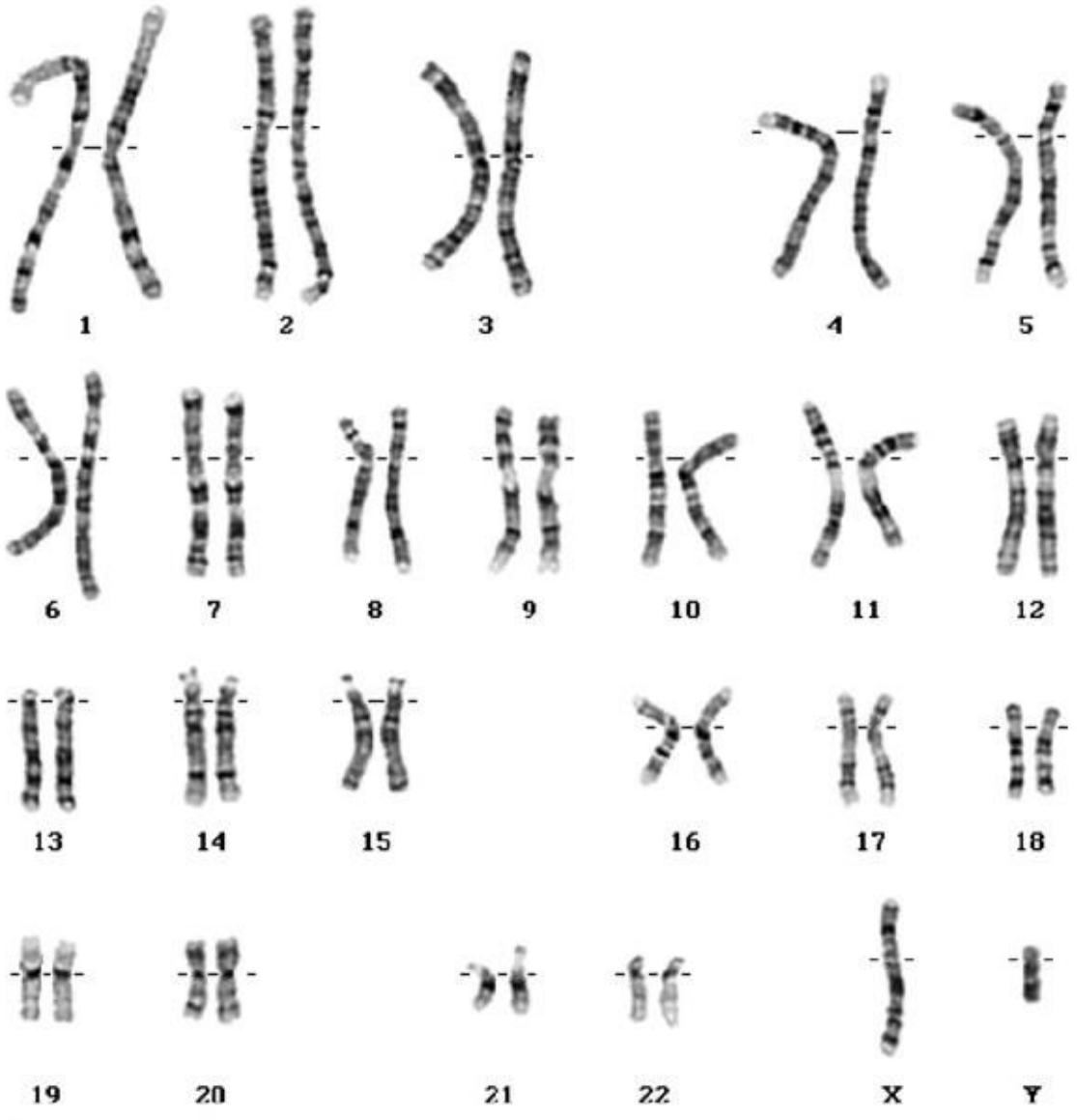
Instytut Informatyki i Matematyki Komputerowej UJ,
opracowanie: mgr Ewa Matczyńska, dr Jacek Śmietański

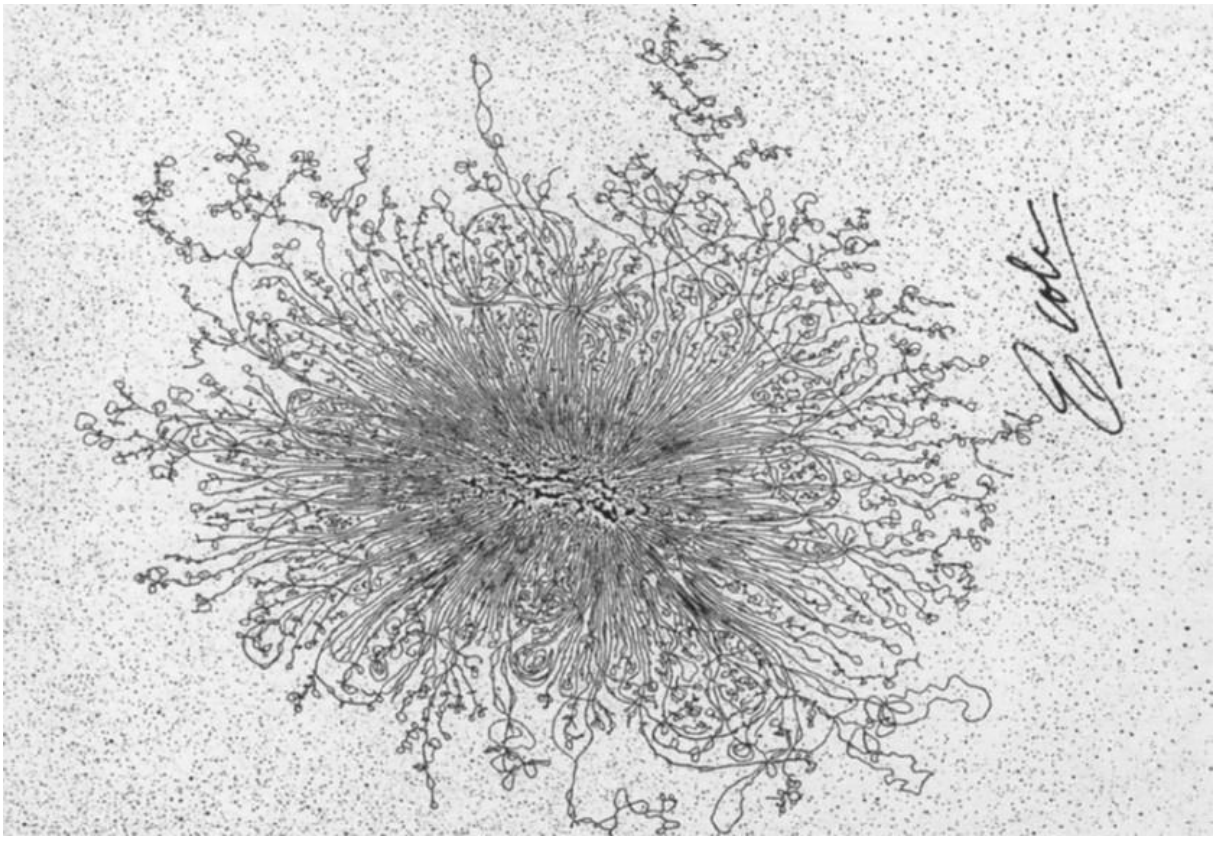
Sekwencjonowanie, przewidywanie genów

1. Technologie sekwencjonowania

Genomem nazywamy sekwencję DNA zawartą w pojedynczym zestawie chromosomów. U wirusów zdarza się czasem, że genom jest zapisywany na nośniku RNA (np. genom wirusa HIV). Mówi się też o genomie mitochondrialnym, zawartym w mitochondrium (fabryce energii organizmu), który jest dziedziczony w całości po matce.



Rysunek 1. Chromosomy człowieka (źródło: http://www.ajnr.org/content/28/3/406/F1.expansion.html)



Rysunek 2: Genom bakterii E.coli - jedna, bardzo poskręcana pętla.

Sekwencjonowanie to proces odczytu informacji genetycznej danego organizmu, czyli odczytanie sekwencji nici DNA, bądź RNA.

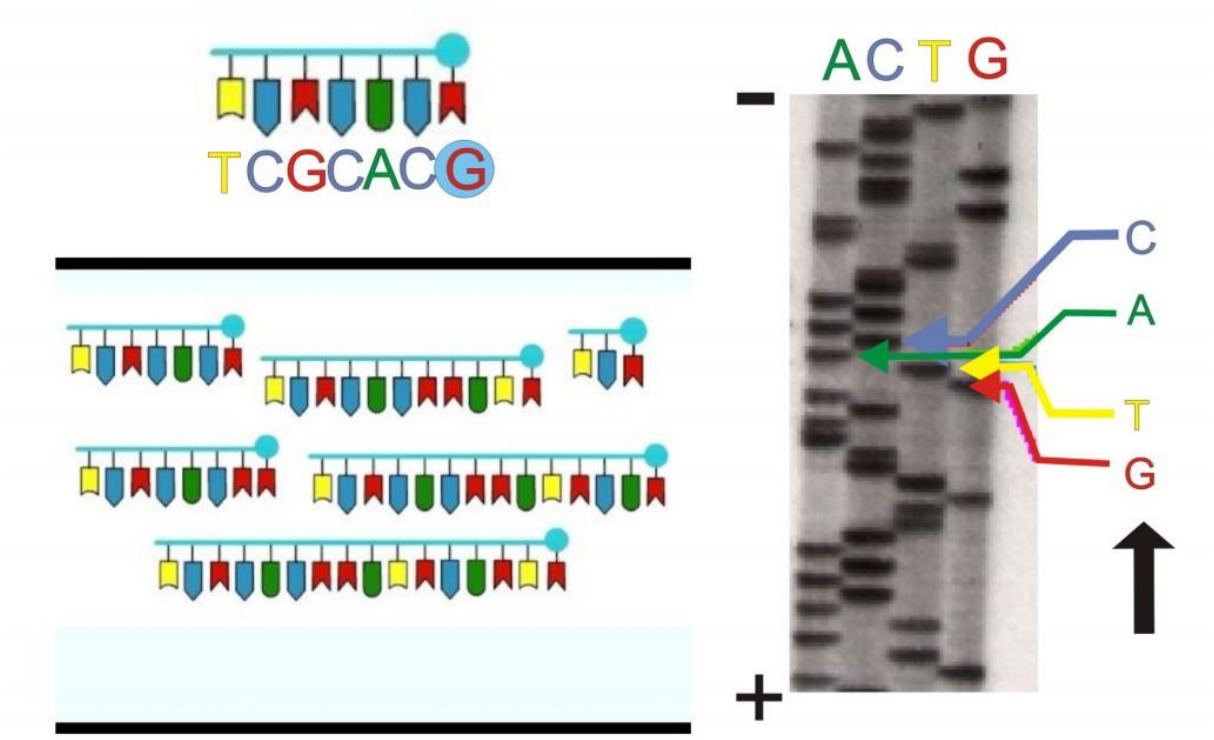
Sanger sequencing

W 1975 Frederick Sanger opracował pierwszą metodę, która pozwalała dość sprawnie odczytywać relatywnie długie fragmenty sekwencji (do kilkuset zasad). Metoda jest używana do dzisiaj, przy czym odczytywać można nieco dłuższe fragmenty długości 1-2 Kbp.

Podstawą metody są specjalne nukleotydy, które nie pozwalają na dalsze dobudowywanie nici komplementarnej. W ten sposób jeśli w danej próbówce mieliśmy naszą nić w wielu kopiach i określony nukleotyd - mamy pewność jaka zasada jest na końcu nici. Następnie sekwencje z każdej z probówek rozdzielamy w żelu przykładając napięcie. Ponieważ nić DNA ma ładunek ujemny, będzie przesuwała się w kierunku dodatnim. Nici dłuższe są cięższe i będą poruszać się wolniej, co umożliwi dokładne rozdzielenie ze względu na długość. Teraz możemy już odczytać sekwencję.

Filmik opisujący metodę Sangera: http://www.youtube.com/watch?v=oYpllbI0qF8

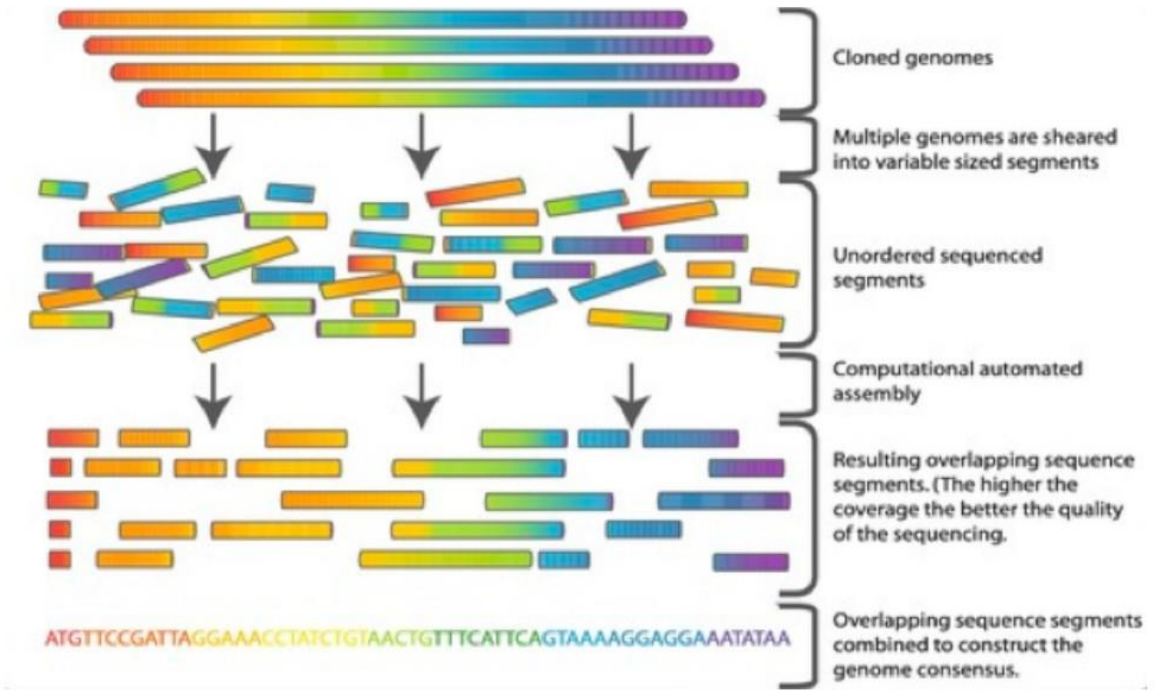
Tą metodą Frederick Sanger odczytał pierwszy genom w roku 1977. Był to bakteriofag ΦX174, którego długość genomu wynosiła 5386 bp. Sekwencję odczytywano fragment po fragmencie, łącząc je w całość, dlatego cały proces był bardzo czasochłonny i wymagał dużego udziału człowieka. Zaczęto sekwencjonować różne organizmy, myśląc powoli o sekwencjonowaniu człowieka, którego genom ma długość ok. 3.2 Gbp.



Rysunek 3: Schemat metody Sanger'a.

Shotgun sequencing

Aby przyspieszyć i zrównoleglić proces sekwencjonowania wprowadzono metodę *shotgun sequencing*, która dzieli genom na losowe fragmenty, a następnie zsekwencjonowane odczyty składa za pomocą komputera - dokładnie za pomocą programów nazywanych assemblerami.



Rysunek 4: Odczytywanie genomu za pomocą shotgun sequencing.

2. Human Genome Project

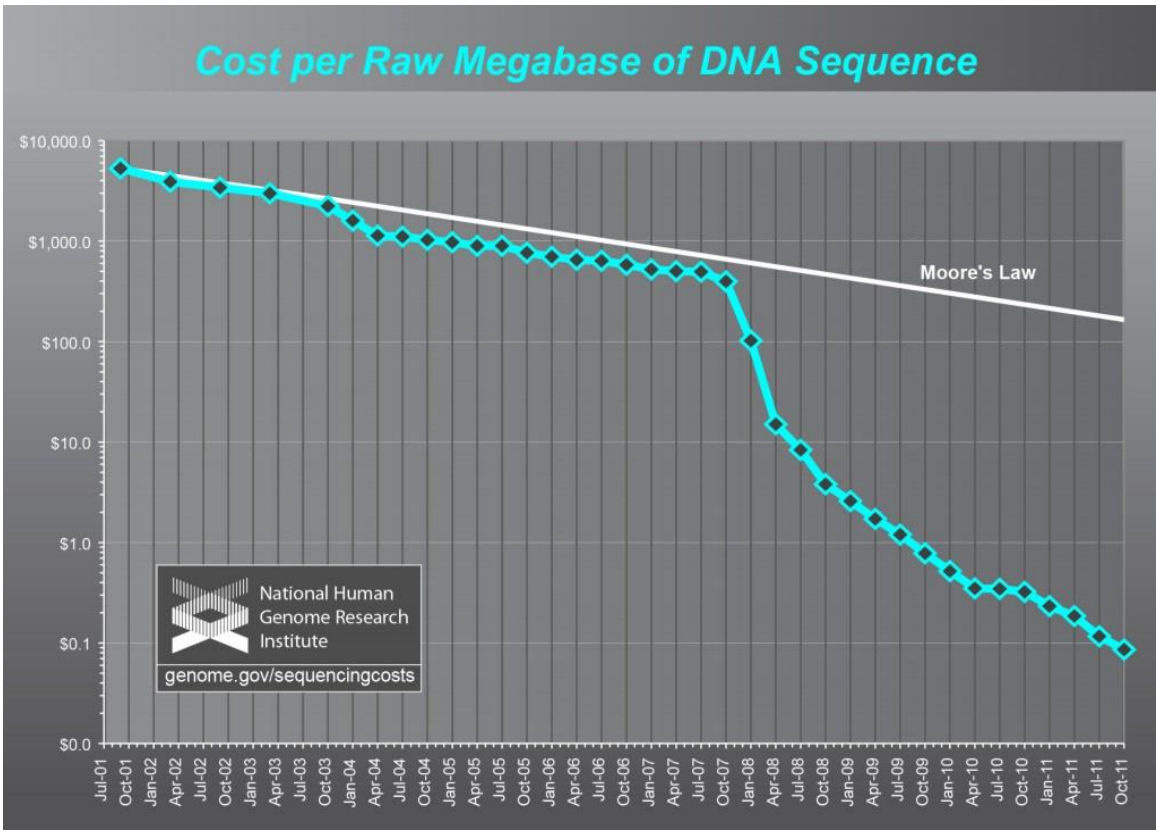
W roku 1990 wystartował *Human Genome Project*, którego celem było odczytanie pełnej sekwencji ludzkiego genomu. Projekt został oficjalnie zakończony w 2003 roku, wcześniej niż planowano, gdyż rozwój technologii, w szczególności metoda shotgun, pozwoliła na przyspieszenie prac nad ludzkim genomem.

W tej chwili mamy dostępne już pełne sekwencje genomów wielu organizmów, w tym wielu ssaków. Oczywiście można je znaleźć w bazach NCBI.

Zadanie 1.
Wejdź na NCBI i znajdź genomy różnych organizmów.
(Resources -> Genomes and Maps -> Genome -> Using Genome -> Download / FTP)
Zobacz jakie genomy są dostępne, znajdź genom człowieka i ściągnij chromosom Y z genomu referencyjnego w formacie fasta.

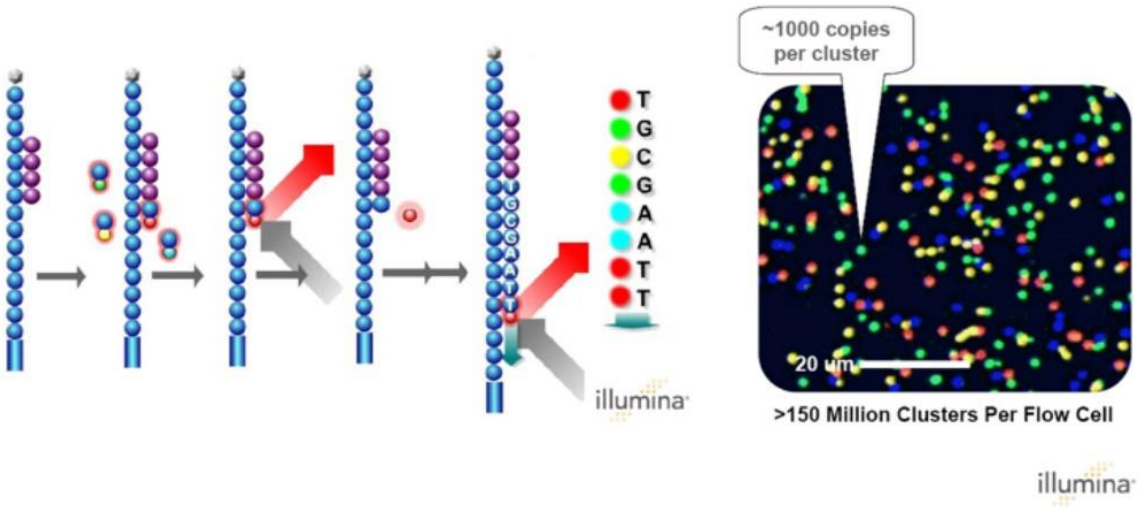
3. Next generation sequencing

Obecnie sekwencjonowanie przebiega bardzo szybko dzięki rozwiniętym technologiom nowej generacji. Zapewniają one dużo mniejszy koszt niż metody używane jeszcze przy HGP.



Rysunek 5: Spadek kosztu sekwencjonowania (koszt 1 Mbp).

Przykładowo, technologia firmy Illumina nieco przypomina podejście Sangera do sekwencjonowania, przy czym proces jest masowo równoległy, odczyty mają ok. 100 bp i może być ich do 3 miliardów z jednego eksperymentu.



Rysunek 6: Przykład technologii sekwencjonowania nowej generacji. W każdym cyklu maszyny dołączany jest jeden nukleotyd, który wzbudzony laserem emituje światło o określonej częstotliwości. Punkty świetlne są skanowane i rozpoczyna się następny cykl.

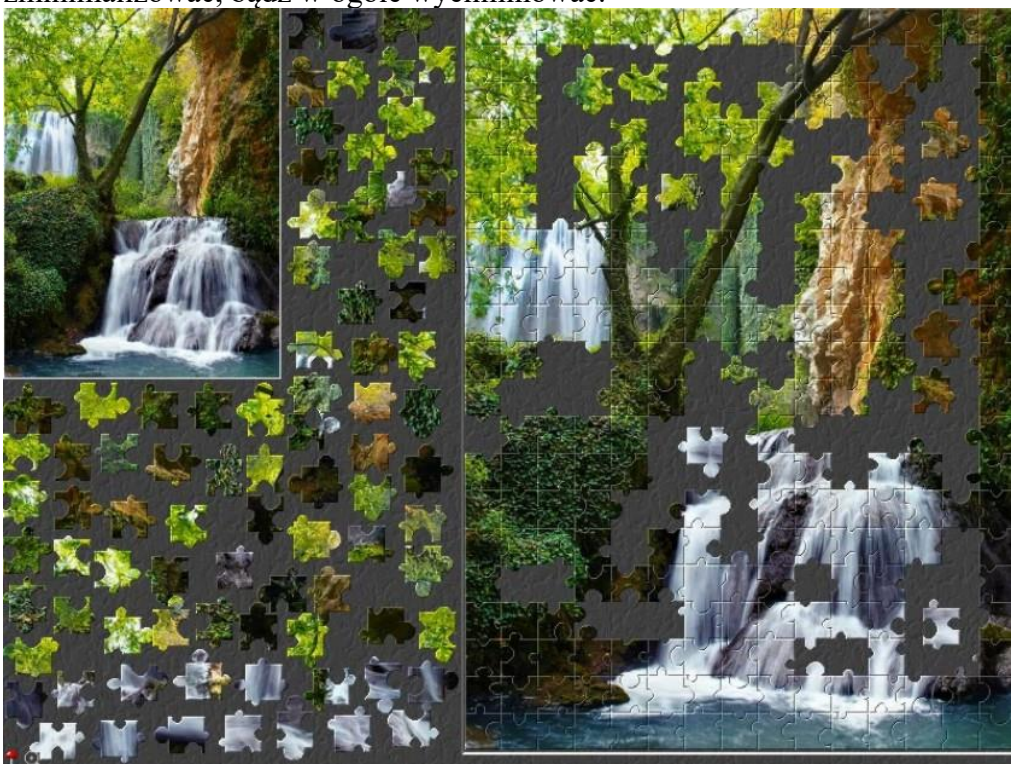
Kolejnym krokiem milowym, zrealizowanym przy użyciu technologii nowej generacji był program zsekwencjonowania 1000 genomów ludzi z różnych regionów świata. Celem projektu było poznanie różnorodnosci genetycznej wśród ludzi różnych ras, żyjących w różnych obszarach klimatycznych. Strona projektu: **www.internationalgenome.org**

4. Genome assembly

Zagadnienie *genome assembly* czyli składania genomu z odczytów sekwencji możemy podzielić na dwa podzagadnienia:
- składanie z wykorzystaniem genomu referencyjnego (o ile jest dostępny) - sprowadza się głównie do używania wydajnych algorytmów dopasowania sekwencji;
- składanie bez genomu referencyjnego - *de novo assembly* - składanie genomu tylko na podstawie odczytów sekwencji, bez dodatkowych informacji o genomie

Terminologia:
- sekwencje, które zostały odczytane nazywamy odczytami (*reads*);
- dzięki assemblerom odczyty możemy połączyć w ciągłe fragmenty - tzw. kontigi (*contigs*);
- STS (*sequence tag sites*) – charakterystyczne miejsca w sekwencji, dzięki, którym można eksperymentalnie zlokalizować dany fragment na chromosomie z dużą dokładnością;
- pokrycie sekwencji - średnia liczba odczytów, w których wystąpił nukleotyd z danego miejsca w sekwencji. Oczywiście im większe pokrycie naszej sekwencji tym dłuższe kontigi uzyskamy. Idealnie byłoby, gdyby assembler zwracał na wyjściu np. cały chromosom jako jeden kontig, jednak tak się z reguły nie dzieje ze względu na poziom pokrycia oraz różne utrudnienia w sekwencji. Dlatego poprawne uszeregowanie kontigów wymaga walidowań

specjalnych, Ten etap jest eksperymentalny i zwykle bardzo drogi, dlatego należy go zminimalizować, bądź w ogóle wyeliminować.

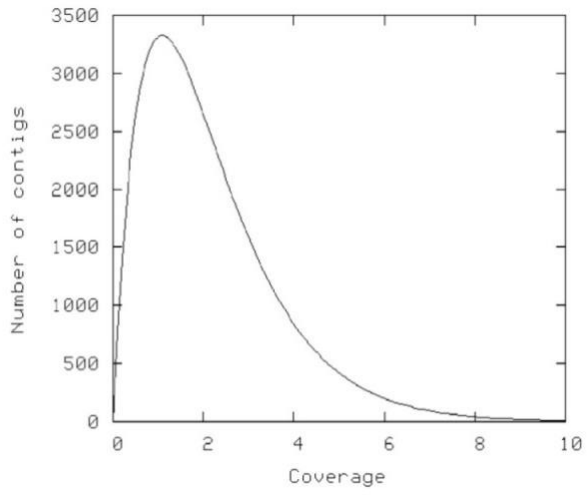


Rysunek 7: Analogia dla sekwencjonowania z genomem referencyjnym i de novo

Zadanie 2.

Otwórz chromosom Y z genomu człowieka i zaobserwuj podział na kontigi.

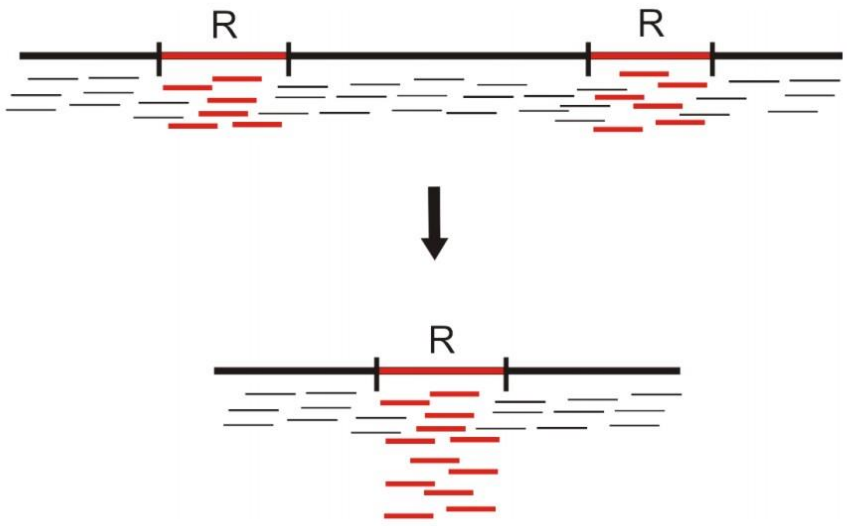
Technologie sekwencjonowania nowej generacji pozwalają na szybsze odczytywanie sekwencji, ale ich długość jest ograniczona do kilkuset zasad, dlatego pokrycie takiej sekwencji musi być duże. Odczyty z nowoczesnych sekwencjonerów są deponowane w *Sequnce read archive* na NCBI, gdzie można znaleźć i pobrać wyniki eksperymentów.



Rysunek 8: Wpływ pokrycia na liczbę kontigów.

Zadanie 3.
Wejdź na NCBI, w wyszukiwarce zaznacz bazę **SRA** *(Sequence Read Archive)* i wpisz w pole wyszukiwania *HIV genome*.
Odnalezione zostaną eksperymenty dotyczące sekwencjonowania wirusa HIV.
Wybierz pierwszy z nich dotyczący analizy pętli V3 - pętla ta koduje białko, które umożliwia zarażenie się wirusem HIV przez człowieka
W tabelce eksperymetów (Run) wybierz jeden z nich, zobacz jak wygląda ten zbiór odczytów (reads).

Sekwencjonowanie bardzo utrudniają tzw. sekwencje repetytywne, czyli powtórzenia, które zdarzają się bardzo często i mogą powodować błędy w składaniu, w szczególności tzw. *repeat collapse*.

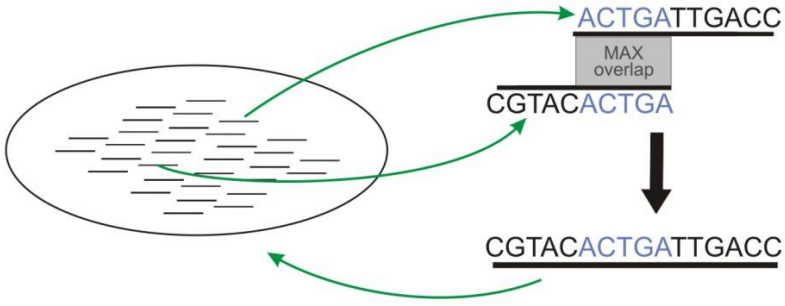


Rysunek 9: Wpływ powtórzeń na efekt składania sekwencji.

5. Algorytmy składania sekwencji

Strategia zachłanna

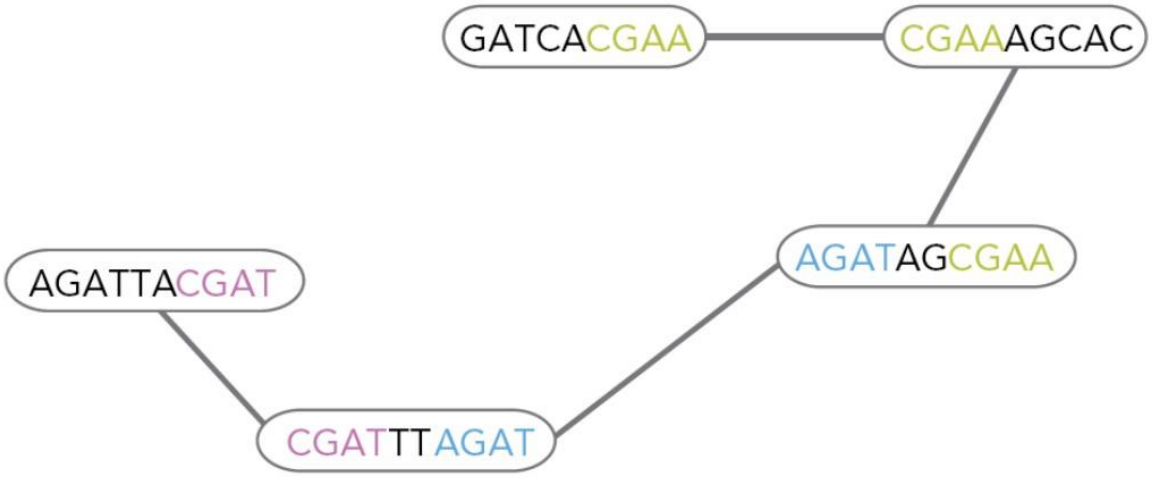
Szukamy dwóch najbardziej nakładających się odczytów w zbiorze, łączymy je i dołączamy do zbioru.



Rysunek 10: Podejście zachłanne.

Overlap-layout-consensus

Tworzymy graf, którego wierzchołkami są odczyty. Krawędź łączy wierzchołki wtedy kiedy odczyty nakładają się. Rozwiązaniem jest ścieżka przechodząca przez wszystkie wierzchołki grafu.

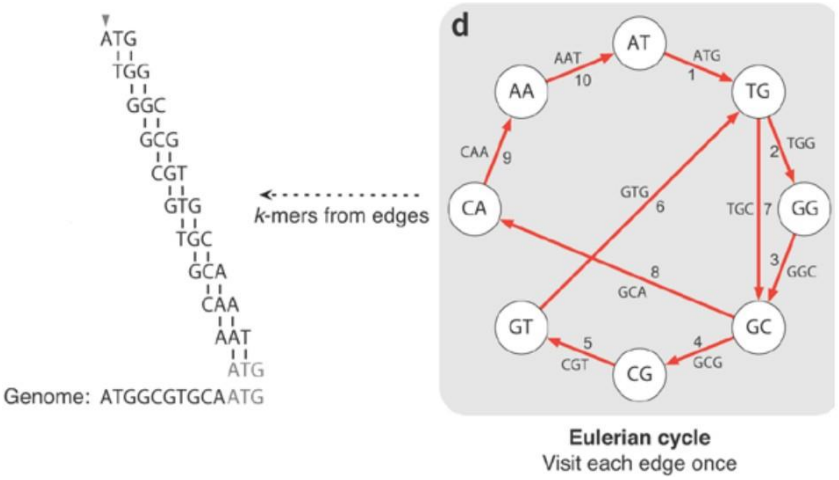


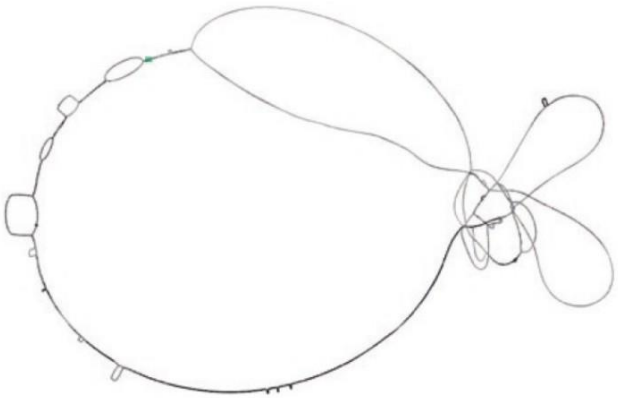
Rysunek 11: Overlap-layout-consensus. Szukamy ścieżki przechodzącej przez wszystkie wierzchołki.

Graf de Bruijn'a

1. Dzielimy odczyty na k-mery:

2. Szukamy ścieżki przechodzącej przez wszystkie krawędzie grafu.





Rysunek 12: De Brujin graph dla genomu pewnej bakterii.

W sieci dostępne są różne asemblery np.:
- Celera Assembler
- Cap3
- Phrap
- Euler
- Velvet

Na użytek ćwiczeń możemy sprawdzić program Cap3, działający według podejścia overlap-layout-consensus. Jest on dostępny online na http://doua.prabi.fr/software/cap3

Zadanie 4.

Na podstawie ciągłego fragmentu chromosomu Y człowieka wygeneruj odczyty ustalonej długości. Sprawdź czy program Cap3 poprawnie złoży je w jeden kontig.

Zadanie 5. (4pkt)

> Rozwiązanie zadanie prześlij mailem do wtorku, **17.12.2019** włącznie, na adres:
> **jacek.smietanski@ii.uj.edu.pl**
> Temat wiadomości proszę opatrzyć przedrostkiem **[Bio] Lab 10**. Rozwiązaniem ma być **tylko jeden plik** – skrypt zgodny z Pythonem w wersji 3.x**,** zawierający wszystkie niezbędne funkcje oraz procedurę wykonawczą. Proszę o nazwanie pliku wg schematu**: Imie.Nazwisko.10.py**.

Napisz:
(a) funkcję implementującą generację losowych odczytów fragmentów sekwencji;
(b) funkcję implementującą prosty algorytm zachłanny składania zadanych na wejściu odczytów w jedną sekwencję.

(c) Przetestuj swoje rozwiązanie na ciągłym fragmencie (ok. 10000 nukleotydów) chromosomu Y człowieka. Czy wynik zgadza się z wyjściową sekwencją?

Ad. a) Z zadanej ciągłej sekwencji wybieramy losowo ciągłe fragmenty o określonej długości.

Parametrami wywołania funkcji powinny być: długość tych fragmentów oraz oczekiwana wielkość pokrycia. Wygenerowane fragmenty należy zapisać do pliku w formacie fasta (każdy fragment traktujemy jako jedną samodzielną sekwencję, wszystkie zapisujemy w jednym pliku).

Ad. b) Dane wejściowe (zbiór fragmentów sekwencji) odczytujemy z (jednego) pliku w formacie fasta. W algorytmie zachłannym rekurencyjnie łączymy ze sobą te odczyty, które mają najdłuższe wspólne sufikso-prefiksy (końcówka jednego odczytu pokrywa się z początkiem drugiego).
W przypadku niejednoznaczności, gdy kilka par odczytów ma jednakowo długą część wspólną, wybieramy losowo jedną z nich.
Program ma zwrócić ciągłe kontigi (w idealnym przypadku jeden) pokrywające wszystkie zadane na wejściu fragmenty sekwencji.

Ad. c) Chromosom Y człowieka możesz znaleźć w bazie „Genome": NCBI → Resources → Genomes and Maps → Genome → Using Genome – Download / FTP.

Na potrzeby testu zastosuj fragmenty długości 200 i pokrycie 5 (parametry domyślne).