پروفایل

- محدودیت زمان: 3 ثانیه
- محدودیت حافظه: 256مگابایت

در این تمرین قرار است پروفایل یک MSA را بسازید و با استفاده از آن پروفایل از یک دنباله‌ی طولانی، آن بخشی که بیشترین شباهت به MSA را دارد پیدا کنید.

برای ساخت پروفایل کافی است مثل ساخت PSSM جلو برید، با این تفاوت که Gap ها را هم در نظر می‌گیرید. برای جلوگیری از log(0) در این پروفایل از pseudocount استفاده کنید (pseudocount را 2 بگذارید). بعد از ساخت پروفایل روی رشته‌ی طولانی داده شده دنبال بهترین زیر رشته که با توجه به پروفایل بالاترین امتیاز را دارد بگردید.

دقت داشته باشید که در این مسئله در جواب شما میتواند Gap هم وجود داشته باشد.

برای ساخت پروفایل فقط از آمینو اسیدهای رشته‌های داده شده استفاده کنید. یعنی سطرهای جدول فقط شامل حروف دیده شده در MSA و Gap باید باشند.

ورودی

در خط اول عدد N را می‌گیرید که نشان دهنده‌ی تعداد رشته‌های درون MSA است (حداکثر طول رشته‌های درون MSA ده است). در N خط بعدی رشته‌های درون MSA را خواهید گرفت. در خط آخر رشته‌ی طولانی‌ای را که باید درونش مرتبط‌ترین زیر رشته را بیابید قرار دارد (حداکثر به طول 100 است و حروفش از حروف درون N رشته‌ی بالایی است)

خروجی

در این بخش کافی است زیر رشته‌ای از رشته‌ی طولانی که، با توجه به پروفایل، بالاترین امتیاز را دارد چاپ کنید (تست‌کیس‌ها به گونه‌ای است که فقط یک جواب درست وجود دارد).

مثال

در اینجا چند نمونه برای فهم بهتر صورت سوال و قالب ورودی و خروجی تست‌ها داده می‌شود.

ورودی نمونه ۱

```
4
HVLIP
H-MIP
HVL-P
LVLIP
LIVPHHVPIPVLVIHPVLPPHIVLHHIHVHIHLPVLHIVHHLVIHLHPIVL
```

خروجی نمونه ۱

```
H-L-P
```

برای اطمینان از پروفایل خود، پروفایل این سوال در اینجا قرار گرفته است(اعداد در log_2 هستند):

Aminos	1	2	3	4	5
H	0.943	-0.378	-0.378	-0.378	-0.378
V	-0.378	0.943	-0.378	-0.378	-0.378
L	0.099	-0.485	0.836	-0.485	-0.485
I	-0.378	-0.378	-0.378	0.943	-0.378
P	-0.485	-0.485	-0.485	-0.485	1.099
M	-0.137	-0.137	0.447	-0.137	-0.137
-	-0.263	0.321	-0.263	0.321	-0.263

به درخواست یکی از دانشجویان، چند مرحله‌ی میانی محاسبه‌ی پروفایل در اینجا قرار گرفته است:

- تعداد تکرارهای هر آمینو در هر ستون + 2 (pseudocount)

Aminos	1	2	3	4	5
H	5	2	2	2	2
V	2	5	2	2	2
L	3	2	5	2	2
I	2	2	2	5	2
P	2	2	2	2	6
M	2	2	3	2	2
-	2	3	2	3	2

- هر ستون را بر (تعداد رشته‌ها + 2(pseudocount) ضرب در 7(تعداد آمینوها و Gap)) تقسیم می‌کنیم. (در اصل داریم هر ستون را بر جمع اعداد آن ستون تقسیم می‌کنیم)

Aminos	1	2	3	4	5
H	0.277	0.111	0.111	0.111	0.111
V	0.111	0.277	0.111	0.111	0.111
L	0.166	0.111	0.277	0.111	0.111
I	0.111	0.111	0.111	0.277	0.111
P	0.111	0.111	0.111	0.111	0.333
M	0.111	0.111	0.166	0.111	0.111

Aminos	1	2	3	4	5
-	0.111	0.166	0.111	0.166	0.111

- هر سطر را بر (جمع آن سطر تقسیم بر 5(تعداد ستون)) تقسیم می‌کنیم (تقسیم بر Overall Frequency)

Aminos	1	2	3	4	5
H	1.923	0.769	0.769	0.769	0.769
V	0.769	1.923	0.769	0.769	0.769
L	1.071	0.714	1.785	0.714	0.714
I	0.769	0.769	0.769	1.923	0.769
P	0.714	0.714	0.714	0.714	2.142
M	0.909	0.909	1.363	0.909	0.909
-	0.833	1.249	0.833	1.249	0.833

- در آخر با log_2 گرفتن از اعداد این جدول، به جدول اولی می‌رسیم

ورودی نمونه ۲

```
4
T-CT
--CT
A-CT
ATCT
ATCCTATATCTTCTCTATACTATCCTTCA
```

خروجی نمونه ۲

A-CT

اشکالات خود را میتوانید به ایمیل مقابل ارسال کنید: mmnafar57@gmail.com