Wprowadzenie do Bioinformatyki
zajęcia laboratoryjne 5

Badanie genomów wirusowych w systemie kompleksowej analizy danych

Bioinformatyka I rok

1 Wprowadzenie

Identyfikacja zmian w strukturze genomu wirusowego jest obecnie jedną z podstawowych metod badania rozprzestrzeniania się infekcji wirusowych. Wraz z upływem czasu wirus akumuluje nowe mutacje co przy założeniu stałego tempa ich powstawania pozwala w przybliżeniu określić jego wiek a także śledzić, jego rozprzestrzenianie, pozwalając na zidentyfikowanie źródła epidemii oraz kierunku jej przemieszczania. Ponadto, identyfikacja nowych mutacji pozwala na kontrolowanie możliwości wykrycia wirusa a także skuteczności opracowanych szczepionek, które zaprojektowane są w oparciu o specyficzne sekwencje białkowe wirusa. Jeśli zmiany w sekwencji aminokwasów będą znaczące to przeciwciała w tego typu terapii mogą okazać się nieskuteczne.

Badania sekwencji genomu wirusowego były kluczowe podczas epidemii Eboli w 2014 [1], obejmującej około 9000 przypadków zarażenia w regionach Gwinei, Sierra Leone i Liberii. Ebola (*Zaire Ebolavirus*) to wirus gorączki krwotocznej, który atakuje wiele kluczowych narządów w organizmie, powodując niekontrolowane krwawienie wewnętrzne. Ebola rozprzestrzenia się poprzez bliski kontakt z płynami ustrojowymi zakażonego pacjenta, takimi jak krew, ślina, mocz i pot, z tego powodu śledzenie epidemii jest bardzo trudne. W ciągu pierwszych kilku tygodni epidemii z 2014 roku naukowcy z Broad Institute zsekwencjonowali genomy wirusa Eboli uzyskane od 78 pacjentów z Sierra Leone, porównując je z sekwencją pacjenta z Gwinei, gdzie wybuchła epidemia. Dane zostały natychmiast umieszczone w publicznej bazie danych dzięki czemu naukowcy na całym świecie mogli wykonywać różnego typu analizy, na przykład oceniając, czy mutacje wirusa mogą wpłynąć na skuteczność eksperymentalnego leku ZMapp, który został użyty podczas tej epidemii. Genom Eboli zbudowany jest z jednoniciowego RNA o ujemnej polarności - ssRNA(-), co oznacza, że wirusowe RNA musi być przepisane przez polimerazę na mRNA przed rozpoczęciem procesu translacji. Genom wirusa Eboli składa się z 7 genów kodujących białka, ich lokalizacja oraz położenie sekwencji kodującej znajduje się w Tabeli 1.

Tablica 1: Położenie pełnej sekwencji mRNA oraz sekwencji kodującej białka (CDS) wirusa Eboli. Współrzędne odpowiadają genomowi KJ660346.2 (C15), w przypadku wirusów uzyskanych od innych pacjentów dokładne położenie sekwencji może być inne.

Symbol	Nazwa	Położenie sekwencji	
		mRNA	CDS
NP	nucleoprotein	56-3026	470-2689
VP35	polymerase complex protein	3032-4407	3129-4151
VP40	matrix protein	4390-5894	4479-5459
GP	virion spike glycoprotein precursor	5900-8305	6039-6923, 6923-8068
VP30	minor nucleoprotein synthesis of viral RNAs	8288-9740	8509-9375
VP24	membrane-associated protein	9885-11518	10345-11100
L	polymerase synthesis of viral RNAs	11501-18282	11581-18219

Ebola, podobnie jak wirus HIV, zawierają w swoim genomie krótkie sekwencje kodujące motyw aminokwasowy "PTAPPEY" odpowiedzialny za przyłączanie białka Tsg101, umożliwiającego opuszczenie przez wirusa zainfekowanych komórek i dalsze rozprzestrzenianie się po organizmie [3]. Podobnie jak inne wirusy Ebola bardzo szybko

mutuje zarówno wewnątrz zarażonego organizmu jak i poza nim. Szacunkowe tempo mutacji to 0.002 substytucji na daną pozycję w roku [2].

2 Wykorzystywane narzędzia i bazy danych

1. **GALAXY** – otwarta platforma internetowa służąca do prowadzenia badań biomedycznych, nie wymagająca wiedzy z zakresu programowania. Galaxy jest systemem zarządzania zasobami w formie danych, oprogramowania oraz mocy obliczeniowej. Ponadto dostarcza interfejsy graficzne dla bardzo wielu powszechnie wykorzystywanych narzędzi bioinformatycznych. Ujednolicony format przesyłania informacji pomiędzy programami pozwala na tworzenie ustandaryzowanych schematów analizy danych w formie przepływu pracy (z ang. workflow). Pozwala także na kontrolowanie wersji użytych baz danych i oprogramowania oraz zachowuje informacje na temat wszystkich parametrów wykorzystywanych podczas analizy. Znacznie ułatwia to tworzenie powtarzalnych analiz danych na potrzeby badań naukowych. Galaxy dostępne jest na zasadzie licencji Open Source i może być zainstalowane na prywatnym komputerze, istnieje jednak kilka publicznie dostępnych serwerów pozwalających na prowadzenie własnych obliczeń z wykorzystaniem nieodpłatnie udostępnionych zasobów. Przykładami takich serwerów są: https://usegalaxy.eu oraz https://usegalaxy.org. Na potrzeby wykonania wszystkich zadań należy utworzyć konto na GALAXY.

2. **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) oraz wizualizacji danych.

3. **ClustalW** – narzędzie do jednoczesnego porównywania wielu sekwencji nukleotydowych. Program umożliwia także tworzenie drzew filogenetycznych pomiędzy sekwencjami pozwalając na określenie hierarchii ewolucyjnej.

3 Zadania do wykonania

Celem ćwiczenia jest analiza podobieństwa pomiędzy sekwencjami określonego genu wirusa Ebola uzyskanymi od pacjentów w Sierra Leone, podczas wybuchu epidemii w 2014 r. w Afryce Zachodniej (dane do pobrania). Zbiór danych obejmuje 7 sekwencji z regionu Sierra Leone oraz jedną z Gwinei (C15) gdzie rozpoczęła się epidemia. Wszystkie sekwencje pochodzą z bazy NCBI. Korzystając z serwera Galaxy, znajdującego się pod adresem https://usegalaxy.org należy wykonać poniższe polecenia. Przed przystąpieniem do zadań rozpakuj plik i załaduj dane na serwer Galaxy. Odpowiedzi do konkretnych zadań zapisz w pliku tekstowym do okazania pod koniec zajęć.

1. Podstawowa analiza sekwencji genomu wirusa Eboli (C15) oraz identyfikacja białka zawierającego specyficzny motyw sekwencyjny (praca z jedną sekwencją).

   (a) Korzystając z narzędzia infoseq należy określić długość sekwencji ssRNA wirusa Eboli z próbki C15. Należy wybrać opcję formatowania danych wyjściowych z wykorzystaniem HTML.

   (b) Korzystać z narzędzia extractseq należy wyciągnąć fragmenty sekwencji RNA próbki C15 w oparciu o współrzędne sekwencji kodującej (CDS) genów, podane w tabeli we wstępie. Dodatkowe regiony można podać oddzielając je przecinkiem (bez spacji), np. “1-10,20-30”. Należy zapisać każdy region jako oddzielną sekwencje. Jaka jest całkowita długość wszystkich regionów?

   (c) Korzystając z narzędzia transeq należy przekonwertować sekwencje nukleotydowe uzyskane w podpunkcie (b) na aminokwasowe.

   (d) Korzystając z narzędzia patmatdb należy określić w którym z białek uzyskanych w podpunkcie (c) występuje motyw w postaci sekwencji aminokwasów “PTAPPEY”. Należy określić z jakiego genu pochodzi dany fragment w oparciu o jego współrzędne dopisane do nazwy sekwencji (patrz Tabela 1).

   (e) W oparciu o współrzędne sekwencji mRNA genu zidentyfikowanego w podpunkcie (d), odczytane z tabeli we wstępie, oraz narzędzie extractseq należy wyciągnąć jego pełną sekwencje nukleotydową z genomu C15.

2. Uzyskiwanie sekwencji RNA genu zidentyfikowanego w etapie 1 z pozostałych genomów dla których znamy jedynie przybliżoną lokalizacje genów.

Należy stworzyć "workflow" (opcja dostępna na głównym, górnym pasku narzędzia Galaxy) służący do selekcji sekwencji mRNA genu uzyskanego w punkcie 1(e) z genomów pozostałych wirusów z zestawu, dla których dokładna lokalizacja genu jest nieznana. Dla wszystkich pozostałych genomów wykonaj poniższy workflow:

(a) Należy wykorzystać narzędzie water do przeprowadzenia algorytmu dopasowania lokalnego (algorytm Smitha-Watermana) w celu identyfikacji położenia genu. Informacyjnie: Sekwencją nr. 1 powinna być sekwencja wynikowa z zad 1(e), sekwencją nr. 2 ma być pełna sekwencja jednego z genomów, innych niż C15 (każdy sprawdzany po kolei). Plik wyjściowy powinien być w formacie FASTA (ustawienia po prawej stronie). **Ważne: dane wejściowe podaje się dopiero po pełnym przygotowaniu workflow!**

(b) Należy użyć narzędzia skipseq do wybrania wyłącznie drugiej sekwencji z pliku wynikowego uzyskanego w poprzednim punkcie (Number of sequences to skip at start = 1), plik wyjściowy powinien być w formacie FASTA.

(c) Korzystając z narzędzia infoseq należy określić długość sekwencji uzyskanej w poprzednim punkcie (format HTML).

(d) Wyjściem "workflow" powinien być plik uzyskany w podpunkcie (b) i informacja o długości z podpunktu (c).

3. Analiza podobieństwa pomiędzy sekwencjami uzyskanymi w kroku 2 oraz budowa drzewa filogenetycznego.

(a) Sekwencję uzyskaną w punkcie 1(e) i sekwencje uzyskane w kroku 2 należy połączyć w jeden plik za pomocą narzędzia Collapse Collection a następnie wykonać dopasowanie sekwencji, typu MSA, z wykorzystaniem narzędzia ClustalW.

(b) W oparciu o informacje uzyskane w podpunkcie (a) (wyniki ClustalW: dnd) i narzędzie Newick Display należy stworzyć drzewo filogenetyczne określające zależność między badanymi genomami (format wyjściowy PNG). Która z nich jest najbardziej podobna do sekwencji referencyjnej C15?

Literatura

[1] Stephen K Gire, Augustine Goba, Kristian G Andersen, Rachel SG Sealfon, Daniel J Park, Lansana Kanneh, Simbirie Jalloh, Mambu Momoh, Mohamed Fullah, Gytis Dudas, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *science*, 345(6202):1369–1372, 2014.

[2] Gareth M Jenkins, Andrew Rambaut, Oliver G Pybus, and Edward C Holmes. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *Journal of molecular evolution*, 54(2):156–165, 2002.

[3] Juan Martin-Serrano, Trinity Zang, and Paul D Bieniasz. HIV-1 and Ebola virus encode small peptide motifs that recruit Tsg101 to sites of particle assembly to facilitate egress. *Nature medicine*, 7(12):1313–1319, 2001.