Projekt zaliczeniowy - Algorytmy uczenia maszynowego

imie i nazwisko ładnie

Problem

Podejmowanym problemem jest znalezienie dopasowań wirusów do ich gospodarzy - bakterii, na podstawie genomów wymienionych organizmów. Poprzez dopasowanie rozumiane jest to, że dany wirus infekuje konkretną bakterię.

Wprowadzenie

Biologiczne znaczenie przewidywania par wirus-host

Większość całej wirosfery stanowią bakteriofagi - wirusy, które w sposób szczególny zakażają bakterie. Specyficzność infekcji i interakcje między nimi można opisać jako dopasowanie wirus-gospodarz. Aby zbadać różnorodność wirusów i ustalić ich taksonomię, do niedawna stosowano przede wszystkim hodowlę laboratoryjną takich par. Jest to również jedyny sposób na kompleksowe scharakteryzowanie cyklu infekcyjnego wirusa lub jego interakcji z komórką gospodarza na poziomie molekularnym. [1] Jednak hodowla takich par jest często trudna, ponieważ oba organizmy wymagają często specyficznych warunków do wzrostu. Poza poznawaniem różnorodności, identyfikacja par pozwala również na: - zrozumienie koewolucji fagów i ich gospodarzy, - projektowanie eksperymentów w celu zbadania interakcji fag-gospodarz, - wnioskowanie o sieciach zakażeń krzyżowych fag-bakteria, - badanie potencjalnej roli fagów w horyzontalnym transferze genów, rozprzestrzenianiu czynników wirulencji i rozprzestrzenianiu oporności bakterii na antybiotyki

Wiele wymienionych powyżej zastosowań bezpośrednio przyczynia się do badań nad ludzkim zdrowiem - na przykład, wiemy, że bakterie stanowią one bardzo ważną część mikrobiomu w jelitach, a bakteriofagi ciągle na niego wpływają. Poprzez właśnie takie interakcje, powodowane są zmiany w składzie i liczebności społeczności bakteryjnej, co jest ściśle związane z takimi chorobami, jak zespół jelita drażliwego (IBS), nieswoiste zapalenie jelit (IBD), rak jelita grubego (CRC), zakażenie Clostridium difficile (CDI), otyłość i zaburzenia neurologiczne. Niestety, siły, które kształtują skład tych społeczności bakteryjnych, pozostają niewystarczająco poznane, co spowalnia rozwój terapii i biomarkerów opartych na mikrobiomie. [2]

W ostatnich latach rozwój sekwencjonowania następnej generacji (next-generation sequencing) był bardzo dynamiczny i wpłynął na odkrywanie wirusów, uniezależniając je od izolacji wirusów w laboratorium. Ze względu na to, że mikroby i wirusy mogą być bezpośrednio sekwencjonowane bez hodowli, ilość danych metagenomicznych do przetworzenia stale rośnie. [3] W związku z tym pojawia się poważny problem - podczas gdy hodowane wirusy są nieodłącznie związane z jednym ze swoich gospodarzy, tj. z tym, z którego wirus został wyizolowany, nie dotyczy to fagów identyfikowanych wyłącznie na podstawie metagenomów. Jeśli więc większość fagów pochodzących z metagenomów nie jest związana z konkretnym gospodarzem, stwarza to zupełnie nowe wyzwanie dla bioinformatyków, polegające na stworzeniu metod efektywnego i dokładnego przewidywania par wirus-gospodarz na podstawie danych metagenomicznych. [1] [4]

Aktualne podejścia przewidywań [1]

Obecnie, nie ma uniwersalnego rozwiązania, które dawałoby znacząco lepsze wyniki niż inne metody. Przyczyna leży w nadal nie do końca poznanej złożoności problemu oraz różnorodności metod, które można wykorzystać do przewidywania. Istnieją trzy główne kategorie metod, a w każdej z nich istnieje wiele cech, które służą do wyszukiwania dopasowań, co stwarza możliwość opracowania różnorodnych podejść. Każda kategoria ma swoje charakterystyczne zalety i ograniczenia:

- **Metody oparte na przyrównaniach sekwencji** wykazują ogólnie wysoką specyficzność/dokładność i są w stanie zidentyfikować gospodarzy fagów niezwiązanych z genomem referencyjnym faga, ale zazwyczaj są silnie uzależnione od bazy danych genomu gospodarza, co często prowadzi do niskiego współczynnika czułości. Dodatkowo, są one bardzo powolne w działaniu, ponieważ rdzeniem przyrównań są narzędzia lokalnego przyrównywania sekwencji pochodzących z biologicznych baz danych, np. BLAST.
- **Metody wolne od przyrównań sekwencji (alignment-free)** są niezależne od referencyjnych baz danych i dlatego są w stanie zidentyfikować gospodarzy dla zupełnie nowych fagów bez powiązanych z nimi genomów referencyjnych. W porównaniu z innymi metodami wiąże się to jednak z wysokim odsetkiem wyników fałszywie pozytywnych i często wymaga dodatkowych testów statystycznych. Natomiast, ich dużą przewagą jest szybkość działania, która jest o rzędy wielkości większa względem metod opartych na przyrównaniu sekwencji.
- **Metody integracyjne** mają większą dokładność i czułość dzięki uwzględnianiu i integrowaniu wielu sygnałów, ale średnio wymagają dłuższego czasu obliczeń, ponieważ wymagają wyników z kilku indywidualnych podejść predykcyjnych, np. mogą używać klasyfikatorów uczenia maszynowego w połączeniu z przyrównaniami sekwencji.

Cel pracy

Celem jest opracowanie metody przewidywania par wirus-gospodarz w zależności od lokalnych fluktuacji składu sekwencji zarówno w sekwencjach gospodarza, jak i wirusa. Metoda ta ma mieć charakter hybrydowy (integracyjny), ponieważ łączy wykorzystanie cech składu sekwencji z technikami uczenia maszynowego. Część uczenia maszynowego jest obsługiwana przez narzędzie o nazwie fastDNA [5], którego oryginalnym zastosowaniem jest wolna od przyrównań klasyfikacja taksonomiczna krótkich sekwencji metagenomicznych. Dodatkowo, opracowywana metoda będzie najbardziej zbliżona do podejścia zastosowanego w narzędziu WIsH (Who is the host?) [6] i celem będzie również uzyskanie lepszej skuteczności przewidywań od tego narzędzia.

Metody

Zestaw danych

Jako zestaw danych wykorzystywany jest zestaw porównawczy składający się z 820 kompletnych sekwencji genomów fagów i 2698 kompletnych sekwencji genomów bakterii, które zostały pobrane z bazy danych NCBI RefSeq, które wykorzystuje Edwards. [7] Informacje o gospodarzu zostały pobrane z pola "host" w rekordzie RefSeq faga, a fagi, których gospodarz nie miał całkowicie zsekwencjonowanego genomu, zostały usunięte. W ten sposób uzyskano 820 fagów ze 153 różnymi gospodarzami bakteryjnymi. W celu lepszego zarządzania tymi plikami, utworzone zostały pliki json z metadanymi tych genomów, w których znajdują się:

- numer akcesyjny z bazy NCBI
- pełna nazwa organizmu
- NCBI Taxonomy ID
- poziomy taksonomiczne
- dla metadanych wirusów - ich prawidłowy gospodarz

Te metadane są ekstensywnie używane na każdym etapie działania programu i co najważniejsze, pozwalają ocenić prawdziwość przewidywań. Do uczenia modelu jak i testów jest używany ten sam zestaw danych - nie jest to problemem, ponieważ model jest uczony jedynie na genomach gospodarzy, a z kolei jego ewaluacja i testy są wykonywane przy użyciu **losowych fragmentów** o określonej długości z genomów wirusów, które mają na celu symulować odczyty z sekwencjonowania metagenomicznego.

fastDNA [5]

fastDNA to narzędzie służące do klasyfikacji taksonomicznej krótkich odczytów z sekwencjonowania metagenomicznego. Nie wykorzystuje ono jednak żadnego z klasycznych podejść obliczeniowych do klasyfikacji taksonomicznej, takich jak lokalne przyrównania sekwencji względem bazy referencyjnej sekwencji (BLAST), mapowanie krótkich odczytów (Burrows-Wheeler aligner - BWA), ale przekształca zadanie w problem klasyfikacji wielu klas z wykorzystaniem uczenia maszynowego, gdzie źródłem danych są przekształcenia k-merów sekwencji. W celu przekształcenia sekwencji do postaci odpowiedniej dla metod uczenia maszynowego fastDNA wykorzystuje ciągłe osadzanie sekwencji. Natomiast wykorzystywaną metodą klasyfikacji jest model liniowy, w którym wykorzystywana jest generalizowana wielowymiarowa funkcja logistyczna.

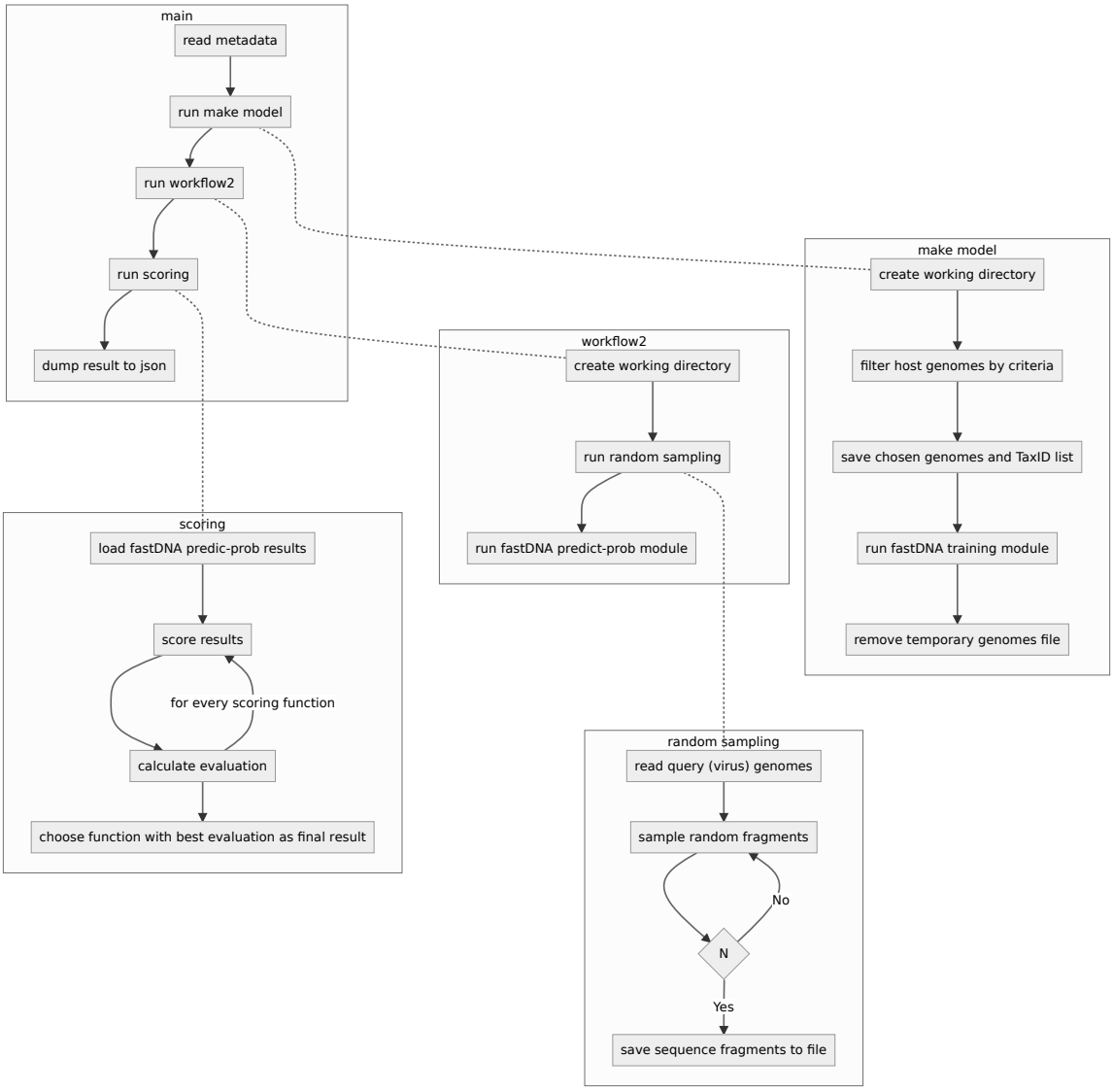
Osadzanie odczytów Sekwencję DNA można zakodować w postaci wektora w N wymiarach ($N = 4^k$, ponieważ na każdą pozycję sekwencji długości k przypada jeden z nukleotydów $\{A, C, G, T\}$). Okazuje się, że nie jest to optymalna reprezentacja, ponieważ stawia duże wyzwanie obliczeniowe zarówno w czasie treningu modeli wykorzystujących dane w takiej formie, jak i w czasie testu - spowodowane jest to tym, że macierz wag o wymiarach $N \times T$ musi być składowana w pamięci dla ewaluacji T gatunków. fastDNA optymalizuje ten krok poprzez osadzanie zbioru odczytów DNA do przestrzeni $\mathbb{R}^d$, gdzie $d \ll N$. W tym celu nadal wyodrębniany jest skład k-merów każdego odczytu, ale zastępowane jest N-wymiarowe kodowanie one-hot każdego k-meru kodowaniem d-wymiarowym, zoptymalizowanym do rozwiązania zadania. Podejście to jest podobne do, na przykład, modelu fastText dla sekwencji języka naturalnego (a sam fastDNA jest w praktyce zmodyfikowaną wersją fastText). [8] [9] Warto zwrócić uwagę, że są wprowadzone pewne zmiany w porównaniu do metod przetwarznania języka naturalnego (NLP) z fastText w celu optymalizacji pod kątem sekwencji DNA:

- w danych sekwencji DNA nie istnieje pojęcie "słowa," czyli grupy liter ograniczonych przestrzenią. W związku z tym stosowana jest rozproszona reprezentacja tylko nakładających się k-merów, a nie słów
- słownictwo jest inne, ponieważ ma ono znaną wielkość (4^k) i jest gęsto reprezentowane (powtarzające się) dla stosunkowo małych wartości k. Przy większych wartościach k k-mery stają się rzadkie, a pojedyncze długie k-mery mogą być dyskryminujące.

Dzięki wyżej wymienionym zmianom, zmniejszone są wymagania pamięciowe dotyczące przechowywania modelu, co przyspiesza czas klasyfikacji, gdy $d < T$, ponieważ macierz $N \times T$ wag jest zastąpiona przez macierz $N \times d$ embeddingów i macierz $d \times T$ wag, które dodatkowo są stałej wielkości (co jest ułatwieniem w implementacji).

Uczenie osadzania Problem jaki podejmuje fastDNA jest zdefiniowany jako wieloklasowy problem klasyfikacji, gdzie każdy z T gatunek bakterii jest traktowany jako klasa i do tych klas musi zostać zaklasyfikowany odczyt. Rozważa się uczenie macierzy M osadzeń i linowego modelu $W \in \mathbb{R}^{T \times d}$, który jest macierzą wag. W tym celu, na podstawie podawanych przykładów zakodowanych sekwencji transformowana jest punktacja przewidywań na prawdpodobieństwa za pomocą funkcji softmax, która jest uogólnieniem funkcji logistycznej dla wielowymiarowych danych. Dzięki temu jest minimalizowane ryzyko empiryczne przy użyciu straty entropii krzyżowej, która mierzy wydajność modelu klasyfikacyjnego, dającego wyniki predykcji z przedziału 0-1. Strata entropii krzyżowej rośnie, gdy przewidywane prawdopodobieństwo odbiega od rzeczywistego zaklasyfikowania, dlatego idealny model miałby wartość straty równą 0. Funkcja celu dla przewidywań jest iteratywnie optymalizowana metodą stochastyczną (stochastic gradient descend - SGD), co jest najbardziej optymalnym podejściem do polepszania wyniku, ponieważ genomy z zestawu treningowego są analizowane po kolei. Warto zauważyć, że gdy $d < T$, to problem jest zwykle niewypukły i SGD może być zbieżne tylko do lokalnego optimum.

Workflow


main
read metadata
run make model
run workflow2
run scoring
dump result to json
make model
create working directory
filter host genomes by criteria
save chosen genomes and TaxID list
run fastDNA training module
remove temporary genomes file
workflow2
create working directory
run random sampling
run fastDNA predict-prob module
scoring
load fastDNA predic-prob results
score results
for every scoring function
calculate evaluation
choose function with best evaluation as final result
random sampling
read query (virus) genomes
sample random fragments
No
N
Yes
save sequence fragments to file

Na powyższym diagramie przedstawiono zarys całkowitego przebiegu działania programu do przewidywania par wirus-gospodarz. Trzeba zaznaczyć, że całkowity przebieg oznacza uwzględnienie etapu treningu modelu, który nie musi zostać przeprowadzany w przypadku końcowego użytkownika, ponieważ będzie on mógł użyć gotowych modeli do przeprowadzenia predykcji.

Trening modelu i optymalizacja Trening modelu jest realizowany poprzez moduł make-model. Umożliwia on filtrowanie zestawu treningowego po poziomach taksonomicznych, dzięki czemu można łatwo wytrenować modele specyficzne do poziomu taksonomicznego. Do treningu potrzebny jest plik FASTA ze wszystkimi genomami bakterii oraz lista NCBI Taxonomy ID tych bakterii. Sam trening jest realizowany przez moduł supervised fastDNA. Moduł ten wymaga zdefiniowania następujących parametrów:

- wymiarowości wektorów,
- długości osadzanych sekwencji,
- minimalnego rozmiaru słowa,
- maksymalnego rozmiaru słowa,
- ilości epok treningu

Wybranie optymalnych parametrów do uzyskania jak najlepszego modelu (tj. dającego możliwie najwięcej prawidłowych przewidywań par wirus-gospodarz) nie jest trywialnym zadaniem. Aby znaleźć jak najlepsze parametry, wykorzystana została metoda optymalizacji Bayesowskiej. Metoda ta polega na skonstruowaniu rozkładu następczego funkcji (procesu gaussowskiego), który najlepiej opisuje funkcję, którą chcemy zoptymalizować. Wraz ze wzrostem liczby obserwacji rozkład potomny ulega poprawie, a algorytm staje się bardziej pewny, które regiony przestrzeni parametrów są warte zbadania, a które nie. [10] W przypadku opisywanego programu, optymalizowaną funkcją jest tak naprawdę cały program, który zwraca przewidywania par wirus-gospodarz. Te przewidywania są oceniane na poszczególnych poziomach taksonomicznych jako procent prawidłowo przewidzianych afiliacji taksonomicznych dla wirusów na każdym z tych poziomów, zatem idealnym wynikiem byłaby wartość 100 na każdym poziomie taksonomicznym. Dodatkowo zadanie optymalizacji jest wymagające obliczeniowo, ponieważ każdy krok po przestrzeni parametrów do optymalizacji jest związany z generowaniem nowego modelu. Samo generowanie modelu nie jest procesem szybkim, a jego wymagania obliczeniowe są najbardziej zależne od ilości rozpatrywanych gatunków organizmów oraz ilości epok treningu. > jakie parametry do treningu

Nadal trwa aktywna praca nad optymalizacją modeli, gdzie poszukiwane są optymalne wartości wielkości słów (przedział przeszukiwnia X-X), wymiarowości wektoró (X-X) oraz długości osadzanych sekwencji (X-X).

Przewidywanie par i punktacja Dopasowywanie par wirus-gospodarz odbywa się poprzez zmodyfikowany moduł predict-prob z fastDNA. Modyfikacja polega przede wszystkim na zmianie formy zwracanych danych, do formatu json. Jak już to było wspomniane, fastDNA nie jest używane dokładnie zgodnie z przeznaczeniem. W normalnym przypadku użycia do modułu predict-prob podawane są krótkie odczyty z sekwencjonowania metagenomicznego, a zwracana jest lista skojarzonych z odczytami gatunków wraz z wartością prawdopodobieństwa przewidywania. W przypadku tego programu, otrzymujemy ten sam wynik, natomiast kontekst jest inny - dla fragmentu wirusa otrzymywany jest skojarzony z nim gatunek bakterii. Co ważne, niemożliwe jest przewidzianie pary na podstawie jednego fragmentu z danego gatunku wirusa. Liczba takich fragmentów nie jest stała i określona, chociaż podczas testów dobrze sprawdzał się przedział wartości 250-500 fragmentów. Z tego powodu, że jest n fragmentów wirusa, to otrzymujemy n potencjalnych dopasowań z określonymi prawdopodobieństwami, wśród których pewne pary mogą się powtarzać. Z tego powodu prawdopodobieństwa m takich samych par są oceniane funkcjami punktującymi. Aktualnie są ekstensywnie prowadzone testy różnych funkcji punktujących (około 30 funkcji), wśród których znajduje się np. średnia geometryczna

prawdopodobieństw trafień, średnia harmoniczna prawdopodobieństw trafień, uśredniany ranking trafień. Jako końcowy wynik, w aktualnej wersji programu, zwracany jest słownik w postaci pliku json, gdzie dla każdego zadanego wirusa, czyli klucza w słowniku, zwracane są dopasowania gospodarza reprezentowane jako krotka z nazwą gospodarza i punktacją przewidywania (**nie** surowym prawdopodobieństwem).

Przykładowy wynik:

```
{
    "NC_000866": [
        [
            "Bacillus_cereus",
            0.1059604194
        ],
        [
            "Peptoclostridium_difficile",
            0.0596440553
        ],
        [
            "Acinetobacter_baumannii",
            0.0539802157
        ],
        [
            "Haliscomenobacter_hydrossis",
            0.0390197225
        ],
...
```

Wyniki

~~**dokładność (accuracy)**~~

~~**Czułość (recall)**~~

~~**precyzja (precision)**~~

~~**specyficzność (specificity)**~~

Krzywa ROC

Precision recall curve

miara F1 (F1-score)

Procenty prawidłowych przewidywań na poziomach taksonomicznych

Porównanie z WIsH

Bibliografia

[1] C. Coclet and S. Roux, "Global overview and major challenges of host prediction methods for uncultivated phages," *Current Opinion in Virology*, vol. 49, pp. 117–126, Aug. 2021.

[2] T. D. S. Sutton and C. Hill, "Gut Bacteriophage: Current Understanding and Challenges," *Frontiers in Endocrinology*, vol. 10, 2019.

[3] D. Paez-Espino, E. A. Eloe-Fadrosh, G. A. Pavlopoulos, A. D. Thomas, M. Huntemann, N. Mikhailova, E. Rubin, N. N. Ivanova, and N. C. Kyrpides, "Uncovering Earth's virome," *Nature*, vol. 536, no. 7617, pp. 425–430, Aug. 2016.

[4] S. Roux, D. Páez-Espino, I.-M. A. Chen, K. Palaniappan, A. Ratner, K. Chu, T. B. K. Reddy, S. Nayfach, F. Schulz, L. Call, R. Y. Neches, T. Woyke, N. N. Ivanova, E. A. Eloe-Fadrosh, and N. C. Kyrpides, "IMG/VR v3: An integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses," *Nucleic Acids Research*, vol. 49, no. D1, pp. D764–D775, Jan. 2021.

[5] R. Menegaux and J.-P. Vert, "Continuous Embeddings of DNA Sequencing Reads and Application to Metagenomics," *J Comput Biol*, vol. 26, no. 6, pp. 509–518, Jun. 2019.

[6] C. Galiez, M. Siebert, F. Enault, J. Vincent, and J. Söding, "WIsH: Who is the host? Predicting prokaryotic hosts from metagenomic phage contigs," *Bioinformatics*, vol. 33, no. 19, pp. 3113–3114, Oct. 2017.

[7] R. A. Edwards, K. McNair, K. Faust, J. Raes, and B. E. Dutilh, "Computational approaches to predict bacteriophage–host relationships," *FEMS Microbiol Rev*, vol. 40, no. 2, pp. 258–272, Mar. 2016.

[8] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[9] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[10] F. Nogueira, "Bayesian Optimization: Open source constrained global optimization tool for Python." 2014--.