Übungen zur Phylogenetik Vorlesung

Universität Bielefeld, WS 2014/2015, Dr. Roland Wittler, Kevin Lamkiewicz

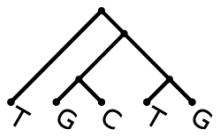
http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2014winter/Phylogenetik

Blatt 4 vom 29.10.2014

Abgabe in einer Woche zu Beginn der Vorlesung oder vorab bei deinem Tutor oder beim Veranstalter.

Aufgabe 1 Small Parsimony – Fitch-Algorithmus. **(3 Punkte)**

(a) Wende den Fitch-Algorithmus auf den rechts stehenden Baum an. Gib eine Lösung und die *Parismony*-Kosten an, die mit dem Algorithmus gefunden werden kann. Gib ebenfalls für jeden inneren Knoten die Kandidatenmenge S an (analog zur Abbildung auf Seite 28 im Skript).



(b) In der Vorlesung haben wir bereits kurz den Algorithmus aus der Originalarbeit von Walter M. Fitch (1971) besprochen. (Er ist auf der Rückseite dieses Übungszettels abgedruckt.) Dieser kann nach der Bottom-Up Phase ausgeführt werden, so dass in einer entsprechend angepassten Top-Down Phase *alle* optimalen Beschriftungen gefunden werden können.
Wende diesen Algorithmus an, um für den oben stehenden Baum die Kandidatenmengen anzureichern. Dies zeigt Lösungen auf, die in Aufgabenteil (a) nicht gefunden werden konnten. Gib eine solche Lösung an und überprüfe die *Parismony*-Kosten. (Die angepasste Top-Down Phase ist hier nicht angegeben, weil neue Lösungen offensichtlich sind.)

Aufgabe 2 Small Parsimony – Sankoff-Algorithmus. **(3 Punkte)**

Wende den Algorithmus von Sankoff mit Einheitskosten auf den Baum aus Aufgabe 1 an, um eine *most parsimonious* Beschriftung der inneren Knoten zu bestimmen. Gib ebenfalls die Werte für $C(u, a)$ für jeden inneren Knoten an (analog zur Abbildung auf Seite 30 im Skript).

Gib alle Lösungen an, die in Aufgabe 1(a) nicht unter den möglichen Lösungen waren.

Für einen Extrapunkt wiederhole die Aufgabe mit der folgenden Kostenfunktion:

cost	A	C	G	T
A	0	2	1	2
C	2	0	2	1
G	1	2	0	2
T	2	1	2	0

Aufgabe 3 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. (Eine Tabellenkalkulation kann hier gute Dienste leisten.)

Bitte wenden!

the data. But this additional information about the upper ancestral node also makes it clear that the first node can not then be a C. The only formulation that will permit the descendent positions to be accounted for in a single replacement requires that replacement to be from A to C in the descent from the first node as shown in Figure 2b (upper right). The elimination of the C from the first node is determined by what may be called the *rule of diminished ambiguity*. Its precise formulation is encompassed in steps I and II of the algorithm, to be presented further on, that contains the complete set of rules for the final phase of reconstructing the nodal sets.

In Figure 2c (middle left) is shown another preliminary phase reconstruction which accounts, using two replacements, for the descent of the characters of the three taxonomic units given. Figure 3d (middle right), however, shows an equally adequate solution which is not encompassed by the possible alternatives available in Figure 3c. Clearly G is a valid alternative for the first node. This case is encompassed by the *rule of expanded ambiguity* which is precisely described in steps III and IV of the forthcoming algorithm.

In Figure 2e (lower left) is shown a third preliminary phase reconstruction that accounts for four descendants using two replacements. In Figure 2f (lower right) is an equally valid solution. Indeed, the C at the lowest node in the preliminary reconstruction is a valid alternative to the A if and only if a C is allowed at the penultimate node above. It is characteristic of this type of case that two nodes, separated by a single node, both contain a nucleotide not present in the intervening node because of the intersection process. Hence, this is called the *rule of encompassing ambiguity* which is formulated as step V of the forthcoming algorithm.

In the preliminary phase, the nodes in Figure 1 were formulated in the order of increasing ancestral remoteness (1→5, with the order for formulating nodes 1 and 2 being arbitrary). In the final phase, the order for correcting the nodal sets must be reversed (5→1).

The preliminary set for the ultimate node is made the final set for that node. We then go to the penultimate node (4 in this case) and proceed according to the following six step algorithm.

I. If the preliminary nodal set contains all of the nucleotides present in the final nodal set of its immediate ancestor, go to II, otherwise go to III.
II. Eliminate all nucleotides from the preliminary nodal set that are not present in the final nodal set of its immediate ancestor and go to VI.
III. If the preliminary nodal set was formed by a union of its descendent sets, go to IV, otherwise go to V.
IV. Add to the preliminary nodal set any nucleotides in the final set of its immediate ancestor that are not present in the preliminary nodal set and go to VI.
V. Add to the preliminary nodal set any nucleotides not already present provided that they are present in both the final set of the immediate ancestor and in at least one of the two immediately descendent preliminary sets and go to VI.
VI. The preliminary nodal set being examined is now final. Descend one node as long as any preliminary nodal sets remain and return to I above.

Figure 1 illustrates the operation of the algorithm. The left hand side (Figure 1a) depicts the preliminary nodal sets. The ultimate ancestral nodal set 5 (AU) is considered the final set and we turn our attention to preliminary nodal set 4. This nodal set does not contain an A and therefore, according to step I, we proceed to step III. Nodal set 4 was not formed by a union and therefore we are directed by step III to go to step V. Following the directions of step V we discover that A is present in both nodal sets 3 and 5 (the rule of encompassing ambiguity) and must therefore be added to nodal set 4. (Mathematically, $((1 \cap 5) \cup (3 \cap 5)) = AU$.

1

[1]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*.