Clustering of sequences

Maciej Sikora

23 listopada 2021

1 Dane do klastrowania

Dane pochodzą z bazy Pfam przez pobranie sekwencji seedu dla losowych rodzin. Wybrane rodziny musiały mieć co najmniej 10 sekwencji, a w przypadku liczby sekwencji większej niż 30 - liczba ta została zredukowana do 30.

2 Metody klastrowania

Klastrowanie zostało wykonane dwoma metodami:

- cd-hit
- CLANS

Klastrowanie cd-hit zostało wykonane z parametrem podobieństwa 40% dla porównywania słów długości 2 (-c 0.4 -n 2)

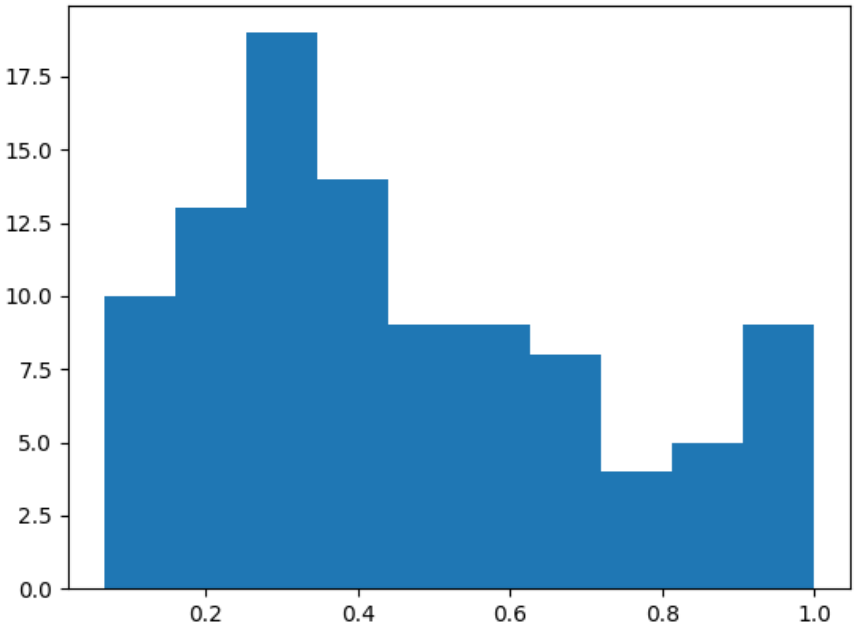
Klastrowanie CLANS opiera się na algorytmie BlastP na podstawie którego liczone są wartości attraction value z zakresu $0 - 1$, gdzie 1 oznacza najbardziej podobne sekwencje. Dalej sekwencje są optymalizowane pod względem attraction value na przestrzeni 2-wymiarowej. Dalej klastry wyznaczane są na podstawie połączeń tak, że najmniejszy klaster może mieć 10 sekwencji.

3 Wyniki i analiza

Jakość klastrowania została wyznaczona przez liczenie wartości indexu Jaccard all vs all w sposób binarny. Dla każdej ze 100 rodzin (true label) zapisany został najlepszy wynik indeksu i sporządzony histogram wartości.

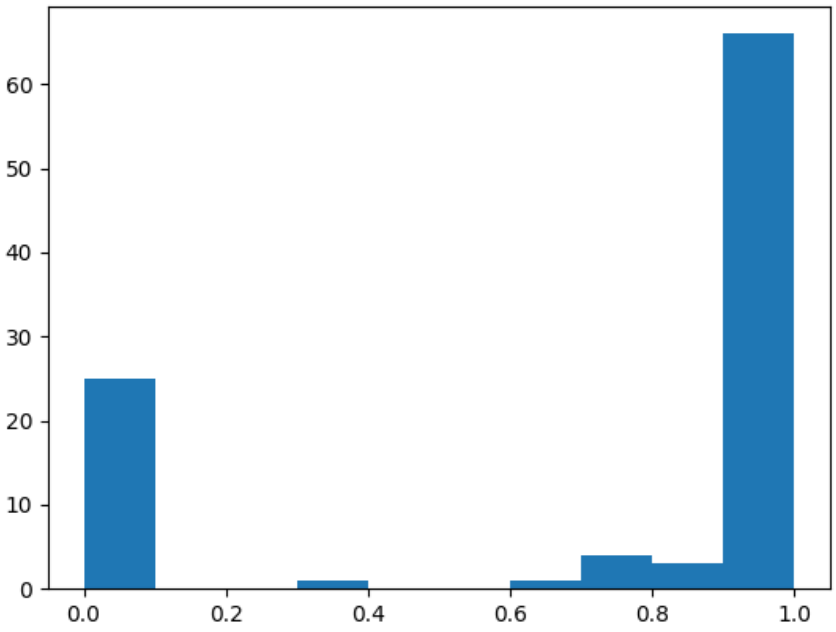
3.1 cd-hit

W tym przypadku cd-hit zwrócił 874 klastry - można więc się spodziewać, że niskie wyniki będą w dużym stopniu wynikały z niekompletności predykcji klastrów. Ok. 15 klastrów posiadało wartości indeksu powyżej 0.9.



3.2 CLANS

CLANS zwrócił 75 klastrów - tu niskie wartości mogą więc wynikać z faktu, że 2 rodziny zostały zaklasyfikowane jako jedna (prawdopodobnie pochodzą z jednego klanu) – potwierdza to fakt, że wiele klastrów ma rozmiar 60, a około 25 klastrów ma wartość 0. Klastrowanie wygląda jednak znacznie lepiej w tym przypadku.



3.3 CLANS dla podziału na klany

W celu zweryfikowania hipotezy dla klastrowania dla klanów przeprowadzone zostało dodatkowe liczenie histogramu. Przy podziale na klany mamy 93 grupy (klany lub rodziny dla nieposiadających klanu). Liczba grup z wartością indeksu 0 nieco spadła, co może potwierdzać hipotezę.