Bảng đặc tả (Character table): dựa vào đặc điểm giải phẫu và sinh lý để lập nên bảng ma trận mô tả tính chất của các đặc điểm.

Vectơ tính chất(Character vector): lưu tính chất của một loài.

Mục tiêu: lập lên cây tiến hoá có các lá biểu hiện cho các chủng loại hiện nay với các vectơ tính chất tương đồng. Ngoài ra, các nút có nút con (interal node) sẽ mang giá trị m để lý giải các tính chất của loài tổ tiên.

Năm 1965, Emile Zuckerkandl và Linus Pauling xuất bản cuốn “Molecules as documents of evolutionary history”, cho rằng chuỗi DNA sẽ cho nhiều thông tin hơn các đặc điểm giải phẫu và sinh lý.

Việc có được đôi cánh dường như đặt ra một thách thức về mặt tiến hóa vì cần có những tương tác sinh lý phức tạp để hỗ trợ cho chuyến bay. Kết quả là chúng ta sẽ tin rằng cánh chỉ tiến hóa một lần ở côn trùng. Lập luận này tương đồng với nguyên tắc không thể đảo ngược của Dollo, một giả thuyết do nhà cổ sinh vật học thế kỷ 19 Louis Dollo đề xuất. Theo nguyên tắc này, khi một loài mất đi một cơ quan phức tạp, chẳng hạn như đôi cánh, cơ quan đó sẽ không xuất hiện trở lại với hình dạng giống hệt ở con cháu của loài đó.

Gắn nhãn cho mỗi lá của cây bằng một hàng có nhiều căn chỉnh và cố gắng suy ra các chuỗi gắn nhãn cho các nút bên trong và tương ứng với các chuỗi tổ tiên ứng cử viên. Cần phát triển một hàm tính điểm để định lượng mức độ phù hợp của cây được gắn nhãn với cách căn chỉnh đa tuyến đã cho. Trong phần tiếp theo, để đơn giản, giả định rằng việc căn chỉnh nhiều lần chỉ chứa các phần thay thế và không có phần indel. Trong thực tế, các nhà nghiên cứu có thể bắt đầu từ nhiều căn chỉnh chứa indel và sau đó loại bỏ tất cả các cột chứa indel.

Điểm trực quan của cây tiến hóa là tổng số đột biến cần thiết để giải thích các chuỗi ở tất cả các nút của cây. Cho một cây T với mỗi nút được gắn nhãn bằng một chuỗi có độ dài m, do đó chúng ta sẽ đặt độ dài của cạnh (v, w) bằng số lượng thay thế (khoảng cách Hamming) giữa các chuỗi có nhãn v và w. Điểm phân tích của T là tổng độ dài các cạnh của nó (xem hình bên dưới).

Giả sử rằng được cung cấp trước cấu trúc của cây nhị phân gốc, trong trường hợp đó chỉ cần gán chuỗi cho các nút bên trong để giảm thiểu điểm phân tích cú pháp.

Vấn đề phân tích nhỏ. Tìm cách ghi nhãn chi tiết nhất về các nút bên trong của cây có gốc.

Đầu vào: Một cây nhị phân có gốc với mỗi lá được đánh nhãn bằng một chuỗi có độ dài m.

Đầu ra: Việc gán nhãn cho tất cả các nút khác của cây bằng các chuỗi có độ dài m nhằm giảm thiểu điểm phân tích của cây.

Cây T có mỗi nút v được gắn nhãn bằng một chuỗi có độ dài m, xác định các cây có nhãn T1, ... , Tm, trong đó Ti có cùng cấu trúc với T và trong đó một nút đã cho được gắn nhãn bởi i- ký hiệu thứ của nút tương ứng trong T.

Vì điểm phân tích của T là tổng điểm phân tích của các cây T1, ... , Tm, nên Bài toán phân tích nhỏ có thể được giải độc lập cho từng cột của căn chỉnh. Quan sát này cho phép chúng ta giả định rằng mỗi chiếc lá được gắn nhãn bằng một ký hiệu duy nhất chứ không phải bằng một chuỗi. Do đó, trọng số của cạnh nối hai nút phải là 0 hoặc 1, tùy thuộc vào việc các nút này được gắn nhãn bằng cùng một ký hiệu hay các ký hiệu khác nhau; cho trước các ký hiệu i và j, chúng ta xác định αi,j = 0 nếu i = j và ﻿αi,j = 1 nếu i ≠ j.

