Algorytm do mapowania długich odczytów z błędami

Joanna Dąbrowska
Anna Szymik

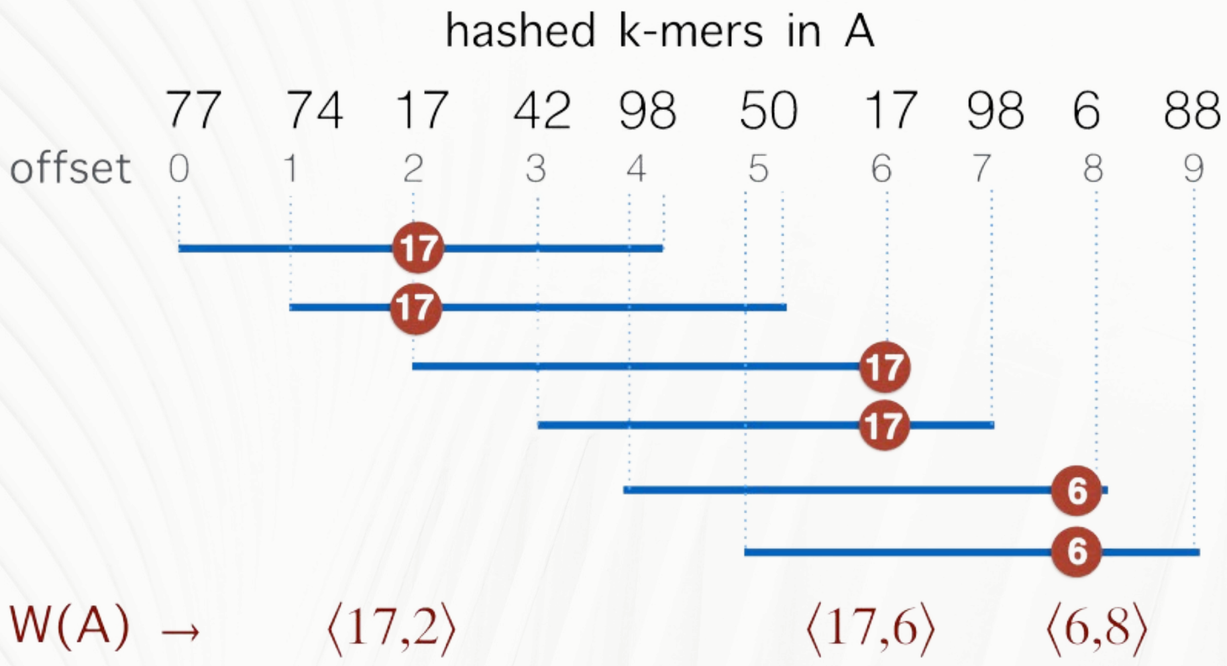
Opis podejścia

Zmodyfikowane podejście minhash + winnowing

1. Preprocessing genomu: znalezienie minimizerów dla zadanego rozmiaru okna
2. Znalezienie minimizerów dla odczytów przy tym samym rozmiarze okna
3. Estymacja indeksu Jackarda przy użyciu sketchu s minimizerów
4. Liczenie odległości edycyjnej dla przefiltrowanych przedziałów, posortowanych według indeksu Jackarda

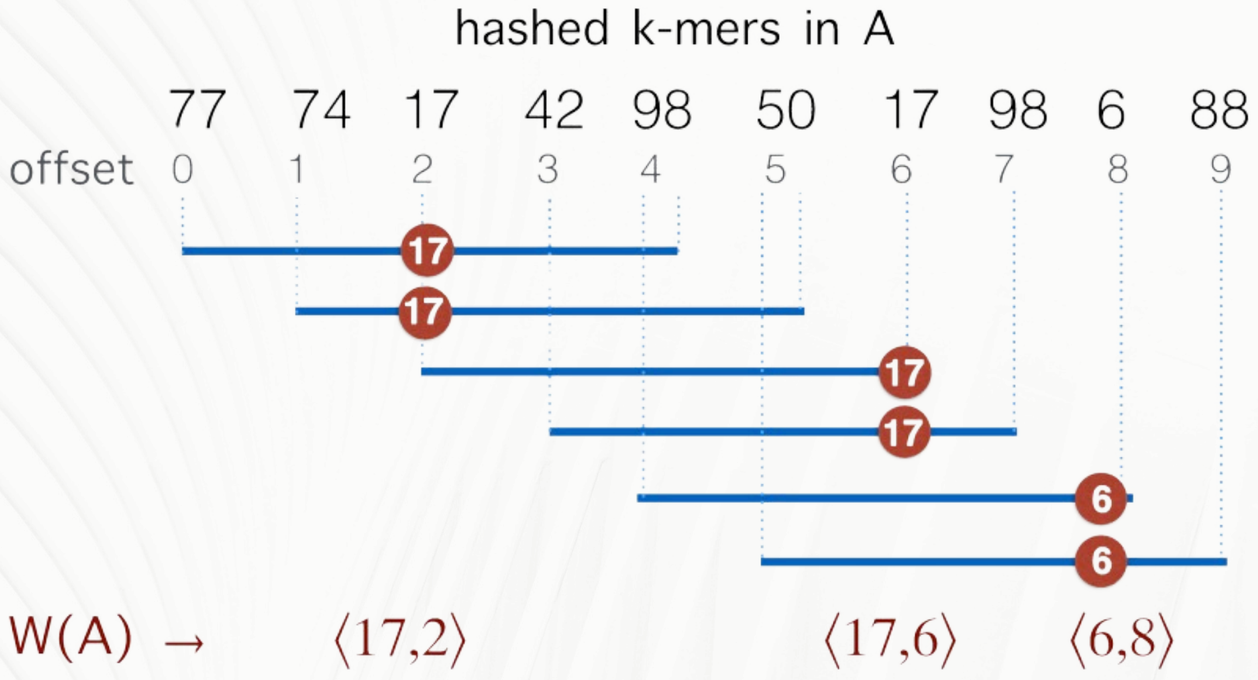
Preprocessing genomu

- **M**: lista minimizerów dla fragmentów genomu o danej długości okna (M[i] – minimizer dla okna zaczynającego się na pozycji i)
- **H**: słownik {wartość funkcji haszującej: lista pozycji minimizerów o danej wartości}



Znalezienie minimizerów dla odczytów

Analogiczne podejście, tylko zamiast listy trzymamy zbiór (dla odczytów nie jest nam potrzebny indeks).



Estymacja indeksu Jackarda

$$\mathcal{J}(A, B_i) = \frac{|S(W(A) \cup W(B_i)) \cap S(W(A)) \cap S(W(B_i))|}{|S(W(A) \cup W(B_i))|}$$

- S(W(A)) – zbiór (sketch) s najmniejszych haszowanych k-merów wybranych metodą *winnowing* z sekwencji A
- w naszym przypadku A to odczyt, a Bi to okno genomu zaczynające się na pozycji i

Filtrowanie

$$\mathcal{J}(A, B_i) \geq \tau \quad \rightarrow \quad |W_h(A) \cap W_h(B_i)| \geq s \cdot \tau \quad \forall i \; 0 \leq i \leq |B| - |A|.$$

$$\tau = \mathcal{G}(\epsilon_{max}, k) - \delta.$$

$$\mathcal{G}(\epsilon, k) = \frac{1}{2e^{\epsilon k} - 1}$$

- G(ϵ, k) – oszacowanie indeksu Jaccarda przy danym współczynniku błędu
- δ – margines błędu w oszacowaniu indeksu Jaccarda przy użyciu 90% przedziału ufności

Ustalanie współrzędnych mapowania

Po przefiltrowaniu pozycji genomu pod względem zawartości k-merów:

- sortujemy kandydatów malejąco pod względem oszacowanego indeksu Jackarda
- zaczynając od pierwszego, liczymy odległość edycyjną mapowania do danego fragmentu
- jeśli jest mniejsza niż 100, zwracamy dane współrzędne mapowania
- jeśli nie osiągniemy mniejszej niż 100, zwracamy najmniejszą

Dziękujemy za uwagę!

Bibliografia

Jain, C. et al. (2018) ‘A fast approximate algorithm for mapping long reads to large reference databases’, Journal of Computational Biology, 25(7), pp. 766–779. doi:10.1089/cmb.2018.0036.