Secuenciando el ADN

Contribución de Facundo Gutiérrez y Agustín Santiago Gutiérrez

Descripción del problema

Durante el siglo XX se produjo una enorme revolución en biología al descubrirse que la información genética de todos los seres vivos se almacena en cromosomas. Los cromosomas son enormes moléculas lineales de ADN formadas por una cadena de componentes llamanos nucleótidos que se encuentran unidos en un orden específico.

Un cromosoma puede describirse mediante una cadena de caracteres donde cada letra representa un nucleótido particular. Por ejemplo, para los seres vivos del planeta tierra hay 4 nucleótidos diferentes y suelen representarse con las letras A, C, G y T. Se denomina *secuenciación* del ADN, al proceso de descubrir esta cadena de caracteres.

En un laboratorio desconocido han desarrollado un peculiar instrumento de medición con el fin de obtener información para secuenciar un cromosoma. Dicho instrumento permite al usuario elegir una cadena de caracteres S totalmente arbitraria y responde con total certeza si la misma S es una *subsecuencia* de la cadena correspondiente al cromosoma.

Una cadena A es subsecuencia de una cadena B cuando es posible obtener A eliminando algunas letras de B sin alterar la ubicación del resto. Por ejemplo, `AGUA` es subsecuencia de `parAGUAs` y también de `inAuGUrAl`. En cambio, `AGUA` no es subsecuencia de `AUGA` ni de `ALGASU`. Siguiendo el mismo razonamiento: toda cadena es subsecuencia de sí misma.

En el laboratorio ya han determinado la cantidad total N de nucleótidos y cuáles son los K nucleótidos diferentes que pueden llegar a existir en el cromosoma. Sin embargo, necesitan tu ayuda: debes escribir una función que reciba N y una cadena s con los K nucleótidos posibles y que, mediante el uso del instrumento de medición, logre secuenciar el cromosoma de ADN encontrado. El puntaje dependerá de la cantidad de veces que se utilice el instrumento (ver Puntajes y Subtareas).

Como en el laboratorio tienen acceso a meteoritos exóticos para analizar, en una de las subtareas se recibirá ADN alienígena, con más de 4 nucleótidos diferentes (ver Puntajes y Subtareas).

Tarea e interacción

Este es un **problema interactivo**: Debes programar una función

`secuenciar(N : ENTERO, s : PALABRA)`

que devuelva una `PALABRA`, con la cadena correspondiente a la secuencia del ADN. Para lograr esta tarea, puedes llamar a la función:

`medir(cad : PALABRA) : ENTERO`

Que devolverá 1 si `cad` es subsecuencia de la cadena de ADN que debes descubrir, o 0 si no lo es.

Ejemplo de interacción

- Se recibe $N = 7$, $s = ABC$ (por lo tanto $K = 3$). El cromosoma en este ejemplo será *AABBABB*.
- Se llama `medir("C")`, que devuelve 0. Entonces el cromosoma no contiene ninguna *C*.
- Se llama `medir("AB")` y devuelve 1.
- Se llama `medir("BA")` y devuelve 1.
- `medir("AAA")` devuelve 1.
- `medir("AABBB")` devuelve 1.
- `medir("AABBBBB")` devuelve 0.
- `medir("AABBABB")` devuelve 1.
- Se devuelve como resultado *AABBABB*, que es correcto.

Evaluador local

El evaluador local lee los datos de entrada por consola en el siguiente orden:

- Una primera línea con la cadena s que se pasará a la función.
- Una segunda línea con la secuencia de ADN que se debe descubrir.

En un caso válido, la secuencia del ADN debería contener solamente caracteres de s, pero el evaluador no se encargará de verificarlo.

El evaluador llamará a la función `secuenciar` con la cadena s indicada, y con N igual a la longitud del ADN. Mostrará por pantalla un mensaje indicando si la función `secuenciar` retorna la secuencia de ADN correcta o no, así como la cantidad total de mediciones realizadas.

Puntuación y subtareas

En todas las subtareas, $N \leq 1000$ y s indica los posibles nucleótidos.

Cada subtarea se evalúa de manera separada. Si en algún caso no se responde la secuencia de ADN correcta, se obtienen cero puntos en la subtarea. De responder correctamente, se obtiene un puntaje que depende de la máxima cantidad de preguntas realizadas.

Las subtareas son:

- $K = 2$, $s = AC$, y se sabe que **todas las A aparecen antes que las C**
  - Hasta 10 usos : 5 puntos
  - Hasta 1100 usos : 2 puntos
- $K = 2$, $s = GT$ con $N = 10$:
  - Hasta 3000 usos : 4 puntos
- $K = 2$, $s = GT$
  - Hasta 1010 usos : 24 puntos
  - Hasta 1100 usos : 20 puntos
  - Hasta 2000 usos : 10 puntos
  - Hasta 3000 usos : 5 puntos
- $K = 4$, $s = ACGT$
  - Hasta 2300 usos : 35 puntos
  - Hasta 3010 usos : 25 puntos
  - Hasta 3100 usos : 15 puntos
  - Hasta 4000 usos : 5 puntos
- $K = 52$, s contendrá las 26 letras del alfabeto inglés, en sus versiones minúsculas y mayúsculas.
  - Hasta 11000 usos : 32 puntos
  - Hasta 45000 usos : 20 puntos
  - Hasta 51000 usos : 10 puntos