Темы исследований для включения в мини-обзор

Выбираете самостоятельно!

Темы, придуманные самостоятельно, принимаются и проверяются. С добавлением баллов, если исследование завершилось успехом

Откройте окошко kodomo Будут упражнения

Здесь обсудим вопросы по заданиям пр.12.all

Консультация

Deadline и штраф за опоздание

Where freedom is given, liberty is taken.

Здесь обсудим задание пр.13

Цель задания – исследовать небольшое число тем, подготовить результаты по ним для включения в мини-обзор.

Набор тем такой, чтобы было достаточно для получения зачёта за мини-обзор

Результаты – в ЭТ таблице с сопроводительными материалами. Формулы должны оставаться для проверки

Смотрим задание, выполняемое в день занятия

Минимум для зачёта мини-обзора

- Две обязательных темы:
  - Размеры геномных ДНК
  - Таблица генов по типам с указанием числа таких генов. Типы генов: белок – РНК; для белков – ген или псевдоген; для РНК – все типы, которые встречаются, tRna, rRNA и т.п.
- >= одной темы с результатами в виде гистограммы (как рисунок)
- >= одной темы с результатами в виде таблицы
- >= одной темы по исследованию генома
- >= одной темы по исследованию протеома
- Сопроводительные материалы в виде ЭТ с анализом данных, включаемых в текст мини-обзора

ИТОГО – не менее … 4х тем!

Предупреждение

- Главная цель студента в первом семестре – получить все зачеты и сдать экзамены с оценкой >2
- Сначала выполните необходимое число простых исследований
- Потом беритесь за сложные и интересные

«Лингвистика» генома

Что можно узнать анализируя только текст - последовательность генома

3. Нуклеотидный состав ДНК генома

- (Борис) Проверить какие буквы встречаются в последовательности геномной ДНК и сколько раз. Верно ли, что только A, T, G, C?
- (ААл) Верно ли, что число букв A примерно равно числу букв T, а число букв G приблизительно равно числу букв C в последовательности одной цепочки геномной ДНК? (Второе правило Чаргаффа)
- Упражнение:

wordcount -wordsize 1 XXXXXXX_genome.fasta stdout

wordcount –help -verbose

4. (Екатерина Тычкова) вычислить частоты комплементарных пар A-T и G-C в геномной ДНК.

- GC – состав генома (%). Почему и зачем?
- Потому, что характеристика генома. Примерно постоянен вдоль генома. Большие участки с отличием GC-состава от среднего требуют поиска причины.
- Для бактерий – подозрение на горизонтальный перенос ДНК

Исследования генома

- 5. (Максим Смирнов) Выполнить анализ k-меров в геноме
  - Разобрать для 2-меров (динуклеотидов). TA и CpG
- 6. (Ума, Дарья Латорцева) Найдите и опишите повторяющиеся последовательности в геноме, появление которых нельзя объяснить случайностью. Источники повторов: паралоги, мобильные элементы, шпильки (инвертированные повторы), узнаваемые последовательности Трудное
- 7. (ААл) Повторы - пока не подобрал какие/ Инвертированные
- 8. (Анастасия) Как определяется начало кольцевой ДНК в файле с последовательностью генома?(ААл) Найдите место начала репликации - origin - в хромосоме вашего генома и место терминации репликации Простое и интересное. Но надо кое-что узнать

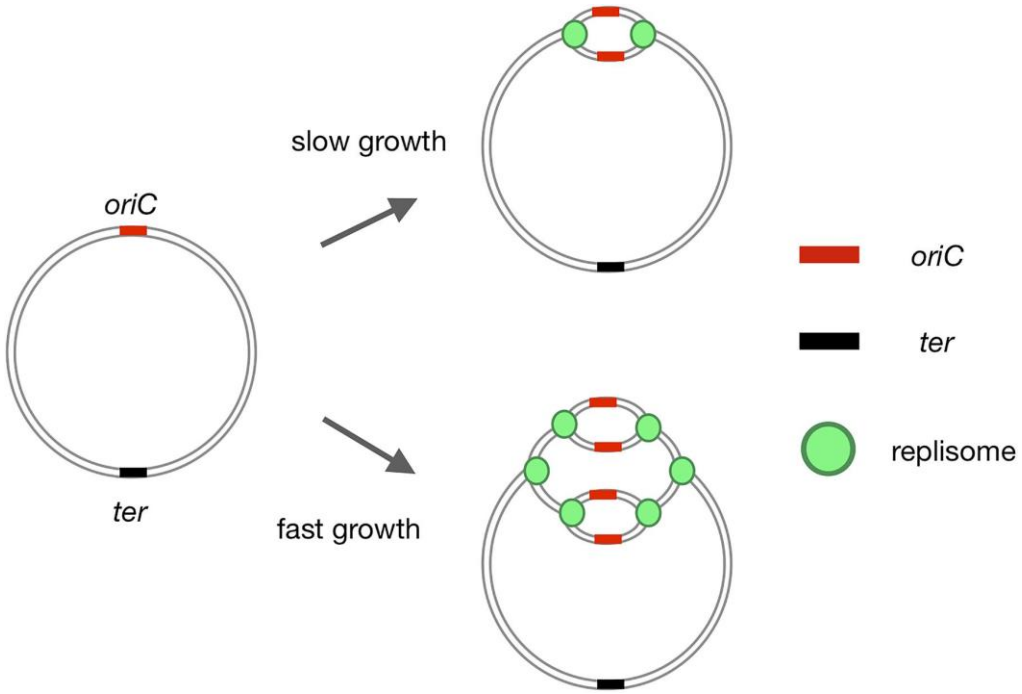
Репликация ДНК у бактерий

У бактерий с кольцевой хромосомой репликация (удвоение ДНК) начинается с определенного места и такое место одно. Оно называется местом начала репликации, по англ. Origin of replication

В этом месте ДНК расплетается.

Начиная от этого места к каждой цепочке начинается достраивание комплементарной цепочки ДНК в обе стороны.

Схема репликации и бактерий



Trojanowski et al.
Where and When Bacterial Chromosome Replication Starts: A Single Cell Perspective
Front. Microbiol., 2018 |
https://doi.org/10.3389/fmicb.2018.02819

Два способа достройки комплементарной ДНК

ДНК достраивается только с 3' конца. Поэтому в одну сторону идет непрерывное достраивание комплементарной цепочки до места терминации. Соответствующая часть ДНК от origin до terminator называется лидирующей цепочкой.

А как же растет комплементарна цепочка ДНК с 5' конца? Не поверите, но по кусочкам, в правильную сторону - фрагменты Оказаки (см. рис.). Эта часть ДНК называется отстающей.

При достраивании комплементарной цепи все идет наоборот.

Лидирующая и отстающая цепочки

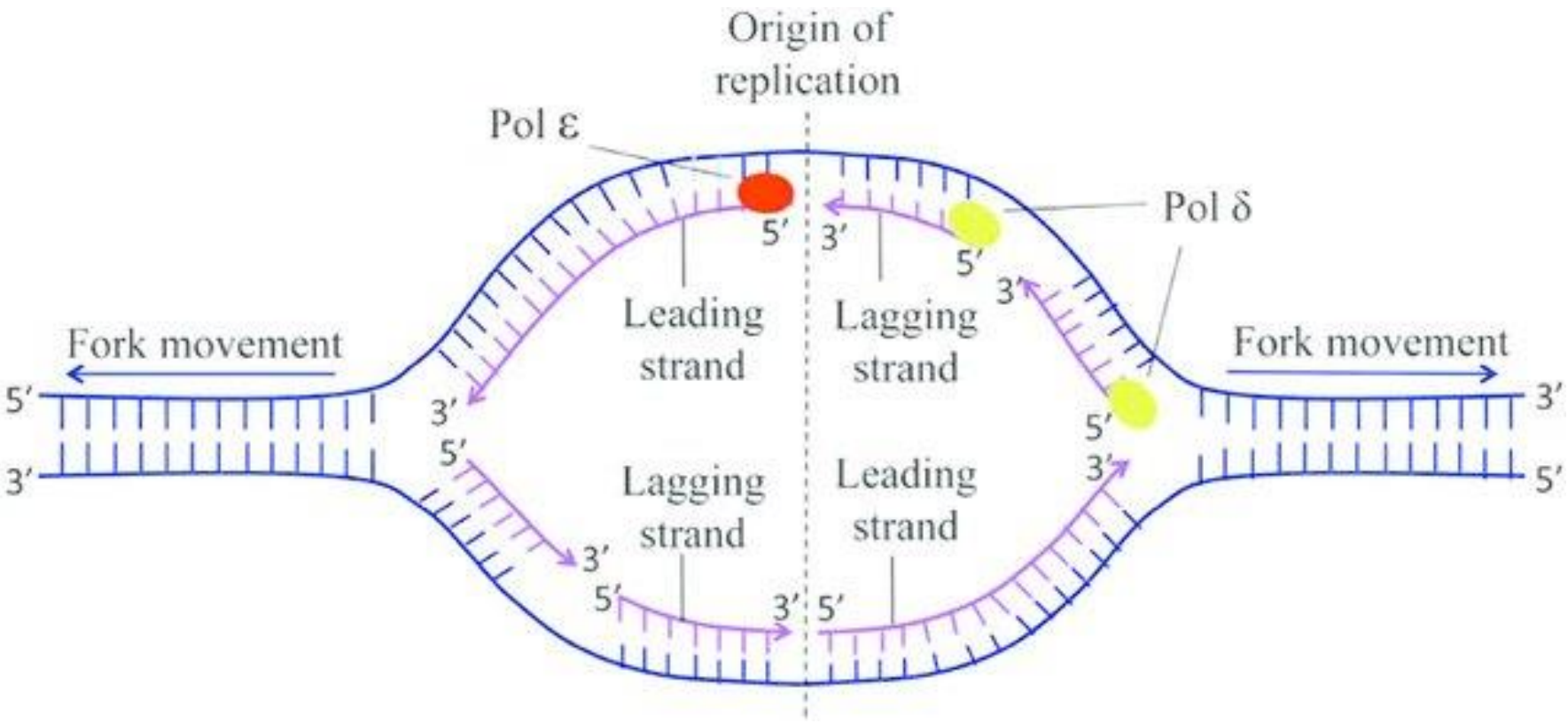
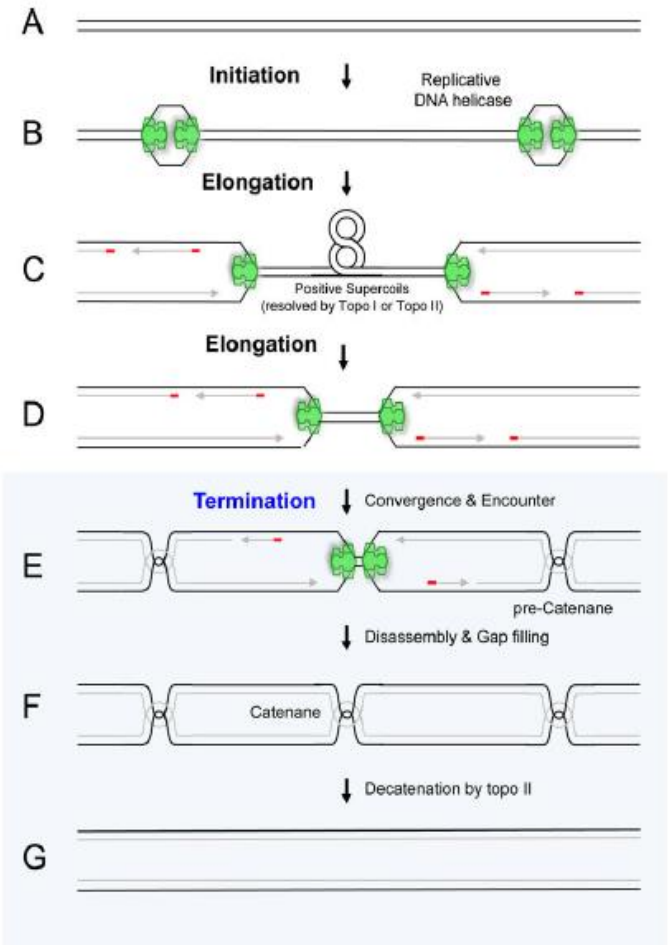


Figure 1: The schematic diagram of origin of replication of human. The process of DNA replication requires two DNA

This figure was uploaded by Wei Chen

Терминация репликации



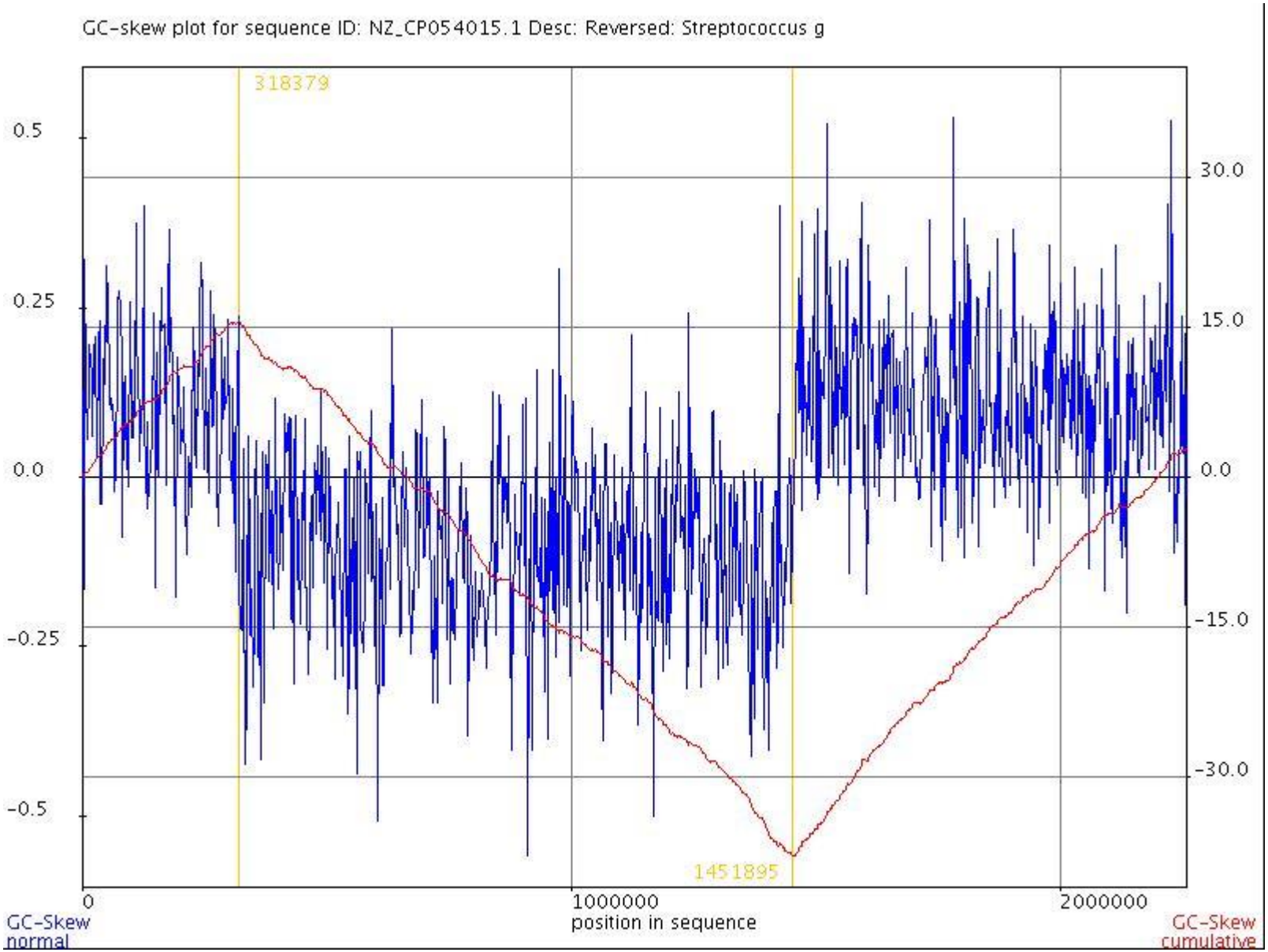
Published in final edited form as:
Nat Rev Mol Cell Biol. 2017 August ; 18(8): 507–516. doi:10.1038/nrm.2017.42.

Mechanisms of DNA replication termination

Неоднородность G и C в одной цепочке ДНК. Эмпирический факт:

- На лидирующей цепи число гуанинов G меньше числа цитозинов C .
- На отстающей цепи число G больше числа C
- В местах oriC и ter происходит резкая смена соотношения G и C
- На этом основан алгоритм определения oriC и ter в кольцевой хромосоме бактерии. Вычисляется величина GC-skew, строится график интегральной величины GC-skew, oriC соответствует минимуму GC-skew cumulative, a ter (терминатор) - максимуму
- Сервис http://genskew.csb.univie.ac.at/ Все объяснено
- В районе oriC есть специальные короткие последовательности, но они разные для разных геномов

Выдача сервера GC-skew



9. Анализ G-квадруплетов в геноме бактерии

- У человека (эукариот?) G-квадруплексы регулируют экспрессию генов на уровне транскрипции (лекция 12)
- То же известно для вирусов человека, что ожидаемо
- У бактерий вопрос изучен плохо (ссылка)
- Постройте гистограмму расстояний до ближайшего гена
- Это не вполне то, что нужно. Гипотеза: у прокариот G-квадруплексы регулирую транскрипцию. Ожидается, что расположены рядом со стартом транскрипции, т.е. перед оперонами. Но информацию об оперонах не легко получить

fuzznuc -complement -pattern "GGGN(1,7)GGGN(1,7)GGGN(1,7)GGG"
<имя входного фаста файла > <имя выходного текстового файла>

Анализ протеома и генов

10. (Михаил Никонов, Артем Васильев) Является ли распределение генов белков по цепям ДНК (прямой + и обратной -) случайным?

- Если из 100 бросаний монеты 44 раза выпал орел и 56 раз – решка это соответствует случайному бросанию или монета кривая? Как решить?
- Случайное бросание монеты значит, что
  - орел выпадает с вероятностью ½
  - результат очередного бросания не зависит от результатов предыдущих бросаний (бросающий не в состоянии обучиться бросать так, чтобы чаще выпадала решка)

Экспериментальный метод решения

1. Бросим монету 100 раз. Запишем сколько орлов выпало
2. Повторим п.1 1000 раз. Посчитаем сколько раз из 1000 выпало 44 орла или меньше, обозначим это число буквой k.
3. Если k = 2 то k/1000 = 0.002 Значит наше наблюдение 44 орла или еще более далекое от ожидаемого числа 50, наблюдается с частотой 2 на тысячу. То же самое можно сказать так: с вероятностью p = 0.002
4. События с вероятностью 0.002 в биологии (и не только) считаются противоречащими гипотезе о случайном бросании. Надо искать причину – кривая монета, или другие условия бросаний некорректны
5. Максимальная вероятность, которую в биологии можно считать противоречащей случайности, равна p = 0. 05 (50 случев на 1000 серий)

Не волнуйтесь – бросать монету не обязательно

1. В Excel можно в ячейке B2 написать функцию =случмежду(0,1) (она же =randbetween(0,1)) 0 – решка, 1 - орёл
2. Растянуть её направо сто раз
3. В ячейке A2 поместим сумму по строке от B2 до конца данных. Это и будет число орлов.
4. Распространим строку 2 вниз 1000 раз. Вот и получем 1000 серий по сто бросаний монеты.
5. Аналогично можно провести эксперименты со случайным распределением генов по цепочкам. Число серий можно сократить до 100 или 200 чтобы поместилось, т.к. генов может быть несколько тысяч. И гены расположить по строкам я одну серию – в столбце

Вероятность можно рассчитать

- Пять орлов при 20-ти бросаниях: случайность или монета кривая?
- Число “успехов” при многих бросаниях имеет распределение Бернулли, оно же – биномиальное
- Найдем вероятность **P** того, что при двадцати *независимых* бросаниях с вероятностью успеха 0.5 выпадет 5 или менее успехов
- В Excel: fx => статистические => биномраспр:
  - число успехов = 5
  - число испытаний = 20
  - вероятность успеха = 0.5
  - интегральная = ИСТИНА (т.к. <= 5 успехов)
    Ответ: **P = 0.02**
- Интерпретация: вероятность получить такой исход 20-ти бросаний при правильной монете равна 0.02.
  Очевидно,
  еще такая же вероятность получить 15 или более решек
- 1000 генов на прямой цепи ДНК из 2200и генов всего:
  - случайность или
  - имеет биологический смысл?

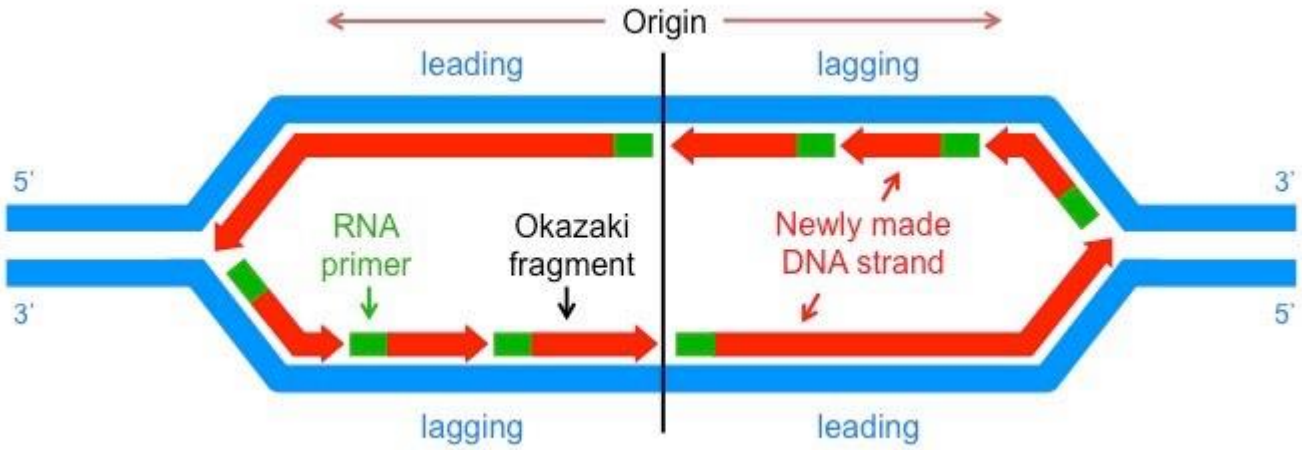
Еще про протеом, гены и ORF

- 11. (ААл) Построить гистограмму длин белков
- 12.(Александр Неверов) Найдите открытые рамки считывания (open reading frame ORF) в вашем геноме
  - a. На странице ЭТ составьте список координат всех открытых рамок считывания в геноме от START до STOP
  - b. Сравнить координаты orf с координатами генов белков. Как минимум привести примеры совпадений и несовпадений. Технически, не простое, но выполнимое задание
- 13. (Артём Васильев) Какую часть генома занимают последовательности, кодирующие белки (CDS)? Предостережение а вдруг гены пересекаются?
- 14.(ААл) Вычислите число гипотетических (hypothetical или ... сами попробуйте выбрать) белков в геноме и процент от всех белков
- 15.(ААл) Составьте таблицу названий и координат (включая ориентацию) рибосомальных (ribosomal) белков и рибосомальных РНК, закодированных в геноме. Хорошо бы сделать общий google doc для них. Был бы частичный ответ на вопрос Софии Наварновой.
- 16. (Анна) Какие нуклеотиды стоят в третей позиции кодонов? (ААл) Составить таблицу частот использования кодонов кодирующих одну и ту же аминокислоту

Темы, пока не сформулированные достаточно конкретно

- 17. Изучите "квазиопероны" в геноме вашей бактерии или археи. Статистика числа генов в квазиоперонах.
  - Оперон – последовательность нескольких генов белков, транскрибирующихся в одну мРНК
    Очевидно, гены в опероне идут подряд на одной цепочке ДНК.
    Расстояние между генами в опероне небольшое
    часто гены в одном опероне связаны по функциям
  - Квазиоперон – максимальная последовательность генов идущих подряд на одной цепочке ДНК с небольшим расстоянием между генами. Например, межгенное расстояние <100 п.н.
    Квазиопероны определяют для предсказания оперонов, так как каждый оперон всегда содержится в квазиопероне
- 18. Почему бы не сравнить числа генов белков в шести рамках считывания? Вдруг что-нибудь неожиданное обнаружится.
- 19. Гистограмма длин межгенных промежутков.
- 20. Статистика белков по категориям достоверности их существования. (Uniprot) Подсказку напишу.

Конец презентации



https://ib.bioninja.com.au/higher-level/topic-7-nucleic-acids/71-dna-structure-and-replic/origins-of-replication.html