pE.DNA 定序

Problem ID: NGS

次世代定序技術（*Next Generation Sequencing*, 簡稱 NGS）的到來，使人類能快速且精準的完成 DNA 定序的工作。舉例而言，人體的全 DNA 定序能在一天內就完成，相比之下，最初的「人類基因體計畫」可花了約十年才結束。

不過，從 DNA 定序儀器所得到的原始定序結果不一定會是完全正確無誤的，因此需先經過簡單的篩選，把不良好的片段刪除，只留下良好的片段才能在後續研究時有更高的精準度。此題即是請你幫忙篩選原始的定序結果。

Phred quality score 是用來標示定序出來每個鹼基對的「品質」，左下方的公式是用來計算 *Phred quality score* 的，其中 P 是該鹼基對錯誤的機率；右下方的表格是一個概略的對照表，代表 *Phred quality score* 所對應到的錯誤與正確率。

$$Q = -10\,log_{10}P$$

Score	錯誤率	精準度
10	1/10	90 %
20	1/100	99 %
30	1/1000	99.9 %
40	1/10000	99.99 %

Phred quality score 會是篩選片斷時的其中一個依據，而篩選片段的方式主要有下列四種：

1. adapter filter
   在做 DNA 定序時，會在 DNA 片段兩端接上 *adapter*，用於將片段 load 到定序儀的特殊平台上。如果一個定序出來的片段中有包含 *adapter* 的 DNA 序列 (此題中，等價於此片段有一子字串為 adapter)，這個片段就需要被刪除。

2. polyN filter
   如果一個片段的多樣性太差，也就是有不少於 85% 的鹼基對是 A,C,T,G 的同一種，那這個片段就需要被刪除。

3. s35 filter
   由於在做定序時，通常越往片段尾端品質會越差，因此可以只考慮此片段前段的品質。具體來說，如果此片段的長度不到 35 個鹼基對，或是前

35 個鹼基對只有不到 25 個鹼基對的 *Phred quality score* 大於等於 30，那這個片段就需要被刪除。

4. Q20 filter
   理由同第 3 點，我們可以改成將品質不好的尾端修剪掉，也就是從尾端往前看，找到第一個 *Phred quality score* 不小於 20 的鹼基對，把那個位置之後 (不含那個鹼基對) 的部份刪除。如果整個片段的 *Phred quality score* 都小於 20 的話，整個片段都會被刪除。

DNA 定序儀器會將 DNA 的定序結果以一種稱作 *FASTQ* 的檔案格式表示。在 *FASTQ* 檔案中，每四行代表定序出的一個片段，例如：

```
@An example
GACTGTGCTTCTGAGCCCCAGAAGGTCATTAATGCAATGT
+
gabhcbdfb[eef__abdc^fhchfXhgcgb_bbhgbeba
```

其中，每一行所代表的意義如下：

- 第一行是以'@' 符號開頭，後面接的是有關這個片段的資訊和註解
- 第二行是這個片段的定序結果 (也就是由 A,C,T,G 四種鹼基對構成)
- 第三行是以'+' 符號開頭，後面會選擇性地複製第一行的註記
- 第四行是一串與定序結果等長的符號，代表片段裡每一個鹼基對的 *Phred quality score*。編碼方式是以'@'(Ascii $= 64$) 當作 Score=0，依 Ascii code 順序 `@,A,B,C,D, ...,f,g,h`，最後一個是'h'(Ascii $= 104$)，代表 Score $= 40$。

現在，給你一份用 FASTQ 格式表示的定序結果，和所要依序執行的篩選 filter，請幫忙計算篩選後剩下片斷的長度和是多少吧！

– 輸入說明 –

輸入的第一行有一個整數 N，代表要使用的篩選 filter 有幾個。接下來 N 行，每行有一個字串 F_i，代表**依序**要使用的 filter。如果 $F_i = "adapter"$，則該行會有第二個字串 S, 代表 adapter 的 DNA 序列。從第 $N + 2$ 行開始都是要被篩選的定序資料，輸入到 EOF 為止。

- $0 \leq N \leq 4$
- $F_i \in \{"adapter", "polyN", "s35", "Q20"\}$
- F_i 皆相異
- $1 \leq |S| \leq 100$
- 定序資料符合 FASTQ 格式，保證行數是 4 的正整數倍
- 定序資料保證不含有空格
- 定序資料總字元數不超過 2×10^5

– 輸出說明 –

輸出只有一個整數，代表經過所給定的 filter 篩選以後，剩下片段的長度總和。

– 範例輸入輸出 –

輸入	輸出
1 adapter GCC @example1.1 TTTTGCCCAT + hcdagf\\bg @example1.2 CAGGGGAA + hg^bgchc	8
1 Q20 @example2 AGGCCAGGTT + DVMCABOE@B	2

輸入	輸出
1 polyN @example3.1 AAAAAAAAAC + _g^_edfegg @example3.2 AAAAATATAAC + Whh_WTfdedh	11
1 s35 @example4.1 GCCTAAAGGTTTGTGTGGCCGCATAGTCGCTCTGA + VLhYV_JOLh\DfV^XeYfZSZ]hXDCKNc]X‘[^ @example4.2 GTGATTACCCACTATTCTTCACATAGGCACAGGGATA + ‘Udhac\‘c[g\‘e\]\^cf_hSeTdfbT‘dggbhgh	37

– 配分 –

分數	說明
18 %	$F_i = "adapter"$
18 %	$F_i = "polyN"$
18 %	$F_i = "s35"$
18 %	$F_i = "Q20"$
28 %	*Original constraints.*

– 範例測資說明 –

第 1,3,4 筆範測中被刪除的都是第一個片斷，第 2 筆測資中第 2 個位置 (score=22) 之後的鹼基對都被刪除。