В последно време, покрай пандемията от Covid19 и новите ваксини срещу този вирус много често се говори за биология, протеини, мембранна пропускливост, РНК и какви ли не още интересни, но магически-сложно-звучащи неща. Всички те всъщност не са толкова сложни и са особено интересни в контекста на математиката и програмирането. Днес ще използваме възможността да научим повече неща в областта на биологията и по-точно как тялото ни произвежда по-малките частички, които ни правят да работим: белтъците. Белтъците представляват, най-просто казано, последователности от аминокиселини - 200, 300, много-много повече - и изпълняват най-разнообразни функции в нашето тяло. Някои от тях са по-скучни от други - чували сме, например, за колаген - той представлява структурен белтък, който поради сложни химични причини има необходимата форма и свойства да изгражда една голяма част от костите и кожата ни (яденето на неща с колаген обаче не е пряко свързано с това - тялото ви ще го разгради и ще използва аминокиселините, за да си направи други белтъци, евентуално пак колаген). Липсата на “лактаза” - друг белтък - характеризира хората, за които казваме, че имат “непоносимост към лактоза”. Лактазата е ензим, който подпомага разграждането на специфичната захар в млечните продукти - лактоза - на прости въглехидрати, които после тялото ни може да използва за енергия.

Тези и, разбира се, много други важни белтъци се произвеждат в нашите клетки по специални рецепти. Всички рецепти за синтез на различните протеини са събрани на едно място - днк молекулите. Те дават инструкциите за това как да бъдат създадени белтъците на далеч по-безславните малки органели (части от клетките ни) - рибозомите. Разбира се ДНК са огромни и синтезирането на отделен белтък става посредством копирани отрязъци от ДНК - по-малки подобни “записи с една рецепта” - РНК молекулите. Отрязъците, които копираме за улеснение наричаме “генетичен код” и представяме като последователности от буквите “A”, “T“, “G” и “C”, с различна дължина - GATTACATGCA е пример за една такава последователност. Тези букви представляват прости нуклеинови киселини и се наричат нуклеотидни бази. Части от този код в последствие се транскрибират до РНК - там “T” става “U”, но няма да се занимаваме детайлно с това. Самата РНК последователност се транслира (превежда, интерпретира) от рибозомите до последователност от аминокиселини, които изграждат белтъците. Самите аминокиселини са повече - 20-ина. Всяка от тези 20 аминокиселини може да се кодира с три от тези букви - за всяка аминокиселина отговарят по три “нуклеотидни бази” - буквичките A, C, T, G. Например CAG (както и CAA) кодира аминокиселината глутамин. Рецептата за синтез на всеки белтък в природата представлява последователност от нуклеотидни бази, започващи със специален старт кодон и завършващи със специален стоп кодон. Кодоните са просто последователности от три нуклеотидни бази (CAG, ATC, TTT, GCT, ...).

Ваша позната от биологическия факултет има за задача по даден отрязък ДНК с големина над 1000 нуклеотидни двойки, да намери дали в него се намират кодирани белтъци, обозначени с уникални номера. Информацията за белтъците е записана във файл. Също така тя трябва да намери и каква е аминокиселинната последователност, която им съответства. Но да се прави това на ръка е лудост! Трябва да ѝ помогнете! Все пак сме в сезона на добрите дела!

Понеже нейни приятели пък имат имат същата задача, но с различни отрязъци от днк, се налага усложняване на подхода. Иначе другите ще ви се сърдят… Така вие се окажете в следната ситуация:

Разполагате с три файла:

- един с дългия отрязък ДНК (последователност от буквите A, T, G и C;
- един със записани на различни редове белтъци от вида ID P, където ID е уникалният номер на белтъка (unsigned long), а P - последователността от нуклеотидни двойки, която го кодира (последователност от символите A, T, G, C);
- един с произволен брой записани на различни редове тройки кодони и аминокиселините във вида **C A**, където **C** е низ от три букви сред A, T, G, C, които кодират една амино киселина **A** (символен низ - името на аминокиселината. Тези тройки са винаги едни и същи и са стандартни, https://www.chemguide.co.uk/organicprops/aminoacids/dnacode.gif, но вашата програма трябва да работи с подадените от този файл.

Вашата задача е да напишете програма, която получава от стандартния вход три символни низа - пътищата към файловете, а също и параметър **Q**, който обозначава броят заявки, които ще бъдат направени. След това **Q** на брой числа, които представляват уникални номера на белтъци.

Като изход за всеки номер на белтък трябва да изведете дали той се намира в отрязъка от ДНК от първия файл и ако да, на коя позиция започва и каква е аминокиселинната последователност, която му съответства.

Задачата не е нова за тази област на информатиката и вероятно никой няма да приеме патента за програмата ви, но някога това е било голяма работа. И в биологическия факултет ще са ви признателни.

Допълнителни изисквания от нас (не трябва да се излагаме, все пак, трябва да защитите доброто име на ФМИ):

- Изисква се сложност по време за всяка заявка O(K + log(N)) или по-малко, където K е дължината на белтъка, който търсим.
- Допуска се извършването на допълнителна предварителна операция, която да ви помогне за заявките. За нея няма ограничение в сложността по време.

Следният файл дава аминокиселината, която съответства на всеки кодон: https://drive.google.com/file/d/1INsFYHnrNR3oclu6sR2_SG1TXaRcBms2/view?usp=sharing

"*" означава "стоп кодон" - той е края на последователността за всеки белтък.

Пример:

Вход:

dna_sequence.txt
proteins.txt
codonToAminoacids.txt
4
34
12
2
52

Изход:

No protein in proteins.txt with id 34
Yes 283 MLLGSFRLIPKETLIQVAGSSPCNLS
No
Yes 731 MTPRLGLESLLE

Файловете, излолзвани в примера, можете да намерите тук.