков

    KEGG – функции генов и сигнальные пути

Типы баз данных

- **Производные базы данных**
  Результат обработки данных
  - SCOP – структурная классификации белков
  - PFAM – семейства белков
  - GO – классификация генов (Gene Ontology)
  - ProDom – белковые домены

- **Интегрированные базы данных**
  Результат объединения всех данных
  - NCBI Entrez – нуклеотидные и аминокислотные последовательности и структуры
  - Ecocyc – гены, белки, метаболизм E. coli

Поиск литературы

https://www.ncbi.nlm.nih.gov/pubmed/

PubMed comprises more than **34 million** citations for biomedical literature from MEDLINE, life science journals, and online books.

Citations may include links to full-text content from PubMed Central and publisher web sites.

PubMed

Начальная страница поиска

Простой поиск в PubMed

drug discovery — 141 566 статей

Точный поиск в PubMed

"drug discovery" — 87 012 статей

Поисковые запросы

AND, & – логическое "И"

OR – логическое "ИЛИ"

diabetes – 856 238 статей

alzheimer – 193 397 статей

antioxidant – 657 255 статей

diabetes OR alzheimer – 1 041 368 статей

diabetes AND alzheimer – 8 267 статей

diabetes alzheimer antioxidant – 763 статьи

Расширенный поиск в PubMed

Advanced

Расширенный поиск в PubMed

Title/Abstract

Расширенный поиск в PubMed

376 статей

Полные тексты статей

164 статьи

Полные тексты обзоров

64 статьи

Обзоры за 5 последних лет

42 статьи

Реферат обзора

Можно получить полный текст

Полный текст обзора

Можно получить полный текст в pdf

Поиск аминокислотных последовательностей

UniProt

https://www.uniprot.org/

227,3 млн. белков

Формат FASTA

W. R. Pearson, D. J. Lipman − 1988

<u>Строка-заголовок (the definition line)</u>

> **[уникальный ID] | [описание]**

Последовательность белка (или ДНК) в однобуквенном коде

```
>sp|P35557|HXK4_HUMAN Hexokinase-4 OS=Homo sapiens
MLDDRARMEAAKKEKVEQILAEFQLQEEDLKKVMRRMQKEMDRGLRLETH
EEASVKMLPTYVRSTPEGSEVGDFLSLDLGGTNFRVMLVKVGEGEEGQWS
VKTKHQMYSIPEDAMTGTAEMLFDYISECISDFLDKHQMKHKKLPLGFTF
SFPVRHEDIDKGILLNWTKGFKASGAEGNNVVGLLRDAIKRRGDFEMDVV
AMVNDTVATMISCYYEDHQCEVGMIVGTGCNACYMEEMQNVELVEGDEGR
MCVNTEWGAFGDSGELDEFLLEYDRLVDESSANPGQQLYEKLIGGKYMGE
LVRLVLLRLVDENLLFHGEASEQLRTRGAFETRFVSQVESDTGDRKQIYN
ILSTLGLRPSTTDCDIVRRACESVSTRAAHMCSAGLAGVINRMRESRSED
VMRITVGVDGSVYKLHPSFKERFHASVRRLTPSCEITFIESEEGSGRGAA
LVSAVACKKACMLGQ
```

Коды нуклеотидов

a	adenine
c	cytosine
g	guanine
t	thymine
u	uracil
r	**a** or **g**
y	**c** or **t**
s	**g** or **c**
w	**a** or **t**

k	**g** or **t**
m	**a** or **c**
b	**c** or **g** or **t**
d	**a** or **g** or **t**
h	**a** or **c** or **t**
v	**a** or **c** or **g**
n	any base
. or **–**	gap

Коды аминокислот

A	Ala	Alanine
B	Asx	Aspartic acid or Asparagine
C	Cys	Cysteine
D	Asp	Aspartic Acid
E	Glu	Glutamic Acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline

Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
X	Xaa	Any amino acid
Y	Tyr	Tyrosine
Z	Glx	Glutamine or Glutamic acid
U	Sec	SelenoCysteine
O	Pyl	Pyrrolysine
*		translation stop
-		gap of indeterminate length

UniProtKB

5-HT$_{2a}$ рецептор человека

5-HT$_{2a}$ рецептор человека

Последовательность 5-HT$_{2A}$ Human

Sequences (2+)[i]

Sequence status[i]: Complete.

This entry describes **2** isoforms[i] produced by **alternative splicing**. ≡ Align   🛒 Add to basket

This entry has 2 described isoforms and 2 potential isoforms that are computationally mapped. 👁 Show all   ≡ Align All

---

Isoform 1 (identifier: **P28223-1**) [UniParc] ⬇ FASTA   🛒 Add to basket

This isoform has been chosen as the canonical[i] sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

« Hide

Length: 471
Mass (Da): 52,603
Last modified: June 1, 1994 - v2
Checksum:[i] EF8AAC0BC5379DA2

BLAST ▾   GO

```
          10         20         30         40         50
   MDILCEENTS LSSTTNSLMQ LNDDTRLYSN DFNSGEANTS DAFNWTVDSE
          60         70         80         90        100
   NRTNLSCEGC LSPSCLSLLH LQEKNWSALL TAVVIILTIA GNILVIMAVS
         110        120        130        140        150
   LEKKLQNATN YFLMSLAIAD MLLGFLVMPV SMLTILYGYR WPLPSKLCAV
         160        170        180        190        200
   WIYLDVLFST ASIMHLCAIS LDRYVAIQNP IHHSRFNSRT KAFLKIIAVW
         210        220        230        240        250
   TISVGISMPI PVFGLQDDSK VFKEGSCLLA DDNFVLIGSF VSFFIPLTIM
         260        270        280        290        300
   VITYFLTIKS LQKEATLCVS DLGTRAKLAS FSFLPQSSLS SEKLFQRSIH
         310        320        330        340        350
   REPGSYTGRR TMQSISNEQK ACKVLGIVFF LFVVMWCPFF ITNIMAVICK
         360        370        380        390        400
   ESCNEDVIGA LLNVFVWIGY LSSAVNPLVY TLFNKTYRSA FSRYIQCQYK
         410        420        430        440        450
   ENKKPLQLIL VNTIPALAYK SSQLQMGQKK NSKQDAKTTD NDCSMVALGK
         460        470
   QHSEEASKDN SDGVNEKVSC V
```

FASTA 5-HT$_{2A}$ Human

```
>sp|P28223|5HT2A_HUMAN 5-hydroxytryptamine receptor 2A OS=Homo sapiens GN=HTR2A PE=1 SV=2
MDILCEENTSLSSTTNSLMQLNDDTRLYSNDFNSGEANTSDAFNWTVDSENRTNLSCEGC
LSPSCLSLLHLQEKNWSALLTAVVIILTIAGNILVIMAVSLEKKLQNATNYFLMSLAIAD
MLLGFLVMPVSMLTILYGYRWPLPSKLCAVWIYLDVLFSTASIMHLCAISLDRYVAIQNP
IHHSRFNSRTKAFLKIIAVWTISVGISMPIPVFGLQDDSKVFKEGSCLLADDNFVLIGSF
VSFFIPLTIMVITYFLTIKSLQKEATLCVSDLGTRAKLASFSFLPQSSLSSEKLFQRSIH
REPGSYTGRRTMQSISNEQKACKVLGIVFFLFVVMWCPFFITNIMAVICKESCNEDVIGA
LLNVFVWIGYLSSAVNPLVYTLFNKTYRSAFSRYIQCQYKENKKPLQLILVNTIPALAYK
SSQLQMGQKKNSKQDAKTTDNDCSMVALGKQHSEEASKDNSDGVNEKVSCV
```

PDB формат

```
HEADER    ModPipe Model of UP B2RAC5              2010-09-2
TITLE     Model of  5-hydroxytryptamine receptor 2A isoform 1 [Homo sa
SOURCE
AUTHOR    URSULA PIEPER, EASHWAR NARAYANAN, ANDREJ SALI
REMARK 220 Original ID: ENSP0000036795
REMARK 220 EXPERIMENTAL DETAILS
REMARK 220 EXPERIMENT TYPE: THEORETICAL MODEL
REMARK 220 METHOD: HOMOLOGY MODELING
REMARK 220 PROGRAM: MODPIPE
REMARK 220 SEQUENCE IDENTITY:          18.00
REMARK 220 GA341 SCORE:                0.29
REMARK 220 EVALUE:                     0
REMARK 220 MPQS:                       0.671047
REMARK 220 zDOPE SCORE:                0.78
REMARK 220 TEMPLATE PDB:               2ziy
REMARK 220 TEMPLATE CHAIN:             A
REMARK 220 TARGET LENGTH:              471
REMARK 220 TARGET BEGIN:               75
REMARK 220 TARGET END:                 439
REMARK 220 TEMPLATE BEGIN:             35
REMARK 220 TEMPLATE END:               369
REMARK 220 MODPIPE RUN:                human_2010_I1
REMARK 220 MODPIPE MODEL ID:           5d6b98e9e00432ae919177b2b38d636c
REMARK 220 MODPIPE ALIGN ID:           ca14ce341fce7d514026f126f0842823
REMARK 220 MODPIPE SEQUENCE ID:        3d8573062bfc3a41e50f3ea5cd5f1570MDILVSCV
EXPDTA    THEORETICAL MODEL, MODELLER SVN 2010/07/22 23:00:26
REMARK   6 MODELLER OBJECTIVE FUNCTION:      3242.1162
REMARK   6 MODELLER BEST TEMPLATE % SEQ ID:  18.209
REMARK   6 GENERATED BY MODPIPE VERSION SVN.r1188:1193
ATOM      1  N   ASN    75      53.564  25.782  27.022  1.00 43.29           N
ATOM      2  CA  ASN    75      53.307  26.958  26.181  0.50 43.29           C
ATOM      3  CB  ASN    75      54.466  27.190  25.199  0.50 43.29           C
ATOM      4  CG  ASN    75      55.636  27.743  25.995  0.50 43.29           C
ATOM      5  OD1 ASN    75      56.654  27.076  26.170  1.00 43.29           O
ATOM      6  ND2 ASN    75      55.487  28.999  26.496  1.00 43.29           N
ATOM      7  C   ASN    75      52.053  26.820  25.395  1.00 43.29           C
ATOM      8  O   ASN    75      51.000  26.472  25.925  1.00 43.29           O
ATOM      9  N   TRP    76      52.159  27.075  24.082  1.00125.00           N
ATOM     10  CA  TRP    76      51.022  27.089  23.216  1.00125.00           C
```

3D-модель 5-HT$_{2a}$-рецептора человека

Поиск нуклеотидных последовательностей
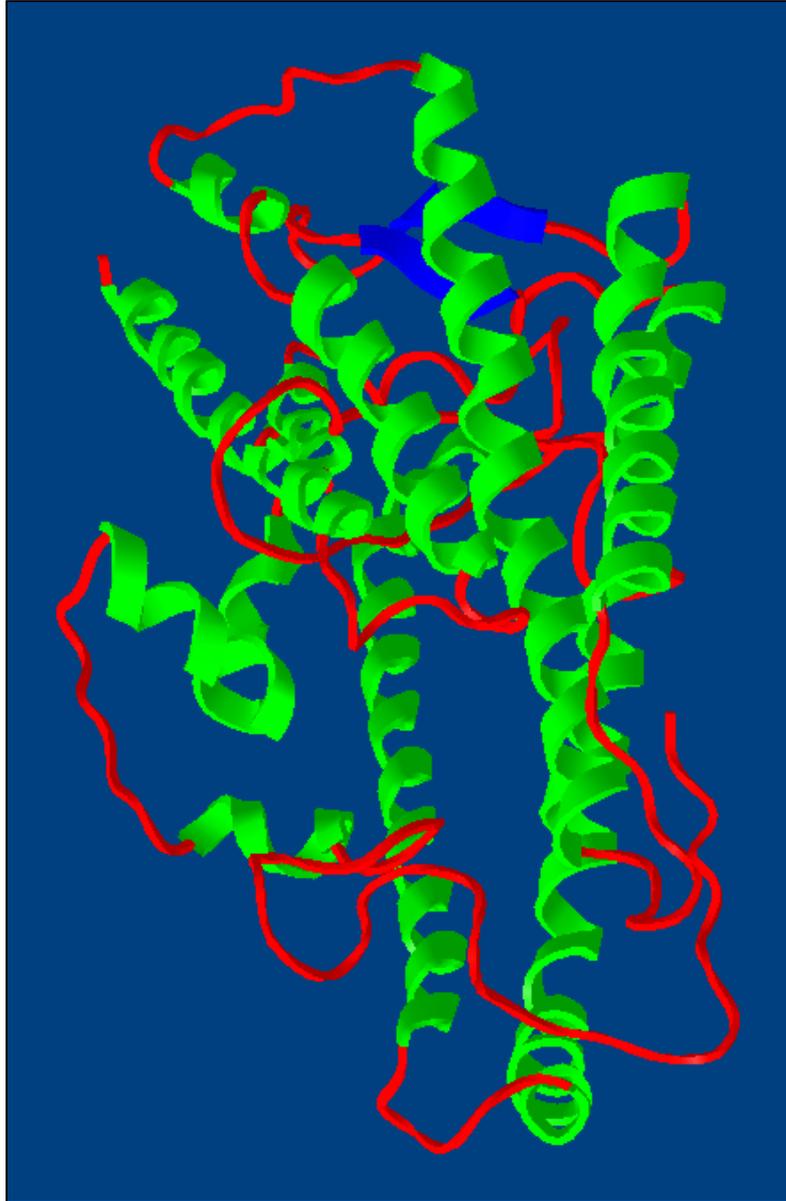
GenBank

https://www.ncbi.nlm.nih.gov/genbank/

234,6 млн. последовательностей

GenBank

245 766 полных геномов человека

GenBank

Genomes → Complete Genome

GenBank

GenBank — WGS Browser

Human whole genome

GenBank — Project: AACC02

Поиск нуклеотидных последовательностей

1000 Genomes

IGSR: The International Genome Sample Resource

https://www.internationalgenome.org/

5019 полных геномов людей разных национальностей

1000 Genomes

1000 Genomes — Map

1000 Genomes — Data Collection

1000 Genomes — Samples

1000 Genomes — HG00152

1000 Genomes — Exome

To be continued ...