Лекція 9. СЕКВЕНУВАННЯ.

Існує декілька методів секвенування ДНК, але найбільш популярним серед них є метод секвенування з обривом ланцюга, вперше застосований в 1978 р. Фредом Сенгером. Цей метод достатньо легко може бути автоматизований, що значно прискорює процедуру.

Сутність метода: На достатньо коротких (до 1000 н) одноланцюгових ДНК здійснюється відпал коротких послідовностей олігонуклеотидів. При цьому в середовище для реакції додано окрім звичайних дезоксирибонуклеотидтрифосфатів (дАТФ, дГТФ, дЦТФ та дТТФ), невелика кількість дидезоксинуклеотидтрифосфатів (ддАТФ, ддГТФ, ддЦТФ та ддТТФ), в яких вуглевод не має гідроксильної групи при 2 та 3 атомах карбону і тому є нездатним формувати фосфодіефірний зв'язок з наступним дНТФ і подовжувати ланцюг. Окрім того, кожен із ддНТФ відмічено флуоресцентною міткою. Фермент полімераза не вибирає дНТФ або ддНТФ приєднувати до ланцюга, тому обрив ланцюга трапляється випадково і в підсумку отримують набір олігонуклеотидів різної довжини, які можна розділити на чотири родини, кожна з якої буде завершуватись або ддАТФ, або ддГТФ, або ддЦТФ, або дТТФ, який присутній в тій же позиції, що і в матричній ДНК.

Для того, щоб визначити цю позицію, ДНК загружають у лунку поліакриламідного гелю (розмір пор дозволяє розрізняти ланцюги від 10 до 1500 п.н., що різняться на 1 нуклеотид) та проводять електрофорез з метою розділення молекул за довжиною. Після розділення молекули рухаються вздовж датчика флюоресценції, який реєструє сигнал від останнього нуклеотида. Далі інформація передається у систему формування зображення і виводиться на екран у вигляді піків, типових для кожного маркера, а отже відповідають положенню конкретного нуклеотида.

|  |  |
| --- | --- |
|  | |
|  | |
|  |  |
| G A C T |  |

До проблем, які можуть виникнути при застосуванні цього метода, можна віднести необхідність попереднього копіювання матричної ДНК, застосування затравок та певних ДНК-полімераз. Також перешкодою можуть стати утворені матричною ДНК петельні структури.

Метод **секвенування ДНК за Максамом -Гілбертом (хімічний метод)** заключається в наступному. Один з кінців фрагмента ДНК мітять за допомогою 32Р. Препарат міченої ДНК ділять на чотири порції і кожну з них обробляють реагентом, який специфічно модифікує одну з чотирьох основ ДНК.

Пуринові основи модифікують диметилсульфатом, при цьому відбувається метилювання аденінових залишків по азоту в положенні 3, а гуанінових – по азоту в положенні 7. Обробка ДНК соляною кислотою при 0С призводить до вищеплення метиладеніна. Наступна інкубація при 90оС в лужному середовищі викликає розрив сахаро-фосфатного ланцюга ДНК в місцях вищеплення основ. Обробка піперидином призводить до гідролізу зразка по остаткам метилгуаніна.

Піримідинові основи модифікують гідразином. Якщо реакцію вести в безсолевому середовищі, то модифікується як цитозин, так і тимідин; якщо обробку вести в присутності 2М натрію хлориду, то модифікується лише цитозин. Розщеплення ланцюгу ДНК на фрагменти і в цьому випадку здійснюється піперидином.

В результаті отримують набір мічених фрагментів, довжина яких визначається відстанню від зруйнованої основи до кінця молекули.

Наприклад, якщо залишки G находяться на відстані 2, 6, 11 та 16 нуклеотидів від міченого кінця, то обробка даного ланцюга реагентами, що руйнують G, призведе до утворення мічених фрагментів довжиною 2, 6, 11 та 16 нуклеотидів (утворені при цьому фрагменти, довжиною 3 та 4 нуклеотида не будуть міченими).

Набори мічених фрагментів, що утворюються в кожній з чотирьох реакцій, піддають електрофорезу у сусідніх доріжках поліакриламідного гелю, при цьому відбувається їх розділення у відповідності до розмірів. Потім здійснюють радіавтографію гелю. Набір полос, що реєструються на рентгенівській плівці, «читають», визначаючи таким чином нуклеотидну послідовність ДНК.

Метод **піросеквенування** предложен Ронагі в середині 1990-х років. Матрицею для секвенування служить одноланцюгова ДНК, яка інкубується з праймером та з ферментами: ДНК-полімеразою, АТФ-сульфурілазою, люциферазою та апіразою. В якості субстратів в реакцію додають аденозин-5’-фосфосульфат (АФС) и люциферин. Для ініціації реакції синтезу другого ланцюга до суміші вводять послідовно кожний дНТФ. Замість дАТФ додають дАТФαS, який не є субстратом для люциферази. У випадку невключення дНТФ до ростучого ланцюга він розщеплюється апіразою. Якщо дНТФ включаеться до ростучого ланцюга, при цьому вивільняється пірофосфат (Р2О74-). АТФ-сульфурилаза кількісно конвертує пірофосфат в присутності АФС в АТФ, який служить джерелом енергії для реакції, яку здійснює люцифераза: перетворення люциферина в оксилюцеферин з утворенням видимого світла, кількість якого пропорційна кількості АТФ. Світло уловлюється камерою, яка з’єднана з комп’ютером. Потім настає черговий цикл – вводиться наступний дНТФ.

Піросеквенування не потребує електрофорезу або іншої процедури розділення фрагментів і, таким чином, працює швидше у 100 разів, ніж методи обриву ланцюга. Однак за один експеримент можна отримати лише декілька десятків п.н. (це дуже актуально, наприклад, при типізації SNP).

До середини 2000-х рр. розробили методи секвенування геномів другого покоління. Для такого секвенування немає необхідності створювати бібліотеку ДНК в клітинах Е. *соli*. Вартість нових методів в кілька разів нижче вартості традиційного методу автоматичного секвенування, а продуктивність набагато вище.

Сьогодні на ринку існує три комерційних системи другого покоління для секвенування ДНК: 454 (Roche Applied Sciences), Solexa (Illumina) і SOLiD (Applied Biosystems). У 2006 р. центром секвенування **454 Life Sciences** запропонований метод масштабного піросеквенування, який дозволяє прочитати більше 4 млн. пар основ ДНК за 10 годин. На першому етапі геномна ДНК розщеплюється на фрагменти довжиною 300-800 п.н., кожен з яких потім зшивається з двома короткими адаптерними послідовностями. Адаптери служать праймерами для ПЛР і секвенування. Адаптер В містить біотин на 5'-кінці для мобілізації бібліотеки ДНК на мікронамистинах, покритих стрептавідином. До кожної намистині прикріплюється по одній молекулі ДНК. Після цього ДНК ампліфіцирують на бусинах методом емульсійної ПЛР - у водних міхурцях інвертної емульсії (вода в олії), що дозволяє уникнути контамінації. В результаті кожна намистина є «клоном», що містить мільйони копій фрагмента. Потім намистини переносяться в лунки розміром 29 мкм оптоволоконного чіпа, на якому відбувається піросеквенування фрагментів ДНК з адаптерної послідовності А. Довжина прочитаних таким способом фрагментів становить 250 -400 п.н.

Методи **Solexa і SOLiD** розраховані на прочитання коротких фрагментів довжиною до 35 пар основ, однак їх вартість на порядок нижче секвенування методом 454. У приладі **SOLiD** реалізована технологія паралельного секвенування на основі лігазної ланцюгової реакції. При цьому підготовку фрагментів ДНК для секвенування здійснюють принципово так само, як в методі 454, але на останньому етапі замість піросеквенування застосовують лігування. Для лігування використовують набір з шістнадцяти зондів. Кожен зонд складається з восьми нуклеотидів: на 3-кінці знаходиться комбінація з 2-х нуклеотидів (24 = 16), наступні 6 нуклеотидів вироджені (тобто можуть спаровуватися з будь-якими нуклеотидами), між 5-м і 6-м нуклеотидами знаходиться сайт, за яким відщеплюється 5-кінець з пришитим до нього флуорофором. Флуорофор відщеплюється тільки в тому випадку, якщо відбулося лігування. Для мічення використовується 4 флуорофора і система флуоресцентного кодування динуклеотидів: один флуорофор відповідає 4 дінуклеотидам. У кожному циклі секвенування читаються два через три нуклеотиди. В кожному наступному раунді використовується зонд, що відпалюється на 1 нуклеотид ближче до мікронамистини. Після п'яти раундів лігувания видається результат секвенування.

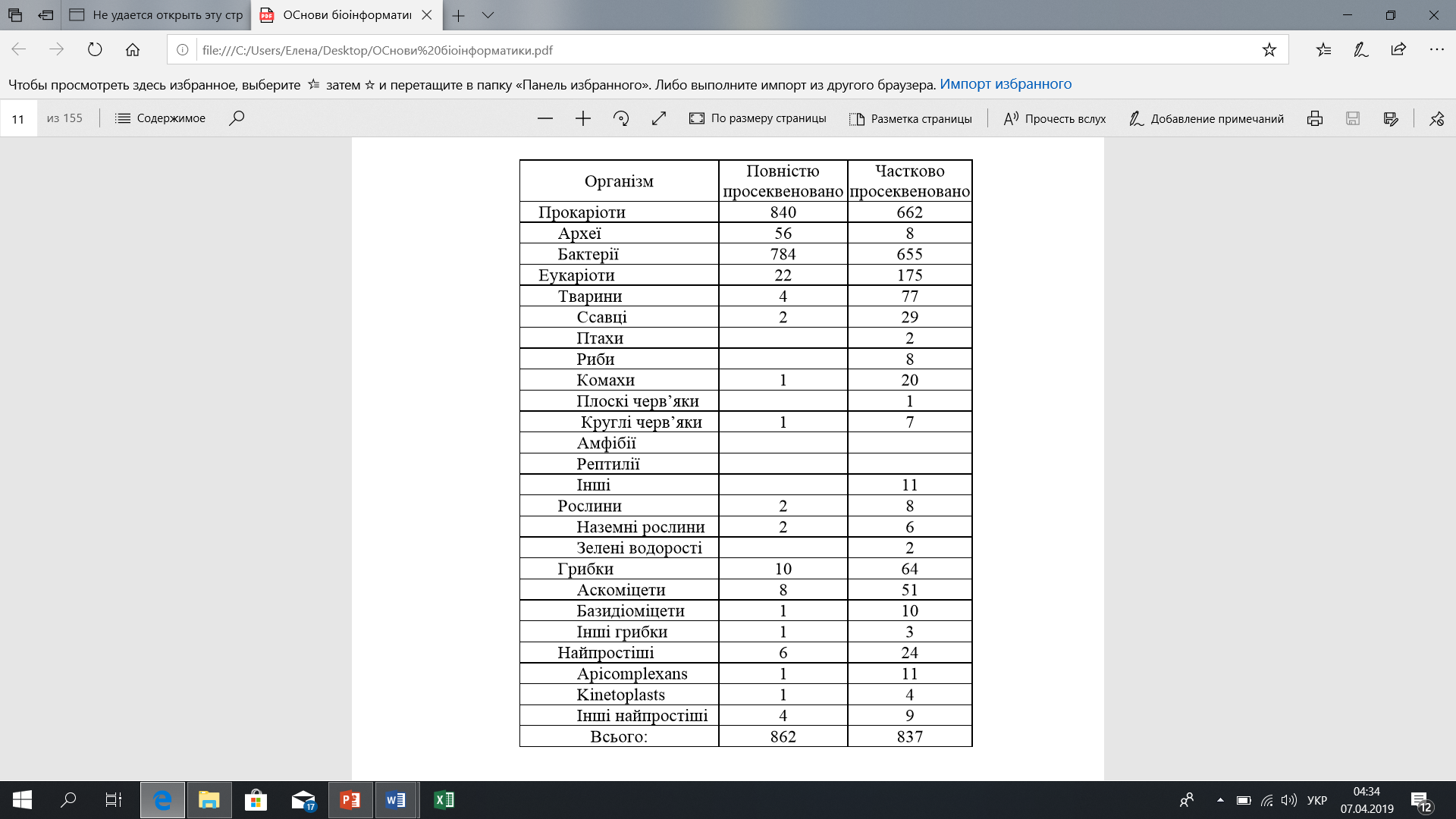
На відміну від методів 454 і SOLiD, в методі **Solexa** ПЛР проходить безпосередньо на носії, вкритому оліонуклеотидами, комплементарними послідовностями адаптерних праймерів. Для секвенування додають все чотири дНТФ, мічені різними флуоресцентними мітками. Від того нуклеотиду, який включився, надходить флуоресцентний сигнал, який і вловлюється детектором.

В даний час з'явилися і активно розвиваються платформи ДНК- секвенування **"третього покоління",** представлені приладами: Helicos Heliscope (Helicos BioScience Corporation), PacBio RS SMRT (Pacific Biosciences) і MinION (Oxford Nanopore). Третє покоління відрізняється від другого тим, що первинна ампліфікація (ПЛР) ДНК більше не потрібна. Досліджувана ДНК секвенується безпосередньо на рівні однієї молекули з використанням спеціально створених полімераз. Перевагою цього підходу є те, що вдається уникнути проблем, пов'язаних з накопиченням помилок, що виникають при ампліфікації ДНК. Наприклад, нанопорне секвенування. Негативно заряджена молекула ДНК приводиться в рух за допомогою прикладеного електричного поля, щоб пройти через нанопори. Процес проходу молекули ДНК через нанопору визначається як транслокація, а відповідний інтервал часу визначається, як час транслокації. В процесі транслокації ДНК, блокада струму значно менша, ніж струм нанопори, так як діелектрична проникність молекули ДНК значно менша, ніж у води. Цей струм визначається як струм блокади. Визначення різниці в блокаді струмів дозволяє розрізняти ДНК на одному рівні з базовою нанопорою. Дослідження блокади струмів є досить популярне в останні роки.

Геномний проект не закінчується з завершення секвенування, а тільки починається з нього. Далі має бути кропітка робота з розшифрування генома: виявлення кодують рамок - генів, складання баз даних, робота з дослідження структури і функції окремих генів і цілого генома. З розвитком геномних проектів з'явилася нова область генетики - геноміка, яка займається вивченням генома в цілому.

**Секвенування геномів.**

У 1978 р Сенгер з співавт. секвенували послідовність генома бактеріофага φХ174, розмір якого перевищує 5 т.п.п. Потім стали розшифровувати геноми інших вірусів, мітохондріальні геноми, геноми хлоропластів. Розробка методу автоматичного секвенування істотно прискорила процес секвенування більших геномів. Розшифровка геномів зажадала об'єднання зусиль різних лабораторій. Виникли так звані геномні проекти. Перший геном бактерії (1,8 млн. пар основ) - *Hemophilus influenza* - секвенували в 1995 р, а в 1996 р. розшифрували перший «еукаріотичний» геном - геном дріжджів S. *cerevesiae* (більше 12 млн.п.н.).

На сьогоднішній день геномні проекти охоплюють понад 2500 геномів вірусів, близько 1600 геномів бактерій і більше 300 геномів еукаріот. Повністю розшифровані нуклеотидні послідовності геномів основних генетичних об’єктів: плодової мушки, нематоди С. *е1еgаns*, людини і хатньої миші.

У світі існує близько 700 центрів секвенування - наукових лабораторій, в яких проводиться секвенування геномів: в США, Великобританії, Німеччини, Франції, Італії, Японії, Китаї та деяких інших країнах. Розвиваються великі міжнародні проекти.

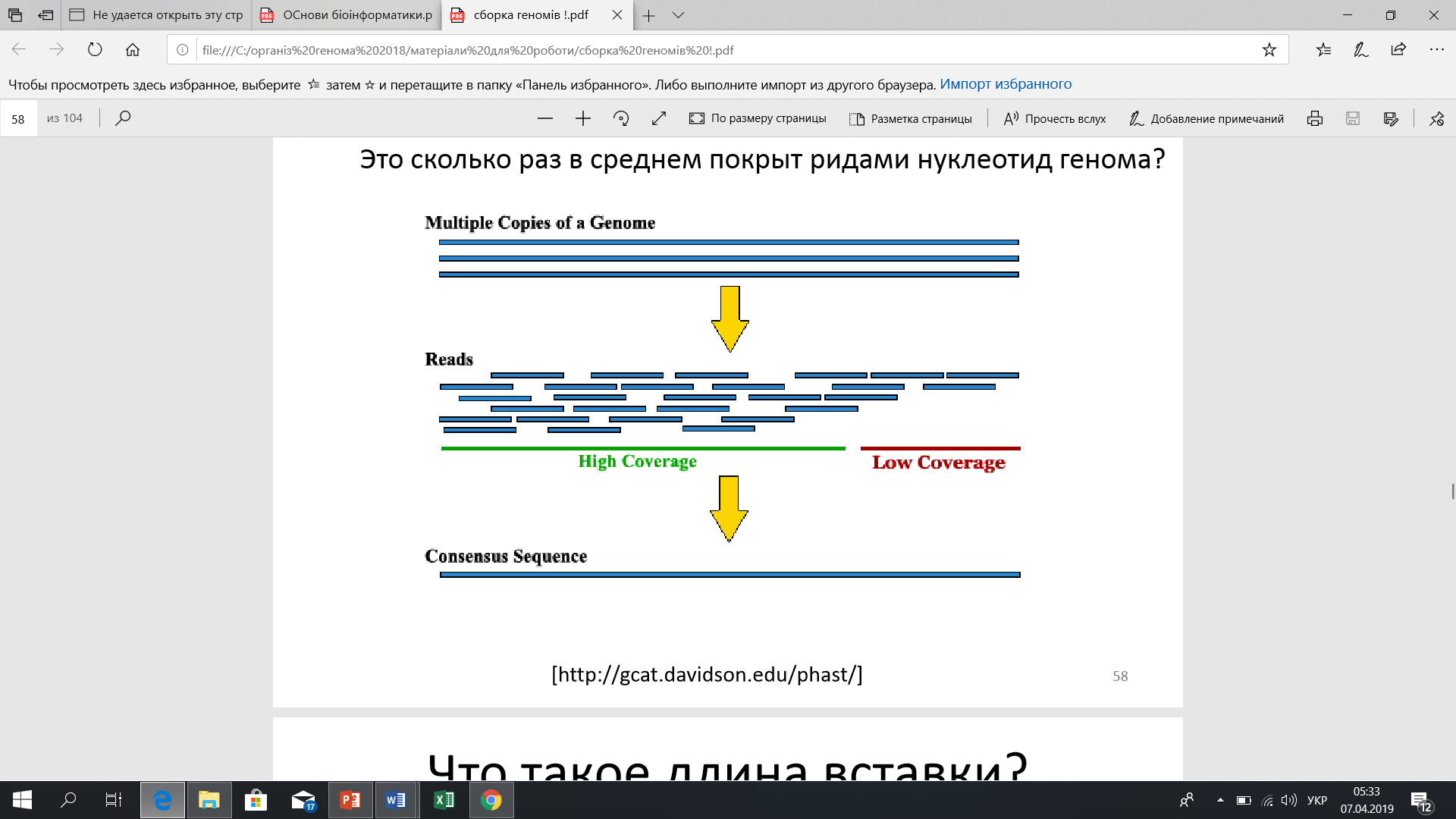
Процес секвенування генома можна умовно розділити на три основних етапи: отримання геномної бібліотеки, секвенування випадкових клонів і об'єднання секвенованих послідовностей.

Геном бактерії Н. *influenza* секвенували методом **"shotgun"** (метод дробовика). Для цього геномну ДНК фрагментували за допомогою ультразвуку. Потім після поділу отриманих фрагментів гель-електрофорезом виділили фракцію фрагментів розміром 1,6-2 т.п.н., яку використовували для клонування в плазмідних векторах. В результаті отримано 19687 клонів і виконано 28643 експериментів з секвенування клонованих фрагментів. За допомогою праймерів, локалізованих всередині векторів, були секвеновані кінці інсерцій. Зіставлення фрагментів, що перекриваються, дозволило реконструювати довгі фрагменти геному **- контигі.** В результаті отримано 140 неперекривающихся контигів. Для заповнення прогалин використовували техніку "primer walking": для секвенування неклонованої послідовностей в якості праймерів використовувалися кінці клонованих послідовностей.

Суть секвенування еукаріотичного генома зводиться до розбиття його на невеликі частини, кожну з яких можна секвенувати за методом "shotgun". Так отримують **читання** або **ріди.**

**Читання (рід) – послідовність ДНК, яка відповідає одному фрагменту бібліотеки.**

Важливим на цьому етапі є **Покриття (глибина секвенування)** – кількість читань (рідів), що містять той або інший нуклеотид генома. Вважають, що для досягнення точності секвенування кожну ділянку генома потрібно секвенувати як мінімум 4 рази.



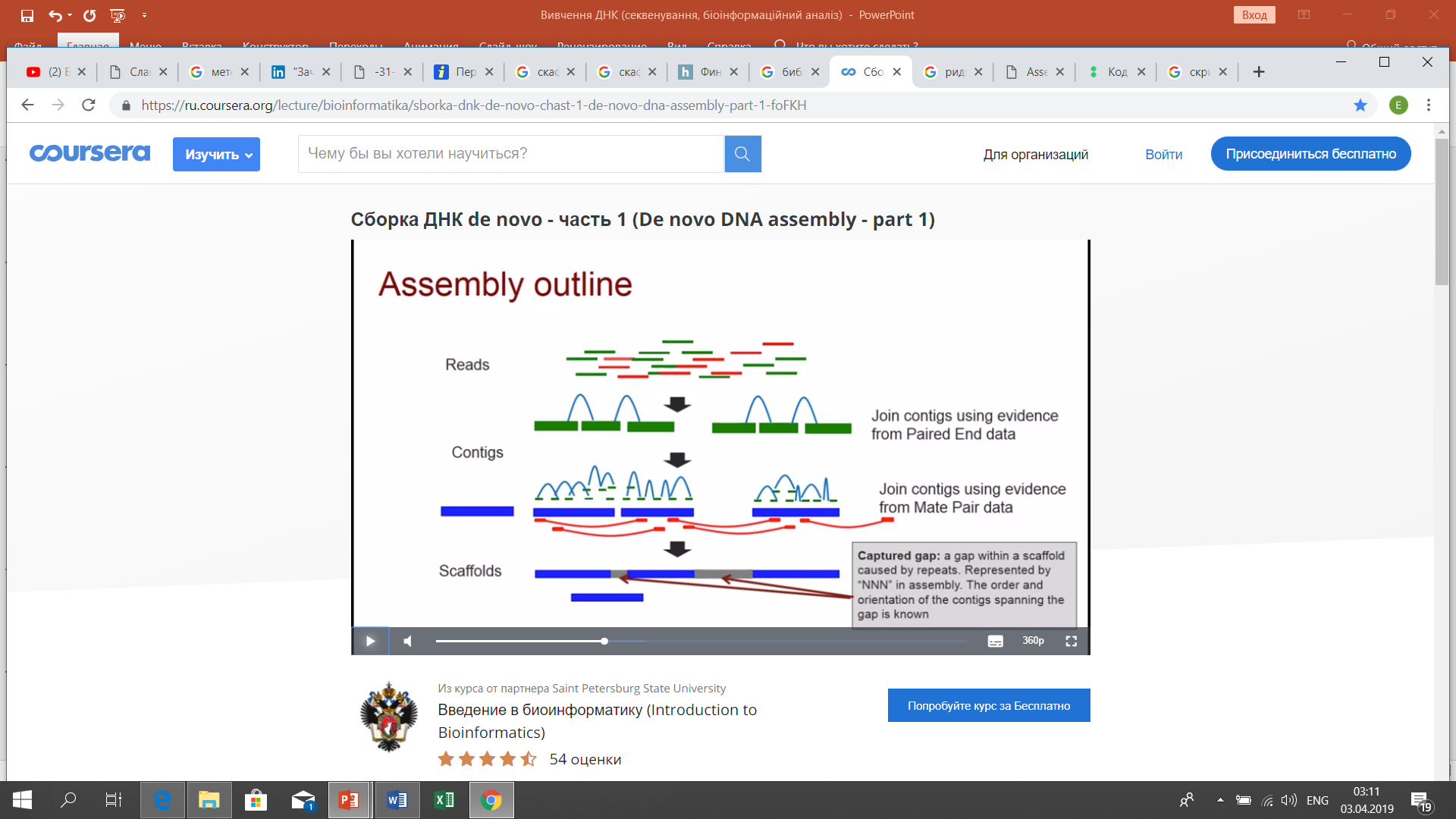
Наступний етап - **Зборка** – відновлення безперервної послідовності ділянки генома шляхом накладення рідів, що перекриваються. Зіставлення фрагментів, що перекриваються, дозволяє реконструювати довгі фрагменти геному **- контигі.**

**Contigs (контігі) - безперервні фрагменти ДНК, які читаються однозначно.**

Самий трудомісткий і останній етап проекту секвенування - об'єднання контігів, заповнення прогалин. Для поетапного збирання геномів використовуються **бібліотеки ДНК різної довжини,** як отримують шляхом розрізання геному на фрагменти та секвенування кінцевих послідовностей. Співставлення цих послідовностей з рідами або контигами дозволяє розташувати їх у більш довгі сіквенси у вірній послідовності. Набір контигів об'єднується в **скаффолди.**

**Скаффолд** - **це серія контигів, розташованих в правильному порядку, але необов'язково сполучених в одну безперервну послідовність.**

Це проміжна неповна структура послідовності, що секвенується, яка допомагає складанню її повної версії. Скаффолди фактично завжди розділені фізичними пробілами – **гепами** (gaps), які важко заповнити, якщо несквеновані послідовності не представлені в бібліотеці.



Серйозна проблема при секвенування еукаріотичних геномів - це величезна кількість повторів. Завжди існує ймовірність, що частина повторів не потрапить в сіквенс або станеться помилкове об'єднання контигів з втратою фрагмента хромосоми між повторами.

Референсний геном – секвенований геном організма того ж виду.

Ресеквенування – секвенування генома виду, для якого існує референсний геном.

Секвенування de novo – секвенування генома вида, для якого не має референсного генома.

**Біоінформатика — молекулярна біологія *in silico.***

Біоінформатика – область науки, що розробляє і застосовує технології інформатики для аналізу, систематизації молекулярнобіологічних даних, використовує фізико-математичні методи для моделювання процесів, що відбуваються на молекулярному рівні з метою виявлення структур, функцій та взаємодії макромолекул (ДНК, РНК, білків).

У вузькому технологічному змісті: під біоінформатикою мають на увазі програмні засоби, спеціально призначені для аналізу «біологічних текстів».

**Основні області досліджень в біоінформатиці:**

1.Аналіз генетичних послідовностей.

Послідовність генома не є самоціль. Головна задача – зрозуміти, що саме містить геном і як геном експресується.

Здійснюється розробка алгоритму обробки великого масиву даних секвенування, збирання геному та автоматичний пошук генів та регуляторних послідовностей в геномі за їх специфічними особливостями.

2. Функціональна анотація геномів (розмітка геномів). Це процес маркування генів та інших об’єктів у послідовності ДНК.

3. Передбачення функцій генів. Практична біоінформатика, спираючись на порівняння з теми генами, функції яких вже визначені, може передбачити функції генів та умови їх регуляції.

Також можлива оцінка ролі окремих ділянок послідовності в функціонуванні білка або ділянок, які відповідають за вирішення різного роду біологічних завдань.

4. Філогенетичні дослідження.

Порівнюючі геноми можно з’ясувати як відбувалась еволюція геномів (змістовна біоінформатика – це наука про молекулярну еволюцію). Порівняння геномів дозволяє вивчати комплексні еволюційні події, зміни ДНК та створювати компьютерні моделі для передбачення подій у часі.

Хоча кожен геном піддається власним видоспецифічним перебудовам, завжди є ділянки, які однаково побудовані у видів, зв’язаних спільною еволюційною історією.

5. Оцінка біологічного різноманіття.

Біологічне різноманіття системи може бути визначено як повна генетична сукупність певного середовища, яка складається з усіх видів. Метагенетика вивчає сукупний геном, наприклад, кишкової мікрофлори чи калюжі.

6. Побудова комп’ютерних молекулярних моделей просторової структури білку або РНК за схожою послідовністю та дослідження механізмів функціонування цієї моделі.

Як продовження – наприклад комп’ютерне конструювання ліків (моделювання у просторі взаємодії різних хімічних сполук з білком-мішенню).

7. Складання генетичних мереж.

Об’єднання біології, хімії, фізики, математики та інформатики дозволяє різнобічно описувати біологічну систему - здійснювати метаболічну реконструкцію, моделювання розвитку та передбачення властивостей організму за його геномом.

З 1982 р. по молекулярній біології, зокрема, нуклеотидним та амінокислотним послідовностям, створено декілька тисяч БД. Серед них є електронні банки даних як загального призначення, так і вузькоспеціалізовані, а також БД наукових робіт з біології та медицини (Medline/PubMed та інші).

Однією з найвідоміших міжнародних організацій з тих, що створюють інструменти для аналізу інформації та здійснюють нагляд за наповненням баз даних біоінформаційного напряму, є **Міжнародна система баз даних нуклеотидних послідовностей** (International Nucleotide Sequence Database Collaboration (**INSDC**) –http://www.ncbi.nlm.nih.gov/collab), яка об’єднує три установи, нуклеотидні БД яких мають різний набір сервісів:

**1 - NCBI** – National Center for Biotechnology Information, USA (БД GenBank містить більше 8×107 записів), http://www.ncbi.nlm.nih.gov/ **, 2 - EBI** – European Bioinformatics Institute, United Kingdom (БД EMBL містить біля 8×107 нуклеотидних послідовностей, а TrEMBL – більше 5×106 автоматично трансльованих з даних EMBL амінокислотних послідовностей), http://www.ebi.ac.uk/ , 3 - **NIG** – National Institute of Genetics, Japan (БД DDBJ містить біля 8×107 записів), http://www.ddbj.nig.ac.jp/ .