Bachelorarbeit Bioinformatik

Eine bioinformatische Studie über Proteinsequenz-Clustering mit DIAMOND

Jasmin Katz

Prüfer: Jun.-Prof. Dr. Andreas Dräger

Betreuer: Dr. Hajk-Georg Drost und Benjamin Buchfink

1) Hintergrund:

Das exponentielle Wachstum an Proteinsequenzen in den letzten Jahren [1] führt dazu, dass diese Daten effizient analysiert werden müssen. Hierbei spielt Proteinsequenz-Clustering eine wichtige Rolle. Denn damit ist es möglich, durch Sequenzähnlichkeiten, Homologierelationen, Orthologie, Proteinfamilien, gemeinsame Domänen oder funktionelle Ähnlichkeiten von Proteinen [1] bestimmen zu können. Das Klassifizieren von Proteinsequenzen in disjunkte Cluster ermöglich somit die Reduktion großer Proteinsequenz-Datensätze.

Um Sequenzähnlichkeiten bestimmen zu können, ist der Abgleichen von Proteinsequenzen gegeneinander notwendig. Um diesen Berechnungsschritt ausführen zu können, wird das Tool DIAMOND verwendet. DIAMOND steht für *'double index alignment of next-generation sequencing data*' und ist ein Sequenz-Aligner [2], der im Besonderen für die Analyse von großen Sequenzdaten entwickelt worden ist. Zusätzlich zum Abgleichen von Proteinsequenzen, soll auch das Proteinsequenz-Clustering mit DIAMOND ausgeführt werden. Der dafür verwendete Clustering-

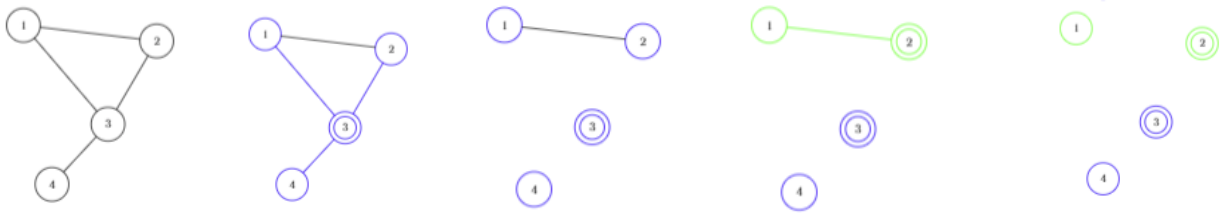

Abbildung 1: Visualisierung des Greedy-Vortex-Cover Algorithmus. Der Algorithmus beruht darauf, dass die Knoten mit den meisten Kanten Repräsentanten eines Clusters werden.

Algorithmus ist ein *‚greedy'* und Graphen-basierter Clustering Algorithmus namens *‚greedy-vortex-cover'* Algorithmus, der in Abbildung 1: Visualisierung des Greedy-Vortex-Cover Algorithmus. grafisch visualisiert ist.

2) Zielsetzung:

In dieser Arbeit soll das Tool DIAMOND [2] und dessen bestehender Clustering-Algorithmus erweitert und verbessert werden. Dabei soll eine kaskadierter Greedy-Vertex-Cover-Algorithmus, auf der Grundlage von Diamond-Alignments, implementiert werden. Außerdem soll der Ressourcenverbrauch durch externes Speichern des Homologiegraphen begrenzt werden. Zudem sollen, durch Vergleichen mit anderen Algorithmen, die Skalierbarkeit und Genauigkeit des Ansatzes untersucht werden. Falls dann noch Zeit bleibt, soll durch, zum Beispiel, die Integration von MinHash, der Algorithmus beschleunigt werden.

3) Voraussetzungen:

Das Aneignen über die Funktionsweise des Tool DIAMOND, sowie Kenntnisse im Programmieren mit der Programmiersprache C++ in der Entwicklungsumgebung Visual Studio. Interesse an der Entwicklungsbiologie und evolutionären Genomik.

Literatur:

[1] G. A. Pavlopoulos. How to cluster protein sequences: tools, tips and commands. *MedCrave.* 2017. DOI: 10.15406/mojpb.2017.05.00174.

[2] B. Buchfink, C. Xie, D. Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods 12.* 2015. https://doi.org/10.1038/nmeth.3176.

[3] A. J. Enright, S. Van Dongen, C. A. Ouzouni. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research. 2002. https://doi.org/10.1093/nar/30.7.1575.