Übungen zur Phylogenetik Vorlesung

Universität Bielefeld, WS 2011/2012, Dr. Roland Wittler

http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2011winter/Phylogenetik

Blatt 4 vom 3.11.2011

Abgabe in einer Woche zu Beginn der Vorlesung oder vorab im Briefkasten bei U10-151.

Aufgabe 1 Walter M. Fitch. **(6 Punkte)**

In dieser Aufgabe betrachten wir die Originalarbeit von Walter M. Fitch von 1971: "Towards Defining the Course of Evolution: Minimum Change for a Specific Tree Topology", publiziert in dem Journal "Systematic Zoology". Der Artikel ist online als PDF verfügbar: http://www.jstor.org/stable/2412116.

Fitch verwendet im Vergleich zur Formulierung im Skript folgende Terminologie: Ein *nodal set* entspricht einem *set S*, ein *immediate ancestor* ist ein *parent node*, mit *immediate descendant* ist ein *child node* gemeint, und der *ultimate (ancestral) node* ist der *root node*. Als *character states* betrachtet Fitch exemplarisch *nucleotides*.

(a) Auf Seite 408 findet sich eine kurze, mathematische Beschreibung, wie das *nodal set* für interne Knoten in der Bottom-Up-Phase gebildet wird (Schritt 1b auf Seite 26 im Skript). Finde den entsprechenden Satz und gib die ersten Worte an.

In der Vorlesung haben wir bereits kurz den Algorithmus auf Seite 410 des Artikels besprochen. Zum Beispiel sorgt Schritt V dafür, dass *alle* Optima gefunden werden können: Ist ein nodal set durch eine Schnittmengenbildung entstanden und die Schnittmenge enthält keine Merkmalsausprägung, die für den Elternknoten in Frage kommt, können alle Elemente in dieser Schnittmenge zu einem Optimum beitragen: Kosten 1 für die Kante zum Elternknoten, keine Kosten für die Kanten zu den Kindknoten. Jedoch kann auch eine Merkmalsausprägung, welche sowohl für den Elternknoten gewählt wurde als auch in einem der nodal sets der Kindknoten enthalten ist, gleiche Kosten verursachen und somit zu einem Optimum führen: keine Kosten für die Kante zum Elternknoten und die Kante zu einem der Kindknoten, Kosten 1 für den anderen Kindknoten. Diesen Schritt V nennt Fitch die "rule of encompassing ambiguity". Fitch beschreibt (auf Seite 410) zwei weitere Regeln.

(b) Erläutere in zwei bis drei Sätzen (ähnlich wie im Beispiel oben) die "rule of diminished ambiguity".

(c) Erläutere ebenso die "rule of expanded ambiguity".

(d) Beschreibe kurz (zwei bis drei Sätze) worum es im Kapitel "Permitted links between nucleotides in successive nodal sets" geht.

Aufgabe 2 Anzahl binärer Bäume. **(3 Punkte)**

Conni Count, der wohl erfolgloseste Bioinformatiker seiner Zeit, möchte einen *most parsimonious* phylogenetischen Baum finden, indem er *alle möglichen* ungewurzelten Baumtopologien aufzählt und für jeden Baum die Parsimony-Kosten berechnet um schlussendlich ein Minimum auszugeben.

(a) Mit seiner Implementierung des Fitch-Algorithmus schafft er es, in einer Sekunde 1 000 000 Bäume durchzurechnen. Conni ist 30 Jahre alt. Wie alt müsste er werden, um das Ergebins für einen Datensatz mit 17 Spezies noch mitzubekommen?

(b) Conni bekommt zu Weihnachten einen Rechencluster geschenkt. Mit einem Terahertz und einer Baumberechung pro Takt, schafft er nun $10^{12} = 1\,000\,000\,000\,000$ Bäume pro Sekunde. Unser Universum ist etwa 15 000 000 000 Jahre alt. Wie viele Blätter hätte Connis Programm bis heute höchstens verarbeiten können, wenn es bereits zum Urknall gestartet worden wäre?

Tipp: Versuche nicht die Formel $U_n = \prod_{i=3}^{n}(2i-5)$ nach n umzuformen. Stattdessen berechne einfach U_n für immer größeres n. (Ein Tabellenkalkulation kann hier gute Dienste leisten.)