

Bioinformatik 15. 7. 2004, M. Sauerhoff

Algorithmus von Gusfield für perfekte phylogenetische Bäume

Gegeben: $n \times m$ -Merkmalsmatrix M , von der bekannt ist, dass sie durch einen perfekten phylogenetischen Baum darstellbar ist.

Aufgabe: Konstruiere einen solchen Baum.

Algorithmus von Gusfield:

- (1) Sortiere Spalten von M absteigend bez. der Anzahl Einsen (BucketSort).
- (2) Erzeuge Wurzel r .
- (3) **for** $i := 1$ **to** n **do**
 - $v := r$;
 - for** $j := 1$ **to** m **do**
 - if** $M_{ij} = 1$ **then**
 - if** Kante (v, w) mit Markierung j existiert **then**
 - $v := w$;
 - else**
 - erzeuge Knoten w und mit j markierte Kante $\{v, w\}$; $v := w$;
 - speichere Taxon i in Knoten v .
- (4) Für jeden Knoten v , in dem mehr als ein Taxon gespeichert ist und der kein Blatt ist:
Hänge neues Blatt für jedes in v gespeicherte Taxon an v an.
- (5) Für jeden Knoten v , der $d > 2$ Nachfolger hat:
Erzeuge Binärbaum mit $d - 1$ inneren Knoten und lauter unmarkierten Kanten,
identifiziere die Wurzel dieses Baumes mit v und die Blätter mit den Nachfolgern von v .

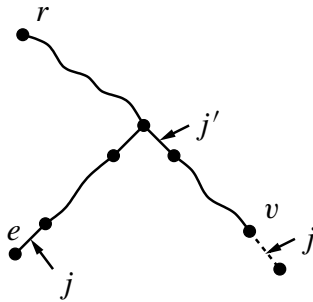
Es ist leicht zu sehen, wie dieser Algorithmus so implementiert werden kann, dass die Rechenzeit $O(nm)$ ist. Es bleibt die Korrektheit zu zeigen.

Für $j \in \{1, \dots, m\}$ sei A_j die von Spalte j in M dargestellte Menge (d. h. die Menge der Taxa mit Merkmal j).

Beweis der Korrektheit des Algorithmus. Es ist klar, dass der Algorithmus einen Binärbaum mit den Taxa an den Blättern produziert. Außerdem sieht man auch sofort, dass die Menge der Merkmale auf dem Pfad von der Wurzel r zu dem Blatt für Taxon i gerade die Menge der Merkmale von Taxon i ist. Es bleibt zu überprüfen, dass es für jedes Merkmal genau eine Kante im Baum gibt, die mit diesem markiert ist. Es ist wiederum klar, dass eine solche Kante existiert; zeige also, dass es auch nur höchstens eine gibt.

Betrachte die Behandlung von Taxon i in Schritt (3). Wir gehen davon aus, dass eine j -Kante an einem Knoten v eingefügt werden soll. Insbesondere ist $M_{ij} = 1$. Wir nehmen an, dass bereits eine andere j -Kante e vorher eingefügt wurde. Weiterhin sei dies das erste Mal, dass bei der Behandlung von Taxon i eine bereits vorhandene Markierung vergeben werden soll. Wir leiten daraus einen Widerspruch her.

Da jedes Merkmal für jedes Taxon nur einmal betrachtet wird, kann die Kante e nicht auf dem Pfad von der Wurzel r zu v vor v liegen und muss für ein anderes Taxon $i' < i$ eingefügt worden sein. Es ergibt sich also die folgende Situation.



Die Kanten an dem Verzweigungsknoten der beiden Pfade sind beide mit irgendwelchen Merkmalen markiert (wir befinden uns noch im Schritt (3) des Algorithmus). Sei j' eine Markierung an der Kante, die zum Knoten v führt. Dann ist aufgrund der Annahme, dass wir das erste Auftreten einer Mehrfachmarkierung bei der Behandlung von Taxon i betrachten, die andere Kante am Verzweigungsknoten nicht mit j' markiert. Damit gilt $M_{i',j'} = 0$ und $M_{i,j'} = 1$. Dies ergibt sich daraus, dass sich der Algorithmus am Verzweigungsknoten bei der Behandlung von i' bzw. i für den linken bzw. rechten Pfad entschieden hat.

Damit folgt $A_{j'} \cap A_j \neq \emptyset$ (wegen i) und $A_j \not\subseteq A_{j'}$ (wegen i'). Schließlich ist $j' < j$, da die Markierungen auf den Pfaden in aufsteigender Reihenfolge vergeben werden. Damit ist $A_{j'} \not\subseteq A_j$ aufgrund der Sortierung der Spalten von M . Dies ist ein Widerspruch zu unserem Kriterium für die Existenz von perfekten phylogenetischen Bäumen. \square