Kürzeste-Wege-Algorithmen für Sequence Alignment

Philipp Seemann

7. Juli 2012

In meiner Bachelorarbeit beschäftige ich mich mit dem Problem des *Sequence Alignment*. Dieses kommt aus der Bioinformatik und dient dort unter anderem zur Bewertung der Ähnlichkeit von Proteinsequenzen oder Genomsequenzen. Man hat dabei eine gewisse Anzahl von Sequenzen, also Zeichenketten sowie eine Metrik (Kostenfunktion) zwischen den verschiedenen Zeichen gegeben. Die Aufgabe ist es nun eine möglichst gute Zuordnung der beiden Sequenzen zu finden, in dem Sinne, dass der durch die Metrik definierte Abstand minimiert wird. Dafür hat man die Möglichkeit die Sequenzen gegeneinander zu verschieben sowie Lücken einzufügen. Nicht erlaubt ist eine Umsortierung der Zeichen, die Reihenfolge muss erhalten bleiben. Die Kostenfunktion hat dabei in der Regel die Eigenschaften, dass zwei gleiche Zeichen den kleinsten Wert und eine Lücke den größten Wert hat.

Beispiel 1. Wir wollen die Sequenzen

$$ACABAA$$
$$CBACCBA$$

optimal alinieren, wobei gleiche Zeichen Kosten 0, verschiedene Zeichen Kosten 1 und Lücken Kosten 2 haben. Die Lösung für dieses Problem ist

$$ACA-BAA$$
$$CBACCBA$$

mit Kosten von 6.

Wir betrachten das Problem zunächst nur für zwei Sequenzen, das sogenannte *Pairwise Sequence Alignment*. Jetzt können wir einen „rechteckigen“ Graphen erstellen, in dem jede mögliche Kombination einen Knoten darstellt und bei (fast) jedem Knoten jeweils genau drei Kanten beginnen und enden, die gemäß der Kostenfunktion gewichtet werden (Stichwort *Needleman-Wunsch*). Auf diesen Graphen können wir jetzt zur Lösung des Problems herkömmliche Kürzeste-Wege-Algorithmen verwenden. Dabei soll vor allem der auf dem Dijkstra-Algorithmus aufbauende $A*-Algorithmus$ untersucht werden. Hierbei werden mit einer Heuristik die Kosten von jedem Knoten zum Zielknoten nach unten abgeschätzt. Dies wird bei der Auswahl des nächsten untersuchten Knotens berücksichtigt. Dadurch kann die Anzahl der untersuchten Knoten deutlich reduziert werden und die Laufzeit somit verringert werden. Aufgabe ist es nun, eine dem Problem angemessene Heuristik zu finden.