ПРО БИОИНФРМАТИКУ КАК НАУКУ

Лекция для 1го курса у21 ААл

Примеры задач и открытий в биоинформатике, она же вычислительная биология

- рутинные методы
- новые методы для старых задач
- открытия

рутинные методы

- Получить из БД последовательности всех белков одного семейства
- Найти все домены в последовательности белка
- Построить выравнивание последовательностей гомологичных белков
- Решить подтверждает ли выравнивание гомологичность
- последовательностей по всей длине, или частично – для общего домена белков
- Найти в выравнивании консервативные участки, т.е. очень похожие у всех последовательностей
- Консервативные участки (мотивы) – значит, важные!
- Исследовать пространственную структуру белка.

новые методы для старых задач

- Найти белки одной бактерии, гены которых когда-то были перенесены в геном этой бактерии из ДНК бактерий другого вида.
- Предки бактерий этих двух видов должны были жить в одном микробном сообществе (микробиом).
- Этот процесс называется горизонтальным переносом генов (ГПГ). «Открыт» японскими учёными в 1959-1960 г.г. когда до определения последовательностей геномов были десятки лет и несколько принципиальных открытий.
- Есть несколько методов нахождения ГПГ, но их применение остаётся творческой задачей – думать надо:)

- Про ГПГ см. https://www.youtube.com/watch?v=1if1-bdE6lo

Ochiai K., Yamanaka T., Kimura K., Sawada, O. Inheritance of drug resistance (and its tranfer) between Shigella strains and Between Shigella and E. coli strains (яп.) // Hihon Iji Shimpor. — 1959. — Т. 1861. — С. 34.

Открытия и достижения

- CRISPR/Cas - я, кажется, рассказывал ранее.
- Открыли тоже японские учёные!
  Японская культура настраивает на наблюдательность и вдумчивое размышление над природными явлениями.
  Это не чуждо и российской культуре (Ломоносов, Лобачевский, Менделеев, Н.Вавилов, ……, упомяну) Кунин&Кацнельсон

- Секвенирование генома человека, нахождение в нем генов (всех ли?) – международный консорциум и институт Крэга Вентера

- У людей обнаружены гены неандертальцев и «денисовского человека»

Данные для биоинформатики получаются с помощью массовых технологий.

Новые технологии ставят новые задачи в биоинформатике. Пример: сшивки РНК-ДНК (Разин – Миронов)

Как выучиться биоинформатике?

- Смотреть и думать
  - Смотреть на доступные данные
  - Думать и задавать вопросы
  - Искать ответы на свои вопросы

Как я (AAл) выучил биоинформатику. И не я один:)

Что можно узнать или выучить имея:

- Геном
- Гены
- Протеом – таблица
- Протеом – последовательности белков

Геном.На примере вируса SARS-CoV-2

>NC_045512.2 Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1,complete genome

ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG
TTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTG
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAA
ATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCAC

CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCAC
CAAATGAATGCAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCA
GACGGGCGATTTTGTTAAAGCCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGAAGGTGCCACT
ACTTGTGGTTACTTACCCCAAAATGCTGTTGTTAAAATTTATTGTCCAGCATGTCACAATTCAGAAGTAG
GACCTGAGCATAGTCTTGCCGAATACCATAATGAATCTGGCTTGAAAACCATTCTTCGTAAGGGTGGTCG
CACTATTGCCTTTGGAGGCTGTGTGTTCTCTTATGTTGGTTGCCATAACAAGTGTGCCTATTGGGTTCCA
CGTGCTAGCGCTAACATAGGTTGTAACCATACAGGTGTTGTTGGAGAAGGTTCCGAAGGTCTTAATGACA
ACCTTCTTGAAATACTCCAAAAAGAGAAAGTCAACATCAATATTGTTGGTGACTTTAAACTTAATGAAGA
GATCGCCATTATTTTGGCATCTTTTTCTGCTTCCACAAGTGCTTTTGTGGAAACTGTGAAAGGTTTGGAT
TATAAAGCATTCAAACAAATTGTTGAATCCTGTGGTAATTTTAAAGTTACAAAAGGAAAAGCTAAAAAAG
GTGCCTGGAATATTGGTGAACAGAAATCAATACTGAGTCCTCTTTATGCATTTGCATCAGAGGCTGCTCG
TGTTGTACGATCAATTTTCTCCCGCACTCTTGAAACTGCTCAAAATTCTGTGCGTGTTTTACAGAAGGCC
GCTATAACAATACTAGATGGAATTTCACAGTATTCACTGAGACTCATTGATGCTATGATGTTCACATCTG
ATTTGGCTACTAACAATCTAGTTGTAATGGCCTACATTACAGGTGGTGTTGTTCAGTTGACTTCGCAGTG
GCTAACTAACATCTTTGGCACTGTTTATGAAAAACTCAAACCCGTCCTTGATTGGCTTGAAGAGAAGTTT
AAGGAAGGTGTAGAGTTTCTTAGAGACGGTTGGGAAATTGTTAAATTTATCTCAACCTGTGCTTGTGAAA
TTGTCGGTGGACAAATTGTCACCTGTGCAAAGGAAATTAAGGAGAGTGTTCAGACATTCTTTAAGCTTGT
AAATAAATTTTTGGCTTTGTGTGCTGACTCTATCATTATTGGTGGAGCTAAACTTAAAGCCTTGAATTTA
GGTGAAACATTTGTCACGCACTCAAAGGGATTGTACAGAAAGTGTGTTAAATCCAGAGAAGAAACTGGCC
TACTCATGCCTCTAAAAGCCCCAAAAGAAATTATCTTCTTAGAGGGAGAAACACTTCCCACAGAAGTGTT
AACAGAGGAAGTTGTCTTGAAAACTGGTGATTTACAACCATTAGAACAACCTACTAGTGAAGCTGTTGAA
GCTCCATTGGTTGGTACACCAGTTTGTATTAACGGGCTTATGTTGCTCGAAATCAAAGACACAGAAAAGT
ACTGTGCCCTTGCACCTAATATGATGGTAACAAACAATACCTTCACACTCAAAGGCGGTGCACCAACAAA
GGTTACTTTTGGTGATGACACTGTGATAGAAGTGCAAGGTTACAAGAGTGTGAATATCACTTTTGAACTT
GATGAAAGGATTGATAAAGTACTTAATGAGAAGTGCTCTGCCTATACAGTTGAACTCGGTACAGAAGTAA
ATGAGTTCGCCTGTGTTGTGGCAGATGCTGTCATAAAAACTTTGCAACCAGTATCTGAATTACTTACACC
ACTGGGCATTGATTTAGATGAGTGGAGTATGGCTACATACTACTTATTTGATGAGTCTGGTGAGTTTAAA

И еще 27 страниц такого текста …

Всего 29 903 букв.

Геном содержит информацию:
инструкцию для клеток организма хозяина (человека) как размножить вирус SARS-CoV-2. При этом хозяин заболеет!!!

Информация: Текст и читатель

Что такое информация?

Армянское радио

- «Правда ли, что Иштоян выиграл в лотерею машину?»

- «**Правда**. Но не Иштоян, а Петросян, не машину, а швейную машинку; не в лотерею, а в карты; и не выиграл, а проиграл»

«Сколько информации в этом сообщении?» И.М.Гельфанд

Протеом – как таблица

# feature	start	end	strand	name	product length
CDS	738410	741130	-	AMP-binding protein	906
CDS	741133	743274	-	aminodeoxychorismate synthase component I	713
CDS	744481	745272	-	aldolase	263
CDS	745381	745743	-	gamma-glutamylcyclotransferase	120
CDS	745916	746992	+	type 1 glutamine amidotransferase domain-conta	358
CDS	747211	748854	+	hypothetical protein	547
CDS	748994	749965	-	DUF5694 domain-containing protein	323
CDS	750284	751528	+	OmpA family protein	414
CDS	751615	751821	+	hypothetical protein	68
CDS	751913	753892	-	TonB-dependent receptor	659
CDS	753981	754877	+	LysR family transcriptional regulator	298
CDS	754893	755579	-	hypothetical protein	228
CDS	755799	756530	-	DUF1080 domain-containing protein	243
CDS	756730	757143	+	hypothetical protein	137
CDS	757190	757717	+	DUF2867 domain-containing protein	175
CDS	757722	758552	-	AraC family transcriptional regulator	276
CDS	758664	759629	+	alpha/beta hydrolase	321
CDS	759936	760382	-	hypothetical protein	148

Протеом – как посл-ти белков

```
AQAAFEASRGQIGATPLQVGGLELSIALHGYEATELTQQPEALLASDTLGFDNLRAVYRR
DGFGTALVAVFPRRSRSDPEAADTPYRDTRYLPVTATLRFDGDGGLDAVLASRTARMDVY
NPYRIDSETIAGRKVPLAANYSAAYGIWLARSELARLSLSSLLRPKQARAFKPRIYLNQP
YDPDKRVIVLVHGLASSPEAWVNLANELLGDETLRKHYQLWQVFYPTNLSILSNRAAIDA
ALTQTFAHYDPEGDDVAGRDAVLVGHSMGGVISRLLVSDSGDRVLDATLQAFDPAAAQRL
RNEPVVRELTVFKAMPQFERAVFLASPHRGAVVTDGWPLRMVRKLIRLPFDVLRETAELA
QRNEVDQDELQKIGFRKGRPPTGPDDLSPNSLFMRSTENLPIESGLPYHTIVGQRDLKLP
LLQSSDGAVPYRSAHLDGALSEKVIPSGHSVQETPQAILELRRILRVDMAEYAKRAKP
>tr|A0A0S2FHC6|A0A0S2FHC6_9GAMM Putative transcriptional regulator, ArsR family protein OS=Lysobact
MSVDSLNATFAALADPTRRAILARLAHGEAGVTELAEPFAMSLPAISKHIKVLERAGLIA
RGREAQWRPCRLETERLQEVSGWLDRYRQIWEQRLDRLDGYLVQLQAAQAERDDPPQRGD
KRDEHDPD
>tr|A0A0S2F6K2|A0A0S2F6K2_9GAMM Glyoxalase/Bleomycin resistance /Dioxygenase superfamily protein OS
MAHRSRLAGFIIDCETGDLDDAANFWSGALGLARVETYDDDGAQYAQLANGPAELHIEVQ
KVAHPSRVHLDIESDDLDAEAARLEALGAKRIAFVKRWWVMQAPTGHRFCIVRMKHPEQG
APPNVWD
>tr|A0A0S2FIB0|A0A0S2FIB0_9GAMM Uncharacterized protein OS=Lysobacter antibioticus OX=84531 GN=LA76
MTEGNPVGRRSYPEAALPLPLPLPLPVLLLCADSH
>tr|A0A0S2FAV6|A0A0S2FAV6_9GAMM Modulator of DNA gyrase family protein OS=Lysobacter antibioticus O
MDRRNFLTMSGLGIAGLMIPYGSAIAAEALLTPLDVAKKKVLADTALTAAKGAGASYCDV
RIGRYLRQFVITREDKVQNVVNTESTGIGIRVLVNGAWGFAATNQLNAKSVAEAAQQAAA
IARANSKTQTQPVVLAKTPGVGEVAWKTPIRKNAMEVPIKDKVDLLLGVNAAAVKAGADF
VNSMLFLVNEQKYFASTDGSYIDQDVHRIWAPMTITAIDKASGKFRTREGLSSPMGMGFE
YLDGAASGKTVSPNGVVNYSNSYDMLEDAVAAAKQARQKLTAPSVKPGKYDLVLDPSHTW
LTIHESIGHPLELDRVLGYEANFAGTSFATVDKVKSGFKYGSDQVTVFADKTQQGSLGAV
GFDDEGVKTKRWNLIKDGILVDYQTIRDQAHILGKSESDGCCYADSWSSVQFQRMANVSL
APGKNKLSVADMIKDVENGIYIVGDGSFSIDQQRYNAQFGGQLFYEIKNGKITGMLEDVA
YQIRTPEFWNSCVAVCDESDYRLGGSFFDGKGQPGQVSAVSHGSSTARFNGVNVINTARK
```

Ниже – приложения.

1. Про открытие CRISP/Cas системы

КОНЕЦ ПРЕЗЕНТАЦИИ

Из лекции на первом занятии первого курса ФББ в 2012 году. ААл

ОТКРЫТИЕ CRISP/CAS

Как бактерии защищаются от вторжения чужеродной ДНК?

- Один механизм изучен в 1960-70х г.г. (Нобелевская премия 1978 г.): системы рестрикции-модификации
- Другой открыт в 2007 г.: **CRISPR**

2021.
Сегодня известно несколько десятков разных типов защитных систем бактерий и архей

Открытие 1.

- (**1987)** Ishino с соавторами обнаружили загадочную последовательность в ДНК кишечной палочки, штамм К12.

Ishino Y et al. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. JBacteriol. 1987

TGAAAATGGGAGGGAGTTCTACCGCAGAGGCGGGGGAACTCCAAGTGATATCCATCATCGCATCCAGTGCGCC (1,451)
(1,452) CGGTTTATCCCCGCTGATGCGGGGAACACCAGCGTCAGGCGTGAAATCTCACCGTCGTTGC (1,512)
(1,513) CGGTTTATCCCTGCTGGCGCGGGGAACTCTCGGTTCAGGCGTTGCAAACCTGGCTACCGGG (1,573)
(1,574) CGGTTTATCCCCGCTAACGCGGGGAACTCGTAGTCCATCATTCCACCTATGTCTGAACTCC (1,634)
(1,635) CGGTTTATCCCCGCTGGCGCGGGGAACTCG (1,664)

consensus: CGGTTTATCCCCGCT$^{GG}_{AA}$CGCGGGGAACTC

FIG. 5. Comparison of direct-repeat sequences consisting of 61 base pairs in the 3′-end flanking region of *iap*. The 29 highly conserved nucleotides, which contain a dyad symmetry of 14 base pairs (underlined), are shown at the bottom. Homologous nucleotides found in at least two DNA segments are shown in boldface type. The second translational termination codon is boxed. The nucleotide numbers are in parentheses.

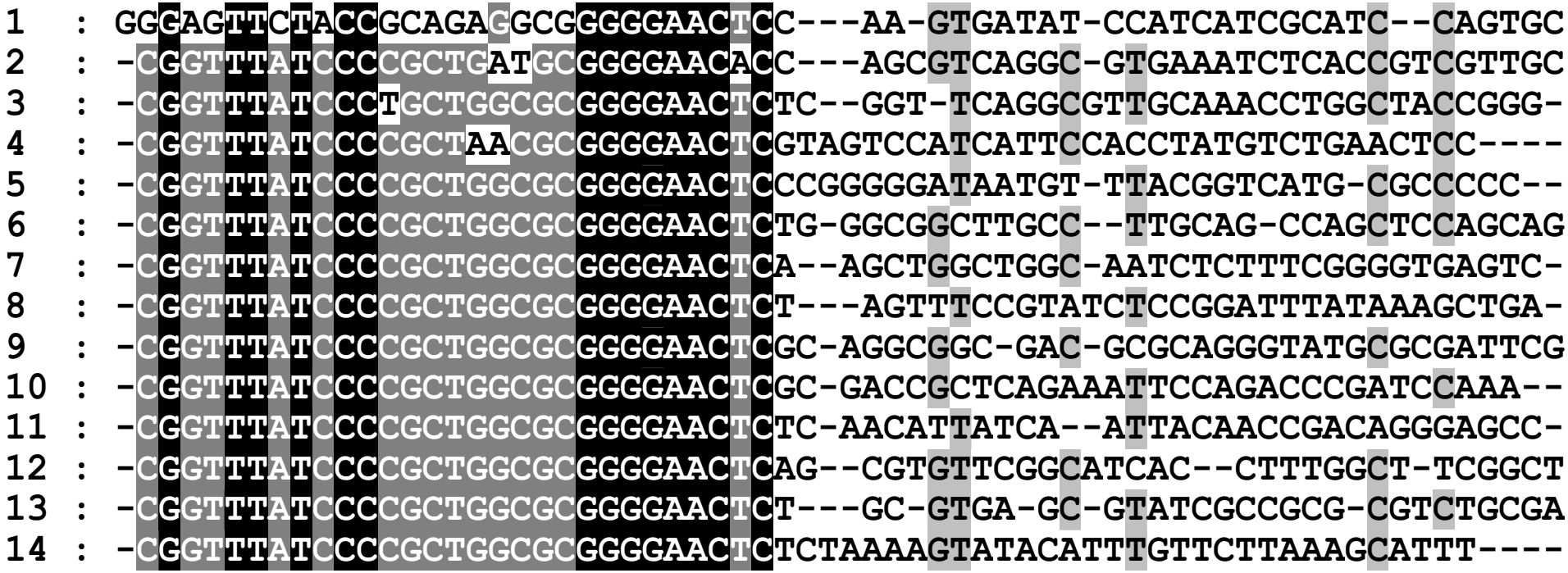
An unusual structure was found in the 3′-end flanking region of *iap* (Fig. 5). Five highly homologous sequences of 29 nucleotides were arranged as direct repeats with 32 nucleotides as spacing. The first sequence was included in the putative transcriptional termination site and had less homology than the others. Well-conserved nucleotide sequences containing a dyad symmetry, named REP sequences, have been found in *E. coli* and *Salmonella typhimurium* (28) and may act to stabilize mRNA (18). A dyad symmetry with 14 nucleotide pairs was also found in the middle of these sequences (underlining, Fig. 5), but no homology was found between these sequences and the REP sequence. So far, no sequence homologous to these has been found elsewhere in procaryotes, and the biological significance of these sequences is not known.

Вот этот фрагмент последовательности ДНК *E.coli.*

>ecoli_crispr1

TGGGTTTGAAAATGGGAGCTGGGAGTTCTACCGCAGAGGCGGGGGAACTCCAAGTGATAT
CCATCATCGCATCCAGTGCGCCCGGTTTATCCCCGCTGATGCGGGGAACACCAGCGTCAG
GCGTGAAATCTCACCGTCGTTGCCGGTTTATCCCTGCTGGCGCGGGGAACTCTCGGTTCA
GGCGTTGCAAACCTGGCTACCGGGCGGTTTATCCCCGCTAACGCGGGGAACTCGTAGTCC
ATCATTCCACCTATGTCTGAACTCCCGGTTTATCCCCGCTGGCGCGGGGAACTCCCGGGG
GATAATGTTTACGGTCATGCGCCCCCCGGTTTATCCCCGCTGGCGCGGGGAACTCTGGGC
GGCTTGCCTTGCAGCCAGCTCCAGCAGCGGTTTATCCCCGCTGGCGCGGGGAACTCAAGC
TGGCTGGCAATCTCTTTCGGGGTGAGTCCGGTTTATCCCCGCTGGCGCGGGGAACTCTAG
TTTCCGTATCTCCGGATTTATAAAGCTGACGGTTTATCCCCGCTGGCGCGGGGAACTCGC
AGGCGGCGACGCGCAGGGTATGCGCGATTCGCGGTTTATCCCCGCTGGCGCGGGGAACTC
GCGACCGCTCAGAAATTCCAGACCCGATCCAAACGGTTTATCCCCGCTGGCGCGGGGAAC
TCTCAACATTATCAATTACAACCGACAGGGAGCCCGGTTTATCCCCGCTGGCGCGGGGAA
CTCAGCGTGTTCGGCATCACCTTTGGCTTCGGCTGCGGTTTATCCCCGCTGGCGCGGGGA
ACTCTGCGTGAGCGTATCGCCGCGCGTCTGCGAAAGCGGTTTATCCCCGCTGGCGCGGGG
AACTCTCTAAAAGTATACATTTGTTCTTAAAGCATTTTTTCCCATAAAAACAACCCACCA
ACCTTAATGTAACATTTCCTTATTATTAAAGATCAGCTAATTCTTTGTTTT

Выравнивание повторов



Почему в названии CRISPR –

Clustered Regularly Interspaced Short Palindromic Repeats

есть слова “палиндромный повтор”?

Похожие повторы были найдены в геномах многих бактерий

- **(1993)** Через 5 лет Groenen с соавторами нашел похожую последовательность в геноме палочки Коха *Mycobacterium tuberculosis*
- **(1995)** Mojica et al. нашел похожую последовательность в ДНК бактерий *Haloferax volcanii* и *Haloferax mediterranei*
- **(1997)** Goyal et al. использовали подобные последовательности для определения штамма бактерий.
- **(2000)** Mojica et al. нашли похожие последовательности в геномах многих бактерий и архей
- Разные авторы называли эти последовательности по разному:
  - TREPs
  - SRSRs
  - SPIDRs
  - **CRISPRs**
  - LCTRs

>AB553331 Streptococcus dysgalactiae subsp. equisimilis DNA, CRISPR2

gatgcaatccactcacccgcgaagggtgagacatgacatccttgacgga
catgccaaaatcagaacatttcaatccactcacccgcgaagggtgagac
caagtaatcagttgagagcaggcagtggattacaatatttcaatccact
cacccgcgaagggtgagacagagataaagaattaacagaaaggcaggtt
tataaaatttcaatccactcacccgcgaagggtgagacggtcgagaaag
tagaatttgctaggttgcaatttatttcaatccactcacccgcgaaggg
tgagacgaggaattgctccttgactttagcaagccacaagatatttcaa
tccactcacccgcgaagggtgagactcttgactgtgatggagactatga
gagagccagaatttcaatccactcacccgcgaagggtgagac

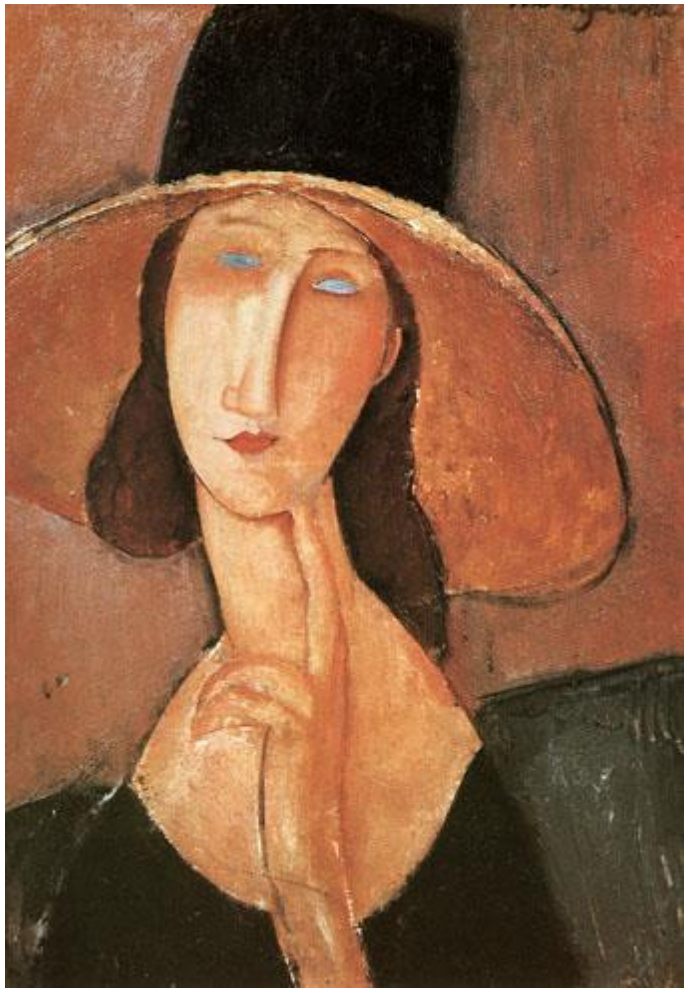
>DQ072993 Streptococcus thermophilus strain JIM 8229 DNA, CRISPR repeat sequence.

tagttaccgtataagatattcccaaacatctgatgaaaaacttttacagaaattt
ttagaaagtaaggattgacaaggacagttattgtttttataatcactatgtgggt
ataaaaacatcaaaatttcatttgaggttttgtactctcaagatttaagtaact
gtacaacgtacttcaaaggttctaactacataacacagttttgtactctcaaga
tttaagtaactgtacaactaaaaccagatggtggttcttctgatactagttttg
tactctcaagatttaagtaactgtacaaccattttcttcagtcaattcgttctca
agcggttttgtactctcaagatttaagtaactgtacaacaaaggacgggggcaa
tgaacaaacgacaacgttttgtactctcaagatttaagtaactgtacaactaat
atcattgatagcttcatcaaaggctgttttgtactctcaagatttaagtaactg
tacaactaaattgttccttgactccgaactgccctgttttgtactctcaagatt
taagtaactgtacaacaaacaatcgtttatctatcctcaaaggatggttttgta
ctctcaagatttaagtaactgtacaacataaaaaaacgcctcaaaaaccgagaca
acgttttgtactctcaagatttaagtaactgtacaacataaaaaaacgcctcaa
aaaccgagacaacgttttgtactctcaagatttaagtaactgtacagtttgatt
caacttaaaaagccagttcaattgaacttggcttttaaaatacgcgatagacat
aaggattgtcaggctgtccgacctctttaacttcagtcaaattgaggataggtag
gctctgtttgagattttgatagta

>emblrelease|GU192460|GU192460 Dickeya sp. 409 CRISPR region genomic sequence.

ccttcagcacccttgttcctgcacttaatcaagatgagacgcagcgctgg
cgccgccggccagccccagtaacagaatgagtgaaaaccgtttttcatga
gagttccttgcaagcctgtcaggcaaaagcgccactgtagcatgccgtttc
tgccgctgccggttttgacccttttttttcggcagctcataactaattgat
ttttaatgacgaaaatattcgactttaaaaaagggttttccaggaaaaatc
cagatttccctttaaaaatcagttaatagacgataaattgctacgtgttca
ctgccgtgtaggcagcttagaaaaagaaagacaggtaaagaaggtattatc
tggcgttcactgccgtgtaggcagcttagaaaggcaaagccggtaagctcc
gccgaacccgcaagttcactgccgtgtaggcagcttagaaaagattgattt
ttgcgtccaagcgctgacgtcggttcactgccgcacaggcagagattgatt
ggtttgctggcgttaaaaactacgctgaggtgggc

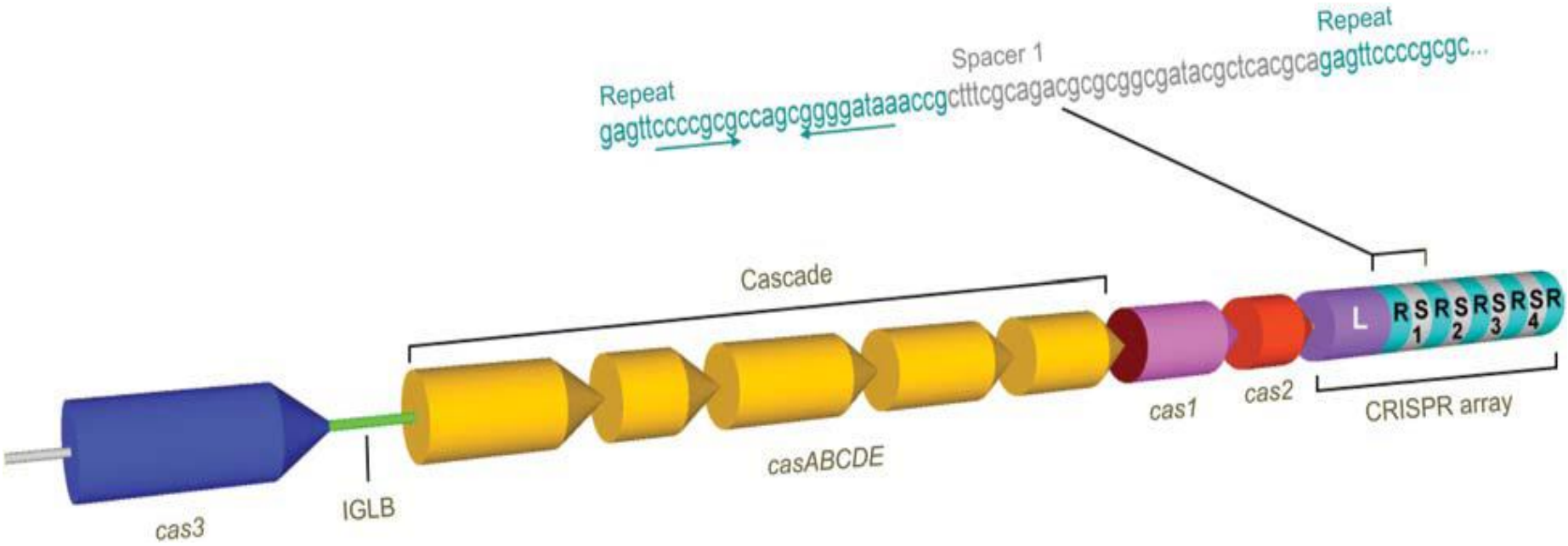
Сходство – одно из основных понятий биоинформатики



Модильяни. Портреты Жанны Эбютернь. 1918

Открытие 2

- **(2002)** Jansen et al. обнаружили, что рядом с CRISPR на ДНК почти всегда *закодированы похожие гены*, названные **cas**.



Современные данные о строении CRISPR кассеты. Схема ДНК.

R - повтор

S - участки между повторами

L – участок между генами и CRISPR

Остальные цилиндры - кодирующие последовательности

Высказывались разные гипотезы о функции CRISPR

- CRISPR отвечают за
  - развитие бактерии (Thony-Meyer и Kaiser, 1993)
  - правильную дупликацию ДНК (репликацию) при делении бактерии (Mojica et al., 1995)
  - исправление “ошибок” в ДНК (Makarova et al., 2002).

Открытие 3 и Гипотеза

- CRISPR содержат участки, очень похожие по последовательности на участки ДНК бактериофагов! (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005)

  Все три группы исследователей предположили, что CRISPR служит для защиты от фагов

- Makarova et al., 2006, собрали все данные о CRISPR в геномах прокариот и обосновали эту гипотезу методами биоинформатики.

Доказательство гипотезы

- Barrangou et al. (2007). Гипотеза доказана экспериментально: наличие в ДНК бактерии CRISP кассеты защищает бактерию от заражения бактериофагом
  (не любым, а тем, кусочек последовательности которого встроен в ДНК бактерии).
- CRISPR/Cas система – активная прокариотическая иммунная система против бактериофагов и других видов чужеродной ДНК (He and Deem, 2010)

“However, conservation of underlying principles of CRISPR immunity in different species was shown recently, by introduction of S. thermophilus CRISPR-3 into E. coli conferring heterologous protection against plasmid and phage”

(Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P. & Siksnys, V. (2011). The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Res)

CRISPR открыты “на кончике пера” –

с помощью анализа последовательностей ДНК и биоинформатики.

Открытие (точнее, предсказание) подтверждено экспериментально.

Сравните с открытием планеты Нептун математиками Леверье (Франции) и Адамсом (Англии), подтвержденное астрономами Галле и д’Аррестом 23 сентября 1846 года!

КОНЕЦ ПРО CRISPR/CAS

Гены и сигналы

2. ЧТО ЗАПИСАНО В ГЕНОМЕ?

Что видим своими глазами

>NC_045512.2 Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1,complete genome **SARS-CoV-2**

ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG
TTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTG
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAA
ATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCAC

Что видим своими глазами

- В геноме четыре буквы A, T, G, C
  (понятно)

- Буквы идут неупорядоченно,
  похоже на случайную последовательность

Лингвистический анализ текста

- Правда ли, что в файле в последовательности генома нет других букв?
- Частоты букв
- Часто и редко встречающиеся слова
- Равномерность частоты букв и слов вдоль текста

Эти вопросы изучаются и имеют биологически смысл! Примеры наблюдений:

- #C ≈ #G, #T ≈ #A (# = число)
- Слов CG *мало* в определенных геномах
- Слов TA *мало* во всех геномах
- В некоторых геномах #C > #G в одной части и #G > #C в другой части («GC skew»)

“Много, нормально, или мало?”

- Чтобы ответить надо знать сколько - нормально: сколько изучаемых слов ожидается, если предположить, что никакой причины, влияющей на число слов нет – чистая случайность!
  - Можно предположить, что буквы A, T, G, C в геноме имеют одинаковую частоту ¼ и проверить так ли это в вашем геноме
  - Можно использовать наблюдаемые в вашем геноме частоты букв и вычислить сколько слов TA ожидается в вашем геноме, если соседние буквы встречаются случайно и независимо друг от друга и сравнить с наблюдаемым в геноме числом слов TA
    - Подсказка: как вычисляется вероятность (частота) двух независимых событий, если вероятности каждого их них известны? Например, события (1) увидеть ворону по дороге в МГУ для сдачи зачёта и (2) получить зачёт похоже независимы.

Всерьёз думают о живых вакцинах, основанных на вирусах с увеличенным числом CG или TA!

Interestingly, most mammalian RNA viruses have low frequencies of CpGs (45,46). Furthermore, viruses with high CpG frequencies may be more recognizable by pathogen innate immune sensors (47–50).

Attenuation of the classical oral poliovirus vaccine is based on very few point mutations, which can revert to virulence after a few rounds of viral replication (144). These pioneering results obtained with recoded polioviruses suggest that codon-usage in recoded viruses may be much more stable than most RNA virus point mutants, and could possibly enable the development of live attenuated RNA virus vaccines with superior genetic stability.

Martinez et al., 2019, NAR