Algorithmen in der Biologie

Dr. Hans-Joachim Böckenhauer
Dr. Dennis Komm

Zusammenfassung des 2. Abends

Zürich, 30. April 2014

1 Alignment-Verfahren

Für einen Überblick über die Alignment-Algorithmen zur Bestimmung der Ähnlichkeit von Biomolekülen siehe die Vortragsfolien. Eine ausführliche Darstellung findet sich auch im Buch „Algorithmische Grundlagen der Bioinformatik“ von Böckenhauer und Bongartz [1].

2 Phylogenetische Bäume

Das Ziel in diesem Abschnitt ist es, einen kleinen Einblick in die Modellierung der Verwandtschaftsbeziehungen zwischen verschiedenen biologischen Arten zu geben und für eines dieser Modelle einen Algorithmus zur Berechnung eines Stammbaums vorzustellen. Eine ausführlichere Diskussion findet sich zum Beispiel in Kapitel 11 des Buchs „Algorithmische Grundlagen der Bioinformatik“ von Böckenhauer und Bongartz [1].

Wir wollen also für eine gegebene Menge heute lebender biologischer Arten einen Stammbaum finden, der die evolutionäre Entwicklung möglichst gut wiedergibt. Ein solcher sogenannter *phylogenetischer Baum* ist schematisch in Abbildung 1 gezeigt.

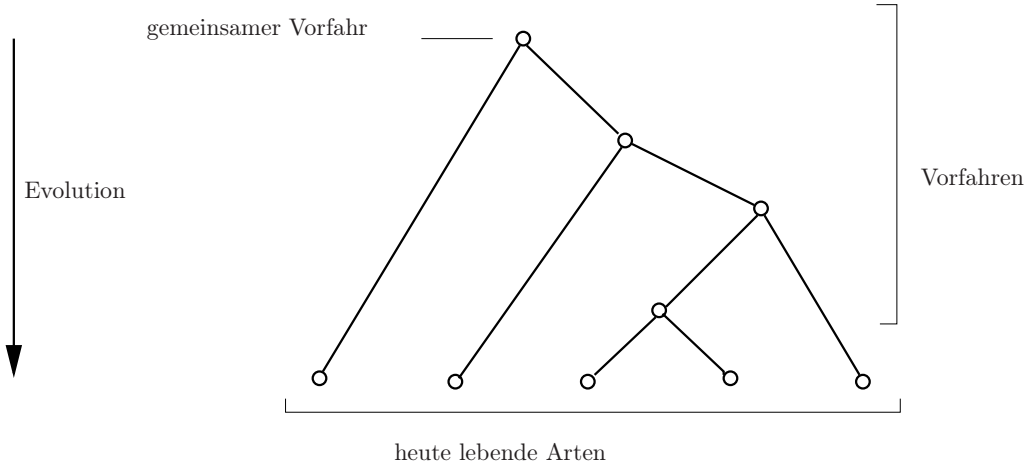

Abbildung 1. Schema eines Stammbaums

Als Eingabe können verschiedene Arten von Wissen über die einzelnen Arten dienen:

- Distanzen zwischen den einzelnen Arten, zum Beispiel die Edit-Distanz der Genome oder einzelner Gene,
- phänotypische Merkmale der Arten, also äusserlich erkennbare Abweichungen der Form, der Farbe, der Grösse etc.,

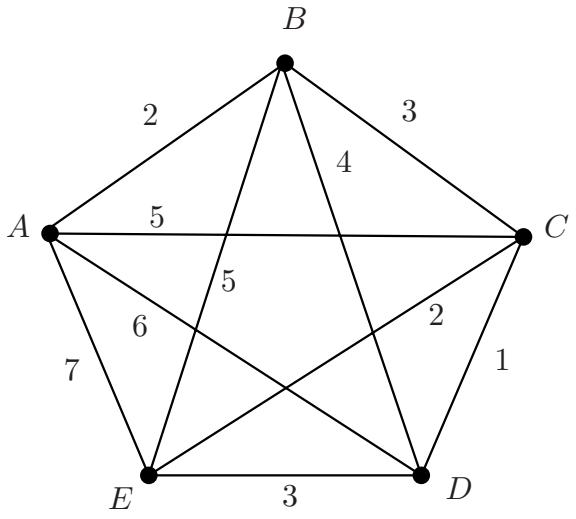

Abbildung 2. Ein Distanzgraph

- genotypische Merkmale, zum Beispiel das Vorhandensein einzelner Gene, Genexpressionsdaten etc.

Gesucht ist nun eine Einordnung der Arten in einen phylogenetischen Baum, so dass die Entfernung zwischen je zwei Arten in dem Baum die Distanzen bzw. die Übereinstimmung der Merkmale wiederspiegelt. Hierfür sind eine Fülle verschiedener Modelle und Algorithmen untersucht worden. Wir wollen hier ein einfaches Modell näher untersuchen, bei dem wir folgende Zusatzannahmen treffen:

- Alle in dem Baum vorkommenden Arten, also auch die bereits ausgestorbenen Vorfahren bis hin zum gemeinsamen Vorfahren aller gegebenen heute lebenden Arten, sind bekannt.
- Für je zwei Arten ist der exakte Wert eines gegebenen Distanzmasses *dist* bekannt.
- Es gab während der Evolution in diesem Stammbaum keine „Rückwärtsentwicklungen;“ das heisst für je drei Arten x, y und z in einer Abstammungslinie (x ist Vorfahr von y und y ist Vorfahr von z) gilt, dass

$$dist(x, y) + dist(y, z) = dist(x, z).$$

  In diesem Fall sprechen wir von einem *additiven Stammbaum.*

Um aus den gegebenen Distanzen einen passenden Stammbaum zu bestimmen, stellen wir die Distanzen zunächst in einem *Distanzgraphen* dar. Die Knoten dieses vollständigen Graphen stellen die Arten dar, die Kantengewichte entsprechen den gegebenen Distanzen, siehe Abbildung 2 für ein Beispiel.

Jeder mögliche Stammbaum für die gegebenen Arten ist ein Baum, der alle Knoten des Distanzgraphen enthält, also ein sogenannter *Spannbaum* des Distanzgraphen. Um jetzt den Stammbaum zu finden, der am besten zu den gegebenen Distanzen passt, suchen wir einen Spannbaum des Distanzgraphen, in dem die Summe der Kantengewichte minimal wird.

Satz 1. *Wegen der Additivitätsbedingung an die Distanzen ist dieser minimale Spannbaum eindeutig bestimmt.*

Beweis. Nehmen wir an, dass es in dem Distanzgraphen zwei verschiedene minimale Spannbäume T und T' gibt. Wir versuchen, diese Annahme zu einem Widerspruch zu führen, damit haben wir dann die Eindeutigkeit gezeigt.

Sei e eine Kante in T' zwischen zwei Knoten x und y, die nicht in T vorkommt. Seien S_1 und S_2 die beiden Teilbäume, die entstehen, wenn man die Kante e aus T' entfernt, sei x in S_1 und y in S_2. Weil auch T ein Spannbaum ist, gibt es in T einen Pfad P von x nach y. Wegen der Additivität der Distanzen sind alle Kanten auf dem Pfad P billiger als die Kante e. Wenn wir also aus T' die Kante e entfernen und dafür eine Kante e' aus P hinzufügen, erhalten wir einen Spannbaum, der billiger als T' ist. Dies ist ein Widerspruch zu unserer Annahme, dass T' ein minimaler Spannbaum ist. Somit muss unsere Annahme falsch sein, es gibt also keine zwei verschiedenen minimalen Spannbäume in einem Distanzgraphen. □

Wir stellen jetzt einen Algorithmus vor, der diesen eindeutigen Spannbaum findet.

Eingabe: Ein Distanzgraph zu einer Menge von Arten.
Schritt 1: Starte mit beliebigem Knoten x.
Schritt 2: Sind alle Knoten im Spannbaum enthalten?
Schritt 3: Falls ja, gib den berechneten Spannbaum aus.
Schritt 4: Sonst füge zu dem bereits berechneten Spannbaum die (eindeutig bestimmte) billigste Kante hinzu, die ihn mit einem noch nicht enthaltenen Knoten verbindet und gehe zurück zu Schritt 2.
Ende

Dieser Algorithmus wird nach seinem Erfinder *Algorithmus von Prim* genannt. Eine Ausführung auf dem Distanzgraphen aus Abbildung 2 ist in Abbildung 3 gezeigt. Wir können diese Ausführung wie folgt beschreiben:

(a) Im ersten Schritt wird der Knoten $x = A$ gewählt. Die billigste Kante von x aus ist die Kante $\{A, B\}$.

(b) Da diese Kante noch kein vollständiger Spannbaum ist, wird nun die billigste Kante von einem der beiden Knoten A und B zu einem noch nicht verbundenen Knoten (also C, D oder E) gesucht. Dies ist die Kante $\{B, C\}$.

(c) Die billigste Kante, die dann A, B oder C mit einem der noch nicht verbundenen Knoten D oder E verbindet, ist die Kante $\{C, D\}$.

(d) Mit der Kante $\{C, E\}$, die den noch nicht verbundenen Knoten E an den bisher berechneten Spannbaum anschliesst, ist der Spannbaum jetzt vollständig.

Dieser Algorithmus ist korrekt, weil man die folgende Beobachtung beweisen kann.

Satz 2. *Sei G ein Graph mit Knotenmenge V, sei $V = U_1 \cup U_2$ eine Aufteilung der Knoten in zwei disjunkte Teilmengen. Weiter seien $u_1 \in U_1$ und $u_2 \in U_2$ zwei Knoten, so dass $\{u_1, u_2\}$ die billigste Kante zwischen U_1 und U_2 ist. Dann gibt es einen minimalen Spannbaum, der $\{u_1, u_2\}$ enthält.*

Beweis. Wir nehmen an, dass kein minimaler Spannbaum die Kante $e = \{u_1, u_2\}$ enthält und führen diese Annahme zu einem Widerspruch. Sei T eine minimaler Spannbaum. Wenn wir e zu T hinzufügen, dann entsteht ein Kreis. Der Pfad P von u_1 nach u_2 in T enthält eine Kante $\{x, y\}$, die von U_1 nach U_2 führt (gegebenenfalls kann es auch mehrere solche Kanten geben), siehe auch Abbildung 4.

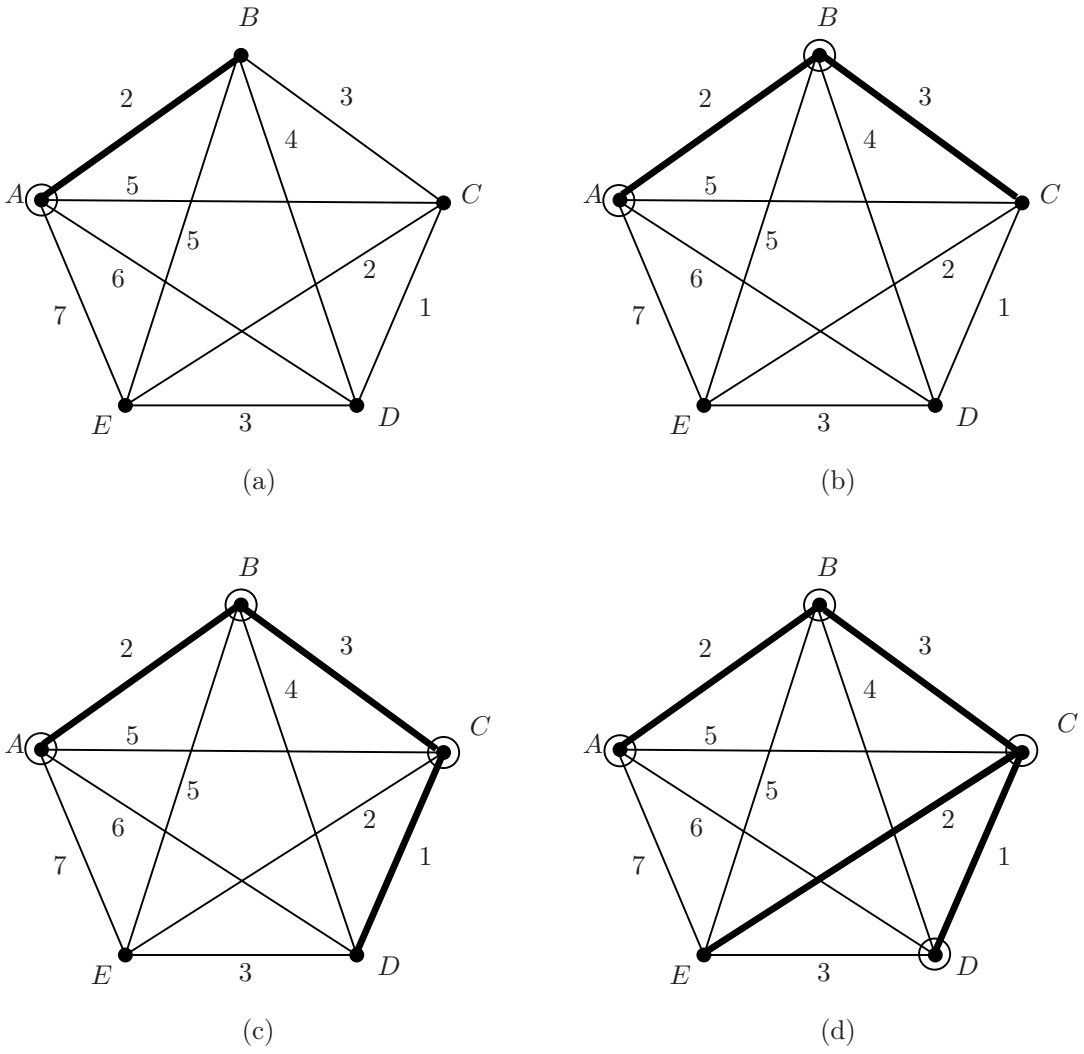

Abbildung 3. Ein Beispiel für die Ausführung des Algorithmus von Prim auf dem Distanzgraphen aus Abbildung 2

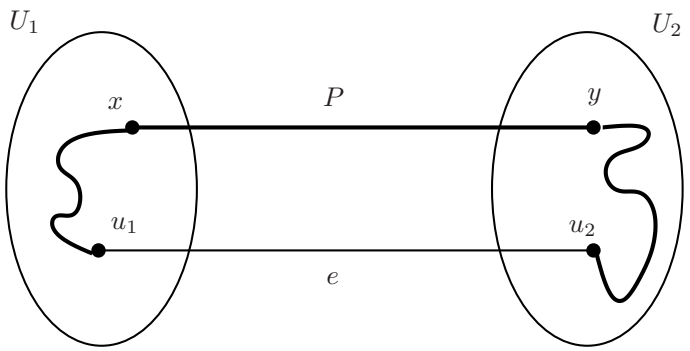

Abbildung 4. Jeder minimale Spannbaum enthält die billigste Kante von U_1 nach U_2.

Weil $\{u_1, u_2\}$ die billigste Kante zwischen U_1 und U_2 ist, ist $\{x, y\}$ mindestens so teuer wie $\{u_1, u_2\}$, damit ist der Spannbaum, der entsteht, wenn man in T die Kante $\{x, y\}$ löscht und die Kante $\{u_1, u_2\}$ einfügt, auch ein minimaler Spannbaum im Widerspruch zu unserer Annahme. Also ist die Annahme falsch und es gibt einen minimalen Spannbaum, der $\{u_1, u_2\}$ enthält. □

3 Weiterführende Literatur

[1] Eine Einführung in die Alignment-Algorithmen findet sich in Kapitel 5 und eine Einführung in die möglichen Modelle zur Stammbaumberechnung in Kapitel 11 des Buchs

H.-J. Böckenhauer, D. Bongartz: *Algorithmische Grundlagen der Bioinformatik*, Teubner-Verlag 2003.