Algorytm do klasyfikacji próbek metagenomowych

Joanna Dąbrowska
Anna Szymik

Inspiracja

Kawulok *et al. Biology Direct* (2019) 14:20
https://doi.org/10.1186/s13062-019-0251-z

Biology Direct

RESEARCH Open Access

Environmental metagenome classification for constructing a microbiome fingerprint

Jolanta Kawulok*, Michal Kawulok and Sebastian Deorowicz

PLOS ONE

RESEARCH ARTICLE

CoMeta: Classification of Metagenomes Using *k*-mers

Jolanta Kawulok*, Sebastian Deorowicz

Institute of Informatics, Silesian University of Technology, Gliwice, Poland

Preprocessing referencji

1. Sketch dla miasta: zbiór haszowanych k-merów (k = 24), połączonych ze wszystkich próbek i ograniczony poprzez losowe wybranie 10^5 k-merów

2. Usunięcie ludzkich k-merów, przygotowanych analogicznie do sketchy miast na bazie kodującego części ludzkiego genomu

Motywacja za usuwaniem genów ludzkiego pochodzenia

Można oczekiwać, że próbki MetaSUB pobrane z różnych miejsc będą zawierały bardzo podobne fragmenty genomu ludzkiego. Fragmenty te nie są szczególnie pomocne w analizie, dlatego postanowiłyśmy, zgodnie z publikacją na której bazowałyśmy, usunąć DNA ludzkie z badanych próbek.

Preprocessing próbek do klasyfikacji

1. Analogiczny preprocessing, z tym że lista zamiast zbioru haszowanych k-merów (k = 24) z zachowaną kolejnością wystąpienia
2. Jeśli odczyt zawiera jakikolwiek ludzki k-mer, nie bierzemy go pod uwagę

Obliczanie wyniku dla odczytu

1. Dla każdego k-meru z odczytu, jeśli znajduje się w referencji, zwiększamy score odczytu o długość nienakładającą się z poprzednio dodanym k-merem

$$\xi_j = k - o_b \qquad o_b = max(k - q + p, 0)$$

ξ – score j-tego k-meru
o – overlap k-meru q z poprzednio dodanym k-merem p

2. Finalny score dla odczytu: dzielimy otrzymany score przez długość

Obliczanie wyniku dla próbki

1. Przyjmujemy threshold wyniku dla odczytu, powyżej którego uznajemy go za przypisany do danej klasy – odczyt może być przypisany do więcej niż jednej klasy

2. Trzy sposoby liczenia wyniku: simple sum, fractional sum i weighted sum

Obliczanie wyniku dla próbki

Simple sum

Klasa (miasto) otrzymuje 1 punkt za każdy dopasowany odczyt, bez względu na to, do ilu klas został on dopasowany.

Fractional sum

Klasa otrzymuje 1/n punktów za i-ty dopasowany odczyt, gdzie n to liczba klas, do których został on dopasowany.

Weighted sum

Wynik dla klasy to wynik odczytu dla tej klasy podzielony przez sumę wyników odczytów przypisanych do tej klasy

Klasyfikacja: przypisujemy próbkę do klasy, która uzyskała najwyższy wynik

Dodatkowe parametry

- ***M*** – maksymalna liczba klas, do których może być przypisany jeden odczyt
- ***ci*** – minimalna liczba wystąpień k-meru w próbce, aby dodać go do sketchu

U nas najlepiej sprawdziło się M = 3 i ci = 4, przy scorze weighted sum.

Dziękujemy za uwagę!

GitHub repository: https://github.com/D4S1/sample-classifier

Bibliografia

1. Kawulok, J., Kawulok, M. and Deorowicz, S. (2019) ‘Environmental Metagenome classification for constructing a microbiome fingerprint’, *Biology Direct*, 14(1). doi:10.1186/s13062-019-0251-z.
2. Kawulok, J. and Deorowicz, S. (2015) ‘Cometa: Classification of Metagenomes using K-MERS’, *PLOS ONE*, 10(4). doi:10.1371/journal.pone.0121453.