Universität Bielefeld
Technische Fakultät
Genominformatik

Dr. Roland Wittler
M. Sc. Tizian Schulz

Sequenzanalyse-Praktikum
Wintersemester 2021/22

Protokoll 0 (Fiktives Beipiel), Ola Nordmann, 10.10.2021

Thema dieses Aufgabenblocks war die (symmetrische) Maximal-Matches-Metrik, die von der Theorie her bereits aus der Vorlesung bekannt war. Es sollte ein Algorithmus zur Berechnung der Maximal-Matches-Metrik implementiert und auf verschiedene Beispieldatensätze angewandt werden.

1. Die erste Aufgabe hatte zum Ziel, eine effiziente Implementierung des aus der Vorlesung bekannten Algorithmus zur Berechnung der Maximal-Matches-Metrik zu erstellen. Die Theorie dazu findet sich im Skript in Abschnitt 3.8: Zunächst wird dort die (nicht symmetrische) Maximal-Matches-Distanz zweier Sequenzen x und y definiert als

   $$\delta(x||y) := \min\{|P| : P \text{ ist Partition von } x \text{ bezüglich } y\}.$$

   Die Maximal-Matches-Metrik ergibt sich dann durch Logarithmierung und Symmetrisierung:

   $$d_{||}(x,y) := \log(\delta(x||y)+1) + \log(\delta(y||x)+1).$$

   Wie im Skript ausgeführt wird, lässt sich die Maximal-Matches-Distanz in einem Durchlauf als Links-Rechts-Partition berechnen. Dies ist sogar in linearer Zeit $O(|x|+|y|)$ möglich, wenn man den Suffixbaum von y zur Verfügung hat.

   Eine solche Implementierung war uns aber zu aufwändig, weswegen wir eine einfachere Variante gewählt haben, bei der an den Stellen, wo das längste Präfix eines Suffixes von x gesucht wird, das in y vorkommt, eine naive Suche durch y verwendet wird. Damit hat unser Algorithmus für die Maximal-Matches-Distanz die Laufzeit $O(|x| \cdot |y|)$. Da dieser für die Berechnung der Maximal-Matches-Metrik genau zweimal aufgerufen wird, hat deren Berechnung dieselbe Laufzeit.

   Die Java-Implementierung haben wir am 10. Oktober 2020 an die Betreuer des Praktikums geschickt.

2. Zum Test unserer Implementierung haben wir sie auf alle möglichen Paare der in der Aufgabenstellung angegebenen kurzen Sequenzen x, y und z angewandt. Es ergaben sich folgende Distanzen:

d	x	y	z
x	0	6	2
y	6	0	5
z	2	5	0

   Manuell konnten wir überprüfen, dass diese Werte korrekt sind.

3. Um die Maximal-Matches-Distanz zwischen den Genomen von *Corynebacterium glutamicum* und *Mycobacterium tuberculosis* zu berechnen, haben wir uns zunächst deren komplette Genomsequenzen von der NCBI-Datenbank (`http://www.ncbi.nlm.nih.gov/genome`) heruntergeladen. Die Berechnung mittels unseres Programms hat 2 Minuten und 52 Sekunden gedauert. Die Distanz beträgt 45.29.