

Банки информации в молекулярной биологии

С.А.Спирин

11/III – 2006

Пример информации

последовательность ДНК

gatcaacactacttgacttcaagacttaccataaagaaaactatagtgtggtattggcaa
aagacaagacaaatagatcaacataacaaaataaagggccatgaaatagaccatatagt
caattgatttttgacaaagaaggattggcaatagaatgggggtaaagatagtcttctcaac
aaacggtaccagaatgactgaatacccacatgcaaaaagaaaaagaaatgaacctagaca
cagatcttatacagttcacaaaaatgtaactcaaatgaatcatagacctaaatataata
ttcaagactataaaaccctaaaatataacataggggaaaatctaaacaatcttgagtttg
ttaatgacttttttagatacaataccaaaggcaggatccaggaaagaatcgataagctggg
cttcattaataaataaataatcttgctctatgaagccactgtcaagagaaggaaaaggca
agccatagactgggagaaaatatttacaanaagacatacatgataaaggactattatccaa
aatgtacaagaactctaaaaacttaacaataagaaaacaaacccaactaaaaactggg
ccaaagatcttaacagatatattaccaagaagatacacagatggcaaataagcataaaa
agattaaccacatcatcgtcattaagaaattgcaattaaaacaacaatgagacaccat
tatacacctagtagaatgacccaaatccagattactgacataatcaaatgctgacaagga
tgtggagaaacaggaactgccattcttggggtgtgggaatgccaaatgggtatgcctgctt
tggaagacagcttggtggtttcttacaacactaagcatactcttaccaaaagatcgagca

Вообще-то ДНК — это молекула...

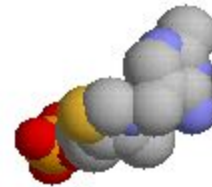
Примеры молекул:



Сульфат

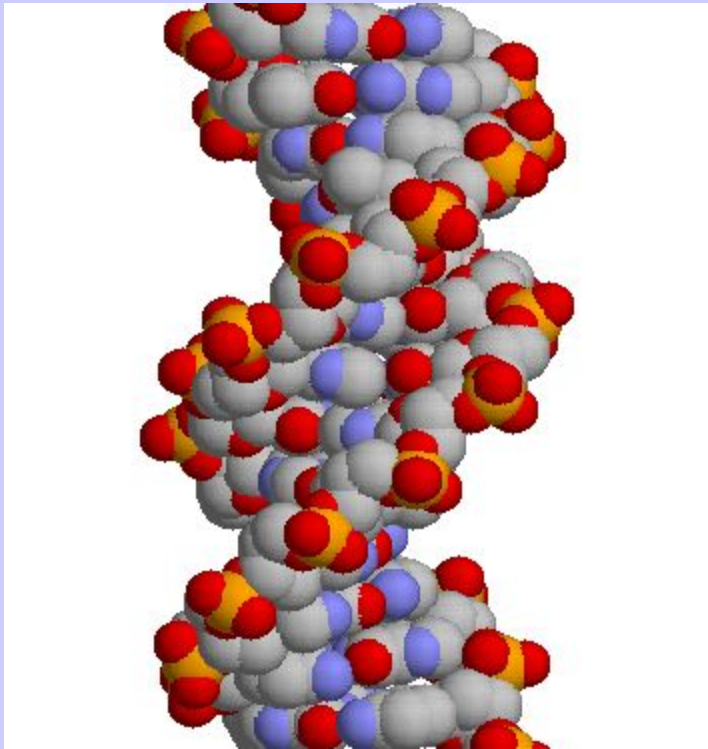


Фенол

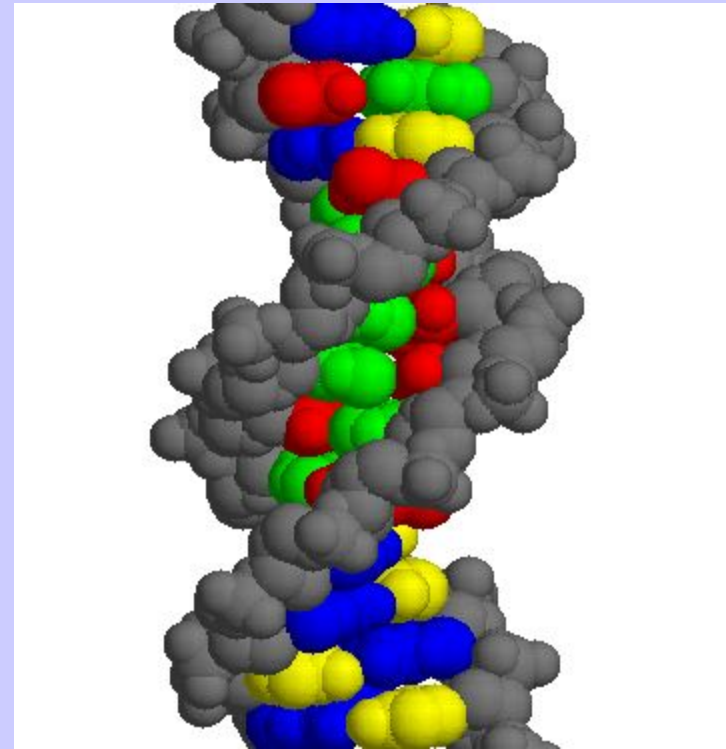


Фосфат тиаминa
(атомы водорода не показаны)

Молекула ДНК

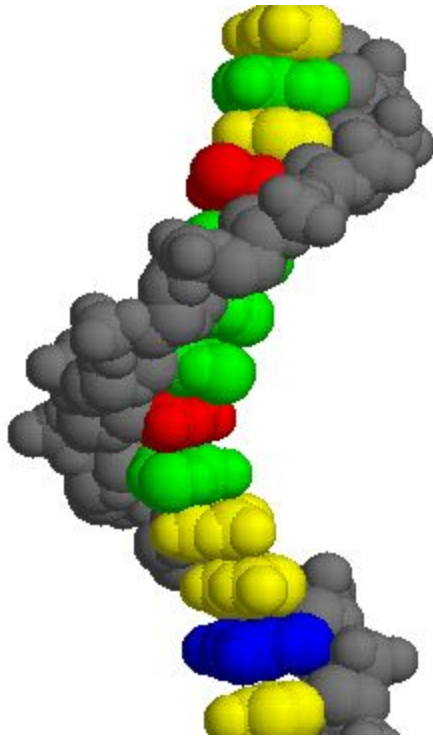


C N O P



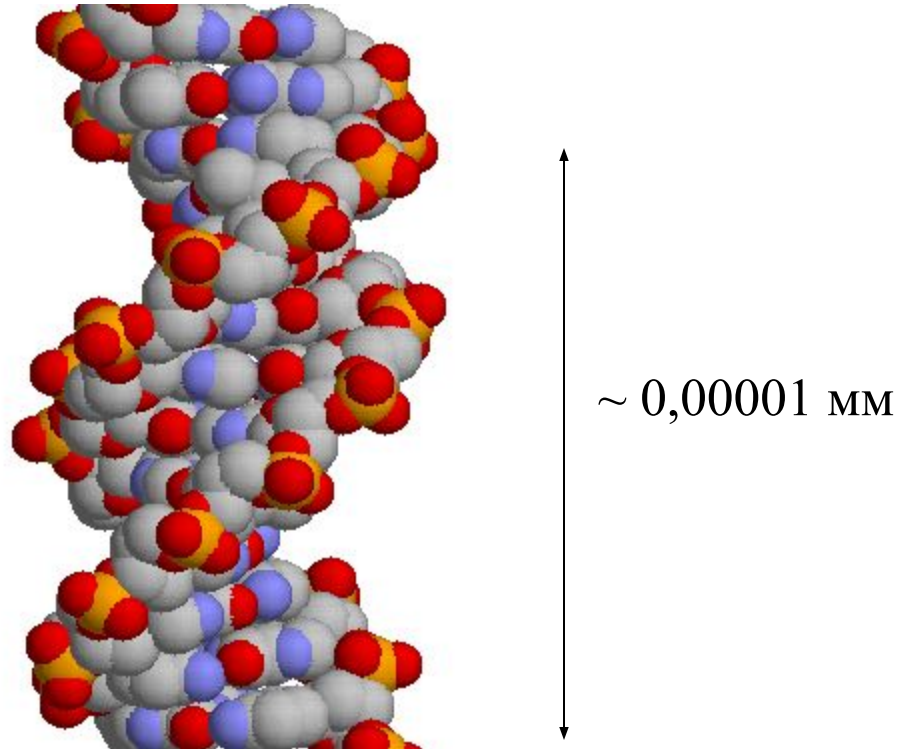
A C G T

Одну нить ДНК можно условно обозначить последовательностью букв



CGCCATAAATCAC

Детали строения молекул в микроскоп не видны!



Существуют сложные и дорогие методы
расшифровки структуры молекул

В конце 1970-х годов был открыт относительно быстрый и дешёвый метод расшифровки последовательности оснований в ДНК



Последовательность ДНК (пример)

gatcaacactacttgacttcaagacttaccataaagaaaactatagtggtggtattggcaa
aagacaagacaaatagatcaacataacaaaataaagggccatgaaatagaccatagat
caattgatttttgacaaagaaggattggcaatagaatgggggtaaagatagtccttctcaac
aaacggtaccagaatgactgaatacccacatgcaaaaagaaaaagaaatgaacctagaca
cagatcttatacagttcacaaaaatgtaactcaaatgaatcatagacctaaatataata
ttcaagactataaaaccctaaaatataacataggggaaaatctaaacaatcttgagtttg
ttaatgacttttttagatacaataccaaaggcaggatccaggaaagaatcgataagctggg
cttcattaaaattaaaatatttctgctctatgaagccactgtcaagagaaggaaaaggca
agccatagactgggagaaaatatttacaanaagacatacatgataaaggactattatccaa
aatgtacaaagaactctaaaaaacttaacaataagaaaacaaacccaactaaaaactggg
ccaaagatcttaacagatatattaccaagaagatacacagatggcaaataagcataaaa
agattaaccacatcatacgtcattaagaaattgcaaattaaaacaacaatgagacaccat
tatacacctagtagaatgacccaaatccagattactgacataatcaaatgctgacaagga
tgtggagaaacaggaactgccattcttggggtgtgggaatgccaaatgggtatgcctgctt
tggaagacagcttggtggtttcttacaacactaagcatactcttaccaaaagatcgagca

Для хранения все возрастающей информации о последовательностях ДНК в 1982 году был основан GenBank

GenBank — хранилище последовательностей нуклеиновых кислот в виде компьютерных файлов

Объем GenBank'а:

1982: 680 338 букв в 606 последовательностях

1992: 101 008 486 букв в 78 608 последовательностях

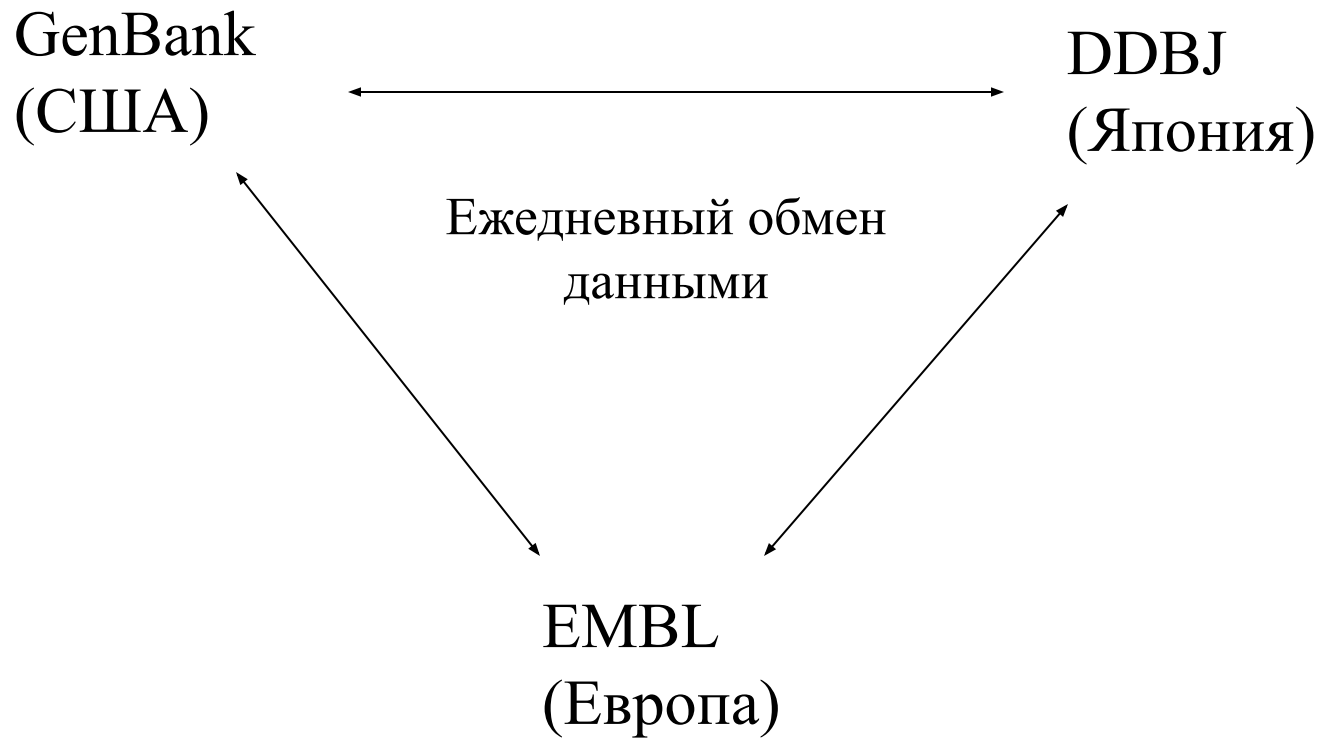
2002: 28 507 990 166 букв в 22 318 883 последовательностях

2004: 44 575 745 176 букв в 40 604 319 последовательностях

2005: 56 037 734 462 букв в 52 016 762 последовательностях
(из ~165 000 организмов)

Размер файлов — 196 Gb

International Nucleotide Sequence Database Collaboration



Структура документа GenBank'a

LOCUS AJ878089 399 bp RNA circular VRL 24-MAY-2005
DEFINITION Chrysanthemum chlorotic mottle viroid, clone CM298 VR.
ACCESSION AJ878089
VERSION AJ878089.1 GI:66571035
KEYWORDS complete genome; hammerhead ribozyme minus; hammerhead ribozyme plus; RZ+ gene; RZ- gene.
SOURCE Chrysanthemum chlorotic mottle viroid
ORGANISM Chrysanthemum chlorotic mottle viroid
Viroids; Avsunviroidae; Pelamoviroid.
REFERENCE 1
AUTHORS Gago,S., de la Pena,M. and Flores,R.
TITLE A kissing-loop interaction in a hammerhead viroid RNA critical for its in vitro folding and in vivo viability
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 399)
AUTHORS Flores,R.
TITLE Direct Submission
JOURNAL Submitted (24-JAN-2005) Flores R., Biologia del Estres, Inst. Biol. Mol. y Cel. de Plantas, Avda. de los Naranjos S/N, Valencia, 46022, SPAIN

FEATURES Location/Qualifiers
source 1..399
/organism="Chrysanthemum chlorotic mottle viroid"
/mol_type="genomic RNA"
/specific_host="Chrysanthemum grandiflora"
/db_xref="taxon:68402"
/clone="CM298 VR"
/lab_host="Chrysanthemum grandiflora cultivar Velvet Ridge"
gene join(392..399,1..53)
/gene="RZ+"
misc_RNA join(392..399,1..53)
/gene="RZ+"
/product="hammerhead ribozyme plus"
misc_feature join(399,1)
/gene="RZ+"
/note="RZ+ self-cleavage site"
gene complement(51..135)
/gene="RZ-"
misc_RNA complement(51..135)
/gene="RZ-"
/product="hammerhead ribozyme minus"
misc_feature complement(126..127)
/gene="RZ-"
/note="RZ- self-cleavage site"

ORIGIN
1 ggcacctgac gtcggtgtcc tgatgaagat ccatgacagc atcgaaacct cttccagttt
61 cggcttgtgc gggagtaaaag ctttcgctct ctccacagcc tcatcaggaa acccacttca
121 ggtctcgact ggagggctct taaacttccc ctctaagcgg agtagaggta aatacctccg
181 tcccaccocg ggaggaaaagg ggtggggacc cggaacagct ctagtcccg tcctttggag
241 tccgtttcta ccggtggaga ttacctccgg gtaagggaga cggggccagt cccagtcggt
301 tcgctctcgt agtcacagcc gctggggaac ctaggcagat ggctggacgg agtcttagtc
361 cactccaaag gaccatgggt tttaaaccct catgaggtc

Описание

Последовательность

GenBank — архивная база данных

Один эксперимент — один документ

Зачем в документе GenBank'а описательная часть?

- Ответы:** 1) чтобы пользователь банка мог найти интересующую его последовательность;
- 2) для хранения дополнительной информации (откуда ДНК, кто проводил эксперимент по секвенированию, биологическая роль данной последовательности и т.д.)

Основная проблема больших банков данных — быстрый поиск нужной информации

Для удобства пользования описательная часть документа GenBank разбита на так называемые **поля** (“fields”)

Общий принцип: любая база данных состоит, с одной стороны, из **записей** (или «документов»), а с другой стороны, из **полей**. Каждая запись есть наполнение содержанием нескольких (или всех) полей.

Пример базы данных — телефонная книга.

Записи соответствуют абонентам.

Примеры полей: фамилия, инициалы, адрес, телефон.

Основная проблема больших банков данных — быстрый поиск нужной информации

Как найти интересующую нас последовательность в GenBank'е?

Существуют специальные компьютерные программы (например, SRS или Entrez), предназначенные для поиска по **ключевым словам** в банках последовательностей.

Пользователь указывает программе, по каким полям нужно искать и какое слово (или слова). Программа выдаёт список записей банка, в которых указанные слова встретились в указанных полях.

Примеры задания на поиск

- “gene” в поле DEFINITION
- “yeast” в поле ORGANISM
- “Ivanov” в поле AUTHORS
- “yeast” в поле ORGANISM
И “tRNA” в поле DEFINITION
- “mouse” **ИЛИ** “rat” в поле ORGANISM

Как искать?

- Перебрать все 52 млн. записей, и в каждой посмотреть, есть ли в соответствующем поле заданное слово.

Это долго даже современному компьютеру!

- Заранее создать **индексную таблицу** каждого из полей и при каждом запросе обращаться к ней

Мораль: при создании программ для работы
с биологическими базами данных
необходимо использовать достижения
теории алгоритмов

Что такое биоинформатика?

- Исследование информационных процессов в биологических системах (клетках, органах, организме, популяции).
- Изучение и внедрение в компьютерную науку «биологических» методов анализа информации (нейросетей, генетических алгоритмов, нечеткой логики и др.).
- Применение компьютерных методов для решения биологических задач.
- Телепатия, парапсихология, информационные поля и т.п.

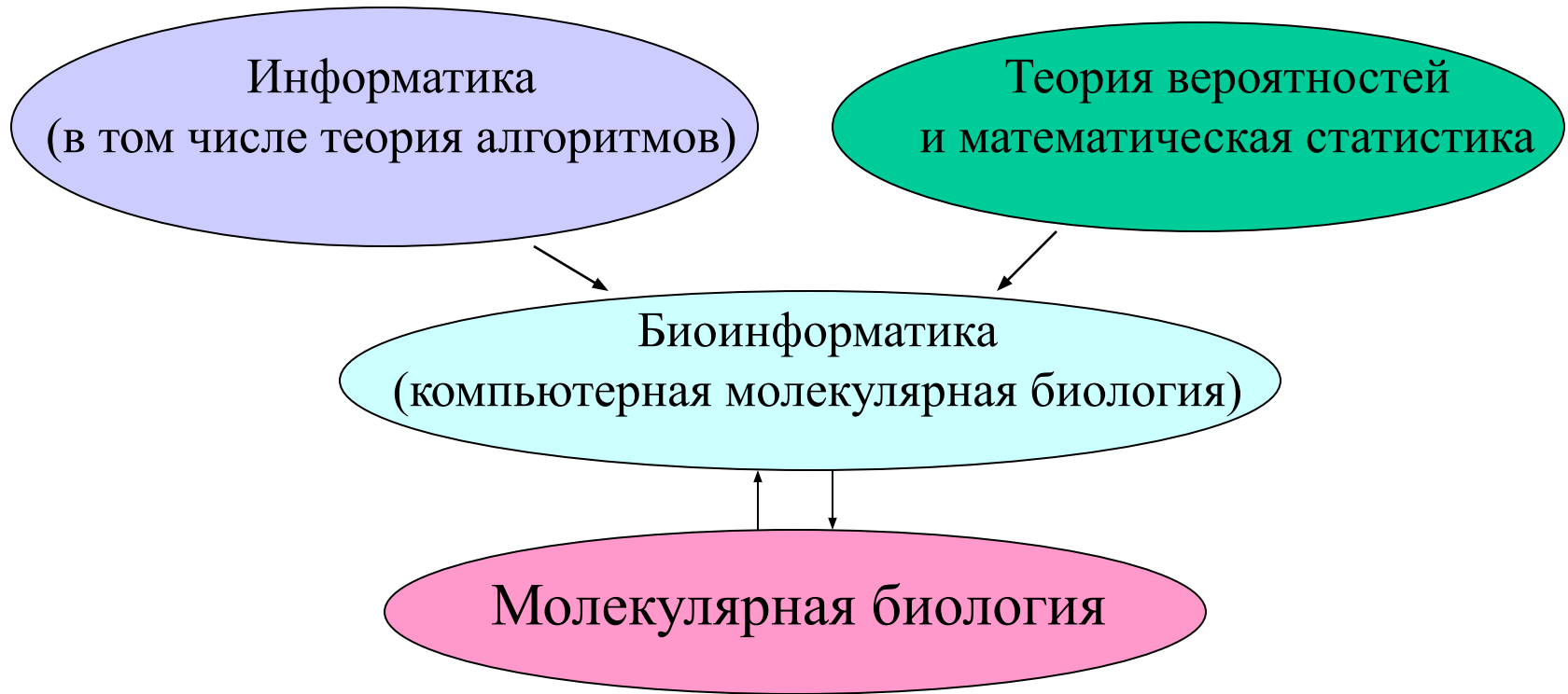


Что такое биоинформатика?

- Исследование информационных процессов в биологических системах (клетках, органах, организме, популяции).
- Изучение и внедрение в компьютерную науку «биологических» методов анализа информации (нейросетей, генетических алгоритмов, нечеткой логики и др.).
- **Применение компьютерных методов для решения биологических задач.**
- Телепатия, парапсихология, информационные поля и т.п.

Биоинформатика

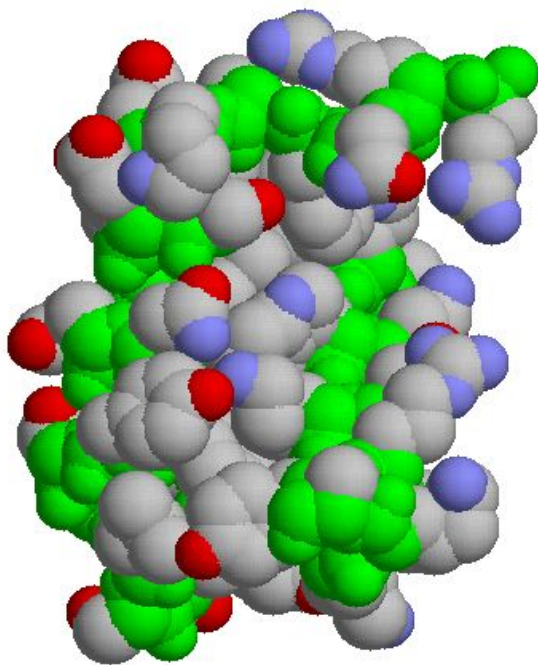
и её связи с другими дисциплинами



Основные объекты современной биоинформатики

- Последовательности нуклеиновых кислот
- Последовательности белков
- Пространственные структуры макромолекул (белков, ДНК и РНК) и их комплексов (друг с другом и с малыми молекулами)

Что такое белок



Пространственная структура

RRNFSKQASE ILNEYFYSHL
SNPYPSEEA EELARKCGIT
VSQVSNWFGN KRIRYKKNI

Последовательность

Банки структурной биологической информации

