**THUẦN CHỦNG**

Gene là một đoạn kết gắn các cặp ADN, mỗi cặp ADN được đặc trưng bằng một chữ cái trong tập {***A***, ***C***, ***G***, ***T***}. Gene thuần chủng là gene hình thành từ một đoạn ADN cơ sở độ dài không quá ***m***, được gắn kết lặp đi lặp lại nhiều lần và ở lần lặp cuối cùng có thể chỉ chứa phấn đầu của đoạn cơ sở. Gene được mô tả dưới dạng xâu ***S*** chỉ chứa các ký tự trong tập nêu trên. Như vậy gene thuần chủng là xâu có thể biểu diễn như tổng của ***k*** đoạn cơ sở (***k*** ≥ 0) và có thể có thêm một đoạn đầu của cơ sở.

Ví dụ, với ***m*** = 10, ***S*** = “***ACATAGACATAGACATAGACA***” là một gene thuần chủng vì có đoạn cơ sở là ”***ACATAG***” và ***S*** = ”***ACATAG***” + ”***ACATAG***” + ”***ACATAG***” + “***ACA***”, nhưng với ***m*** = 5 thì ***S*** không phải là gene thuần chủng.

Cho gene ***S*** và giá trị ***m***. Hãy xác định ***S*** có phải là gene thuần chủng hay không và đưa ra *đoạn cơ sở ngắn nhất* nếu S là gene thuần chủng hoặc đưa ra thông báo “***NO***” trong trường hợp ngược lại.

***Dữ liệu:*** Vào từ file văn bản PURE.INP:

* Dòng đầu tiên chứa số nguyên ***m*** (1 ≤ ***m*** ≤ 106),
* Dòng thứ 2 chứa xâu ***S*** độ dài không quá 106 và chỉ chứa các ký tự trong tập đã nêu.

***Kết quả:*** Đưa ra file văn bản PURE.OUT đoạn cơ sở ngắn nhất tìm được hoặc thông báo ***NO***.

***Ví dụ:***

|  |  |  |
| --- | --- | --- |
| PURE.INP |  | PURE.OUT |
| **10**  **ACATAGACATAGACATAGACA** |  | **ACATAG** |
|  |  |

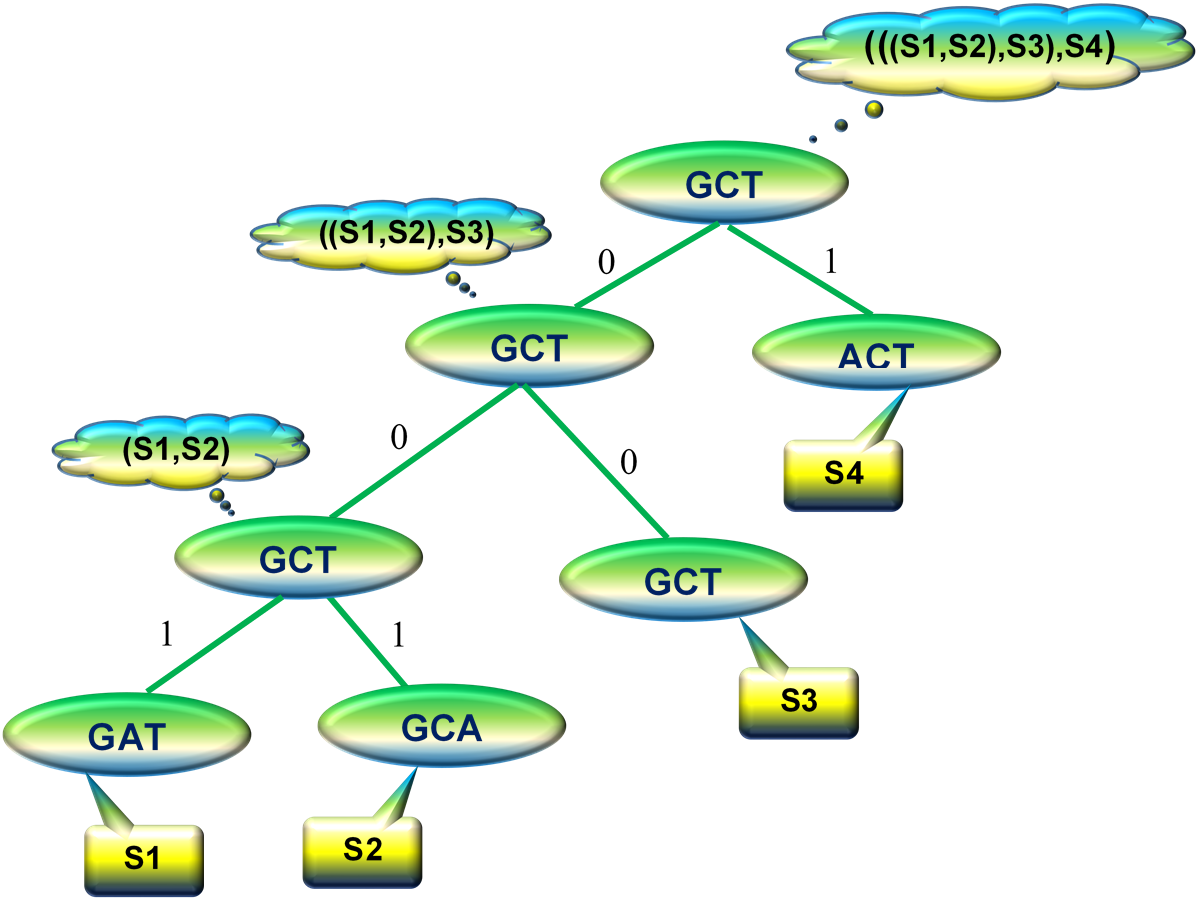
**TIẾN HÓA**

Theo học thuyết tiến hóa của Darwin, các loài sinh vật tiến hóa từ một tổ tiên chung. Quá trình tiến hóa của các loài sinh vật có thể được xác định bằng cách phân tích các chuỗi protein của chúng. Chuỗi protein là một đoạn gắn kết các axit amin. Có 20 loại axit amin khác nhau, mỗi axit amin được biểu diễn bởi một kí tự trong tập {**A**, **R**, **N**, **D**, **C**, **Q**, **E**, **G**, **H**, **I**, **L**, **K**, **M**, **F**, **P**, **S**, **T**, **W**, **Y**, **V**}. Độ dài của một chuỗi protein là số lượng các axit amin trong chuỗi đó.

Cho hai chuỗi protein và có cùng độ dài *k*, khoảng cách giữa hai chuỗi ***p*** và ***q*** được tính bằng số lượng vị trí mà khác . Ví dụ, khoảng cách giữa hai chuỗi protein ***p***=**GAT** và ***q***=**GCA** là 2.

Quá trình tiến hóa của các loài sinh vật có thể biểu diễn bởi một cây nhị phân có gốc. Giá trị mỗi nút của cây là chuỗi protein của sinh vật tương ứng với nút đó. Các nút lá chứa

chuỗi protein của các sinh vật hiện tại, các nút trong của cây chứa chuỗi protein của các sinh vật tổ tiên. Nếu và là hai chuỗi protein của hai loài sinh vật có chung một tổ tiên trực tiếp, thì tổ tiên chung của chúng được biểu diễn bởi *.* Biểu diễn của sinh vật tổ tiên tại nút gốc cho biết cấu trúc của cây (xem *Hình bên dưới*).



Giáo sư Haeseler muốn nghiên cứu quá trình tiến hóa của *n* loài sinh vật hiện tại, được biểu diễn bởi *n* chuỗi protein , S2, S3, . . ., . Các chuỗi protein đều có cùng độ dài là *k.* Một khó khăn trong quá trình nghiên cứu là Giáo sư không có thông tin về các chuỗi protein của các loài sinh vật tổ tiên nằm ở các nút bên trong của cây. Mục tiêu của Giáo sư là xác định chuỗi protein cho các loài sinh vật tổ tiên, để tổng độ dài các cạnh của cây tiến hóa là nhỏ nhất (độ dài cạnh nối hai loài sinh vật được tính bằng khoảng cách giữa hai chuỗi protein biểu diễn hai loài sinh vật đó).

***Dữ liệu:*** Vào từ file EVOLUTION.INP:

* Dòng đầu tiên chứa hai số nguyên dương *n* và *k* ( .
* Dòng thứ *i* trongsố *n* dòng tiếp theo chứa chuỗi protein có độ dài *k* biểu diễn cho sinh vật thứ *.*
* Dòng cuối cùng chứa xâu biểu diễn của sinh vật tổ tiên ở nút gốc của cây.

***Kết quả:*** Ghi ra file EVOLUTION.OUT một dòng duy nhất là tổng độ dài các cạnh của cây tiến hóa.

***Ví dụ:***

|  |  |  |
| --- | --- | --- |
| EVOLUTION.INP |  | EVOLUTION.OUT |
| **4 3**  **GAT**  **GCA**  **GCT**  **ACT**  **(((S1,S2),S3),S4)** |  | **3** |
|  |  |