Marvin Döbel
Rabanus Derr

1	2	3	Σ

Übungsblatt Nr. 4

(Abgabetermin 14.05.2018)

Aufgabe 1

Mit den folgenden Werten (Match: 2, Mismatch: -2, Gap penalty: 2) ergibt sich folgendes Profile Alignment:

Tabelle 1: Profile Alignment Score Berechnung

				V V	C C	L M	W F	C C
			0	-2	-4	-6	-8	-10
I	I	I	-2	-12	-14	-16	-18	-20
E	E	E	-4	-14	-24	-26	-28	-30
C	C	C	-6	-16	-2	-14	-26	-16
I	M	I	-8	-18	-14	-10	-22	-28
E	Q	E	-10	-20	-26	-22	-22	-30
C	C	C	-12	-22	-8	-20	-32	**-10**

Daraus resultiert:

Tabelle 2: Profile Alignment der Profile

I	E	C	I	E	C
I	E	C	M	Q	C
I	E	C	I	E	C
-	V	C	L	W	C
-	V	C	M	F	C

Aufgabe 2

1. **T-COFFEE:**
   Dieses MSA-Programm zeichnet sich durch die Integrierbarkeit weiterer MSA-Methoden aus. Mit Hilfe von M-COFFEE, einem Teil der T-COFFEE Distribution, können verschiedene Alignments aus unterschiedlichen Algorithmen z.B. ClustalW, MAFFT zu einem Alignment kombiniert werden.

   Der eigentliche Algorithmus T-COFFEE selber produziert für eine moderate Laufzeit deutlich bessere Alignments als vergleichbare Methoden. Der progressive Ansatz wird auch hierbei verwendet. Eine Alignment-Informationes-Bibliothek aus globalen und lokalen Alignments wird zuvor hergestellt. Weitere strukturelle, heterogene Vergleichsinformationen können weiterhin integriert werden. Diese Bibliothek wird

danach für die Kombinationen der Profile Alignments verwendet. Dabei werden insbesondere nicht nur die im Moment kombinierten Alignments betrachtet, sondern die gesamte Information über alle Alignments aus dem Pre-prozess.

Notredame C, Higgins DG, Heringa J, T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000 Sep 8;302(1):205-17.

2. **Kalign:**
Dieses MSA-Tool basiert ebenfalls auf dem progressiven Ansatz. Die Distanzen der Sequenzen um den Guide Tree zu erhalten werden aber nicht über paarweise Alignments sondern durch den Wu-Manber string-matching Algorithmus berechnet. Dieser Algorithmus hat eine schnellere Laufzeit als die paarweisen-Alignment Ansätze, welche in $O(n^2)$ liegen. Danach werden die Alignments über Profile Alignments kombiniert, wobei dabei Wu-Manber Ankerpunkte aus dem vorausgehenden Algorithmus genutzt werden können, um bessere Alignments zu erhalten. Dabei zeichnet sich dieses Programm durch die Robustheit und die akzeptablen Laufzeiten bei sehr vielen Sequenzdaten aus. Die Genauigkeit ist bei größeren und differenten Datensets erhöhter im Vergleich zu zuvor etablierten Methoden.

Timo Lassmann,Erik LL Sonnhammer, Kalign – an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics. 2005; 6: 298. Published online 2005 Dec 12. doi: 10.1186/1471-2105-6-298

3. **MAFFT:**
Das MSA-Tool nutzte bei seiner erstmaligen Publizierung progessive Methoden (FFT-NS-1, FFT-NS-2). Heutzutage sind zum progressiven Ansatz zwei iterative Methoden hinzugekommen: der iterative Ansatz (FFT-NS-i, NW-NS-i) und den iterativen Ansatz der den weighted sum-of-pairs score (WSP) und consistency scores verwendet (L-INS-i, E-INS-i, G-INS-i). Die progressive Methode nutzt klassische Algorithmen wie CLUSTALW mit einigen veränderungen, die die schnelligkeit des Algorithmus deutlich verbessern. Zusammengefasst läuft der Algorithmus wie folgt ab: (1) Berechne eine ungefähre Distance Matrix aufgrund von physikalisch-chemischen Eigenschaften der Aminosäuren, hierbei spielen besonders das Volumen und die Polarität eine große Rolle. Der Prozess dieser Berechnung wird mithilfe von Fast Fourier Transform (FFT) beschleunigt. Es werden gemeinsame 6-Tupel zwischen jedem Sequenzpaar gezählt, wobei die 20 Aminosäuren auf ein Alphabet von 6 reduziert werden (k-mer counting). (2) Berechne den Guide Tree mithilfe von einer modifizierten Version des UPGMA. (3) Aligne die Sequenzen. Danach kann, um die genauigkeit des Algorithmus zu erhöhen FFT-NS-2 angewendet werden, der diesen Algorithmus um (4) Guide-Tree wird erneut berechnet aufgrundlage von dem ersten Alignment und (5) das zweite progressive Alignment erweitert. Die beiden iterativen Methoden werden erst nach dieser Berechnung angewandt und verbessern nochmals die genauigkeit des Alignments. FFT-NS-i, NW-NS-i nutzen den WSP-Score und verbessern so das Alignment so lange bis sich am WSP-Score nichts mehr ändert oder die Schleife 1000 (FFT-NS-i) oder 2 (NW-NS-i) mal wiederholt wurde. L-INS-i, E-INS-i, G-INS-i nutzen zusätzlich zum WSP-Score einen COFFEE-like score, der die konsistenz zwischen dem MSA und dem PSA darstellt.

Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyataa - MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. 2002 Jul 15; 30(14): 3059–3066. und https://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html

Aufgabe 3

- Algorithm used: MAFFT (v7.397) (Siehe: https://mafft.cbrc.jp/alignment/server/)
- Method: L-INS-i
- GUIDANCE alignment score: 0.985562
- Coloring scheme: Taylor
- gap scores: affine
- gap opening penalty: 1.53
- Scoring matrix for amino acid sequences: BLOSUM62
- Guide Tree: Optimized UPGMA

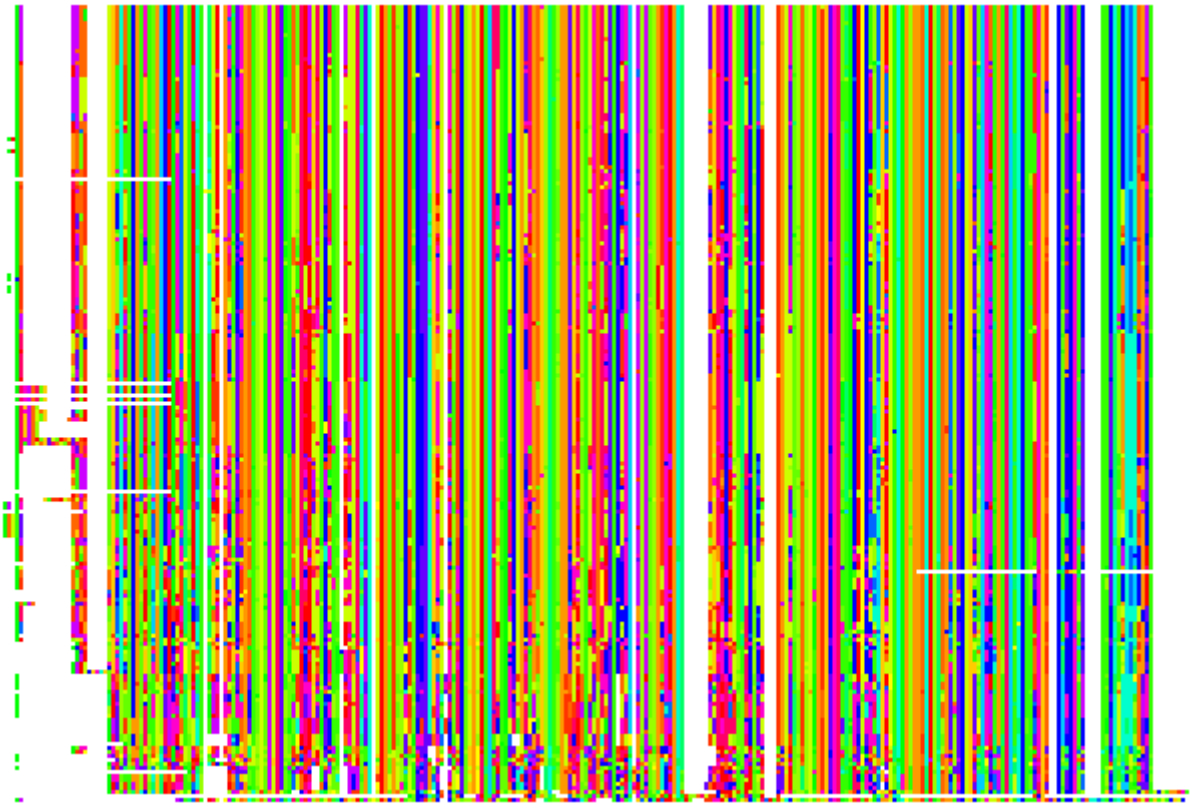

Abbildung 1: Zusammenfassung des Alignments. Hier ist das komplette Alignment zu sehen.

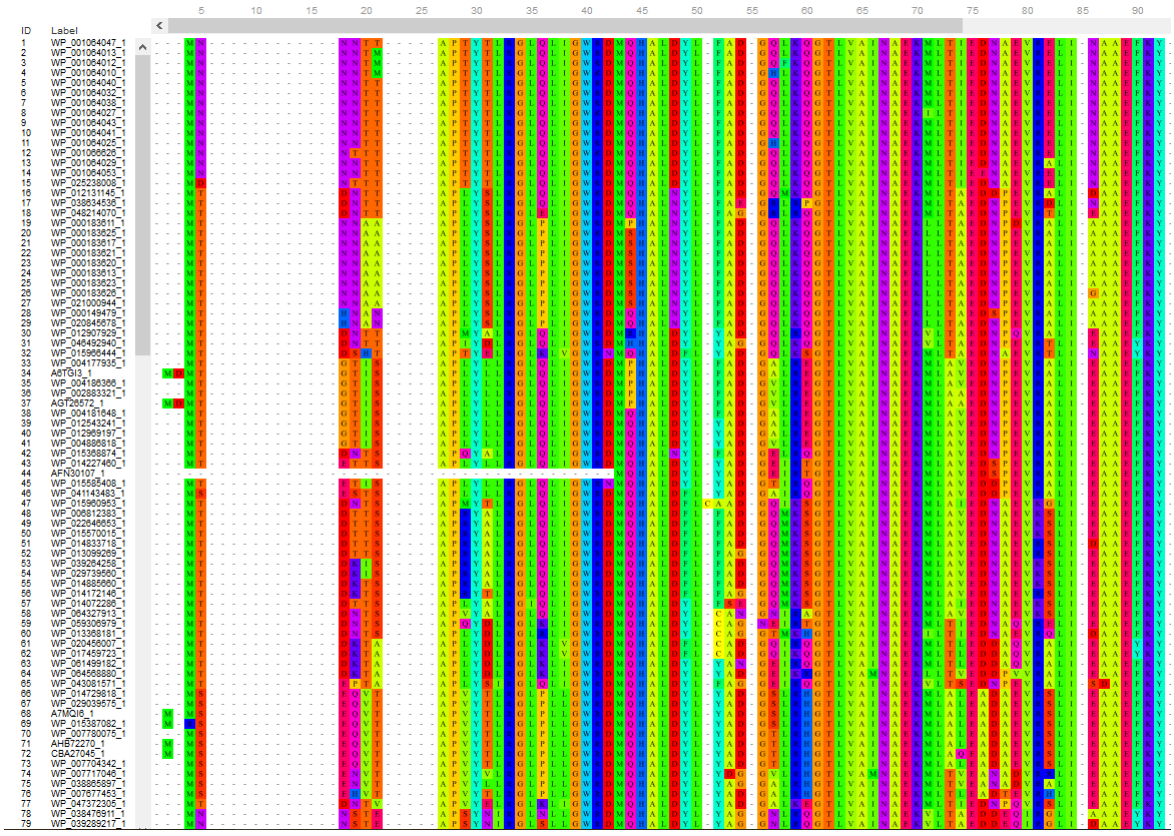

Abbildung 2: Beispiel vom Alignment. Zu sehen sind die ersten 80 Sequenzen und deren erste 150 Aminosäuren.