# Wprowadzenie do Bioinformatyki
# zajęcia laboratoryjne 2

## Podstawy pozyskiwania oraz analizy danych genomicznych i proteomicznych

## 1 Wykorzystywane narzędzia i bazy danych

1. The Basic Local Alignment Search Tool (BLAST) – program służący do odnajdywania regionów podobieństwa pomiędzy sekwencjami biologicznymi. BLAST można wykorzystać do wnioskowania o funkcjonalnych i ewolucyjnych związkach między sekwencjami, a także do identyfikacji podobnych genów, należących do jednej rodziny. Program występuje w wielu wariantach, spośród, których najczęściej wykorzystywane są:

   - blastn (nukleotyd-nukleotyd) – w oparciu o podaną sekwencję DNA (łańcuch nukleotydów), zwraca najbardziej podobne sekwencje DNA z bazy danych określonej przez użytkownika.
   - blastp (białko-białko) – w oparciu o podaną sekwencję białka (łańcuch aminokwasów), zwraca najbardziej podobne sekwencje białkowe z bazy danych określonej przez użytkownika.
   - blastx (nukleotyd-białko) - porównuje możliwe produkty translacji sekwencji nukleotydowej podanej przez użytkownika (obie nici) z bazą danych sekwencji białek.
   - tblastn (białko-nukleotyd) - porównuje zapytanie w formie sekwencji białkowej ze wszystkimi ramkami odczytu bazy danych sekwencji nukleotydów.

   Program ten można pobrać w formie plików wykonywalnych uruchamianych z poziomu wiersza poleceń lub skorzystać z wersji online pod adresem: https://blast.ncbi.nlm.nih.gov/Blast.cgi

2. NCBI Entrez search engine – system wyszukiwania informacji w bazach danych NCBI (National Center for Biotechnology Information) obejmujących literaturę naukową, bazy danych genomów, genów, białek oraz wielu innych http://www.ncbi.nlm.nih.gov/

3. EMBOSS (European Molecular Biology Open Software Suite) - zestaw narzędzi opracowany na potrzeby biologii molekularnej, obejmujących ponad 150 programów z dziedziny:

   - Dopasowania sekwencji (alignment).
   - Wyszukiwania wzorców sekwencji w bazach danych.
   - Identyfikacji motywów białkowych, w tym analiza domen.
   - Wizualizacji danych.

   Programy łączy zbliżona funkcjonalność, polegająca min na rozpoznawaniu tych samych formatów danych, także ich interfejs tekstowy zbudowany jest w oparciu o podobne zasady. Programy uruchamiane są z poziomu linii poleceń jednak istnieje kilka publicznie dostępnych stron internetowych oferujących graficzny interfejs, np. http://www.bioinformatics.nl/cgi-bin/emboss

4. GenomeNet Database Resources – zbiór narzędzi i baz danych ułatwiających zrozumienie funkcji systemów biologicznych na poziomie komórki, organizmu czy ekosystemu, w oparciu o informacje na poziomie molekularnym, szczególnie te dostarczane przez sekwencjonowanie genomu i inne wysokowydajne technologie eksperymentalne. Baza dostępna jest pod adresem: http://www.genome.jp/

## 2 Zadania do wykonania

U pacjenta przyjętego w centrum epidemiologicznym wykryto infekcję wywołaną przez nieznanego wirusa. W celu identyfikacji zdecydowano się na sekwencjonowanie fragmentu jego materiału genetycznego metodą Sangera. W wyniku sekwencjonowania uzyskano fragment sekwencji nukleotydowej jednego z genów. Celem ćwiczenia jest uzyskanie jak największej ilości informacji na temat wirusa oraz samego genu w oparciu o metody bioinformatyczne. Odpowiedzi do poniższych zadań zapisz w pliku tekstowym do okazania pod koniec zajęć. Część uzyskanych informacji będzie potrzebna do wykonania kolejnych zadań.

Dla podanej sekwencji nukleotydowej wykonaj poniższe zadania (sekwencja, zwana w dalszej części badaną sekwencją):

1. Korzystając z narzędzia NCBI BLAST, należy określić do jakiego wirusa należy uzyskany fragment badanej sekwencji. Dopasowanie należy wykonać względem standardowej bazy danych obejmującej wyłącznie organizmy modelowe (opcja Standard Databases). Należy wybrać wirusa, o największej zgodności sekwencji. Wykorzystaj ID *Accession* znalezionej sekwencji aby przejść do strony GenBanku. Podaj podstawowe informacje dla tej sekwencji (locus, definition, accession, version, organism). Następnie wykorzystaj zakładkę "Related information" aby znaleźć sekwencję genomu referencyjnego (wybierz pierwszy rekord z "RefSeq Genome Sequences") i zapisz jego ID uwzględniając wersję. Jaki jest jego całkowity rozmiar znalezionego genomu referencyjnego oraz ile zawiera genów?

2. Korzystając z narzędzia BLAST dostępnego poprzez stronę NCBI Virus, należy powtórnie wykonać dopasowanie badanej sekwencji jednak tym razem względem wszystkich dostępnych genomów danego wirusa, uzyskanych od różnych pacjentów. W tym celu należy zawęzić wyniki w panelu "Refine Results" przez zaznaczenie opcji "Nucleotide Completness → complete", oraz zawężenie pole "host" do wyników dla człowieka (*Human*). Należy posortować tabelę z wynikami według kolumny "score", która domyślnie jest ukryta (przycisk "select columns" nad prawym górnym rogiem tabeli). W jakim kraju został wyizolowany najstarszy znany przypadek tego wirusa o najwyższej zgodności (score) z badaną sekwencją?

3. Dla przypadków uzyskanych w punkcie 2, o największym podobieństwie sekwencji (score), należy wygenerować drzewo filogenetyczne (przycisk "Build Phylogenetic Tree"). Co pokazuje uzyskane drzewo? Podaj przynajmniej jeden przykładowy węzeł drzewa, gdzie prawdopodobnie doszło do zarażenia się wirusem pomiędzy osobami z różnych krajów.

4. Korzystając z programu BLAST serwisu GenomeNet (wariant BLASTN) należy określić, wewnątrz jakiego genu zidentyfikowanego wirusa położona jest badana sekwencja.

5. Należy pobrać pełną sekwencję nukleotydowa i aminokwasową genu odnalezionego w punkcie 4. Ponadto korzystając z nazwy zidentyfikowanego genu (NCBI-GeneID, Gene name lub innej) należy odnaleźć informacje na jego temat w bazie NCBI EntrezGene (symbol, opis, typ). Korzystając z informacji w sekcji "Genomic regions, transcripts, and products" należy określić jakie inne geny leżą w bezpośrednim sąsiedztwie odnalezionego.

6. Korzystając z dowolnego narzędzia, na przykład GC content calculator należy określić zawartość nukleotydów GC oraz długość obu sekwencji nukleotydowych: podanej przez prowadzącego i uzyskanej w punkcie 5. Jaki procent całkowitej sekwencji tego genu obejmował fragment uzyskany w badaniu metodą Sangera?

7. **[Zadanie dodatkowe dla chętnych]**

   Korzystając z sekwencji aminokwasów z pkt. 5 oraz programu blastp należy odszukać 5 najbardziej podobnych białek na przestrzeni wszystkich dostępnych organizmów w bazie danych RefSeq_protein.