دانشگاه صنعتی امیرکبیر
(پلیتکنیک تهران)
دانشکده مهندسی کامپیوتر

درس بیوانفورماتیک
تمرین چهارم۴

امیرمهدی زرین‌نژاد
۹۷۳۱۰۸۷

نیم‌سال دوم ۱۴۰۰

سوال ۱)

الف)

می‌توان از ماتریس blosum استفاده مرد. به این صورت که برای هر کاراکتر رشته داده‌شده، مقادیر ماتریس بلوسام مربوط به آن کاراکتر را قرار دهیم(ستون امتیازات مربوط به آن کاراکتر در ماتریس بلوسام). به این صورت ماتریس از امتیازات بلوسام کاراکترها پر می‌شود.

ب) profile drift:

یک مشکلی است که در psi blast می‌تواند رخ بدهد. و آن اینست که چون روند به صورت اتوماتیک انجام میشود و پروفایل‌ها به صورت خودکار ساخته می‌شوند، اگر در مرحله‌ای اشتباه صورت گیرد و توالی‌های نامناسب انتخاب شوند(تعدادی سیکوینس که هومولوگ نیستند اضافه شوند-سیکوئنس false-positive داشته باشیم) و یا پروفایلی غلط تولید شود، این مشکل بسط پیدا می‌کند و در پیمایش‌های بعدی نیز باقی می‌ماند و تجمیع و بیشتر می‌شود و هی سیکوئنس‌های مشابه آن سیکوئنس‌های غلط اضافه می‌شود. نهایتا این اتفاق باعث تجمیع خطا و low selectivity می‌شود.

ج)

با کاهش آستانه در قسمت جست و جوی این الگوریتم، مشکل profile drift بیشتر می‌شود. زیرا اگر حد آستانه را کاهش دهیم، حساسیت برای میزان هومولوگ رشته‌های انتخاب شده کم‌تر می‌شود و امکان اضافه شدن توالی‌های FP (توالی غیر همولوگ) و متفاوت‌تر بیش‌تر می‌شود. در نتیجه مشکل profile drift شدتش بیش‌تر می‌شود.

سوال ۲)

الف) ماتریس PSSM

Pseudocount = 0, background prob = randome chance = 0.25, تا ۳ رقم اعشار

ATGCCG
AAGATT
TACTCA
CTGAGG
CACCTG

	1	2	3	4	5	6	Overall freq
A	2/5	3/5	-	2/5	-	1/5	8/30
C	2/5	-	2/5	2/5	2/5	-	8/30
G	-	-	3/5	-	1/5	3/5	7/30
T	1/5	2/5	-	1/5	2/5	1/5	7/30

Normalizing: values/background prob(0.25)

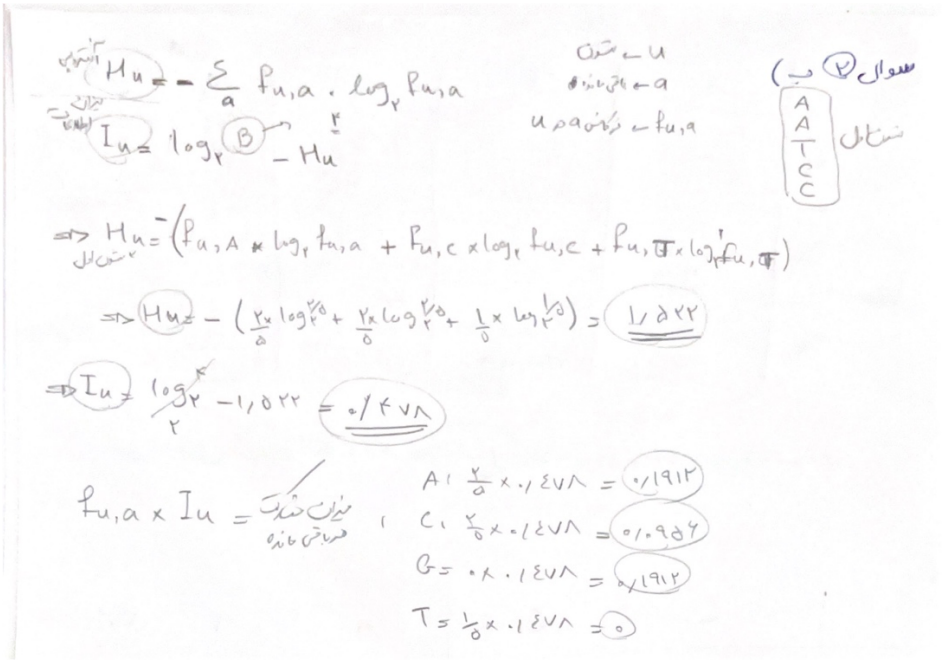
	1	2	3	4	5	6	Overall freq
A	1.6	2.4	-	1.6	-	0.8	8/30
C	1.6	-	1.6	1.6	1.6	-	8/30
G	-	-	2.4	-	0.8	2.4	7/30
T	0.8	1.6	-	0.8	1.6	0.8	7/30

Log2

	1	2	3	4	5	6	Overall freq
A	0.678	1.263	-	0.678	-	-0.322	8/30
C	0.678	-	0.678	0.678	0.678	-	8/30
G	-	-	1.263	-	-0.322	1.263	7/30
T	-0.322	0.678	-	-0.322	0.678	-0.322	7/30

ب) logo

تنها برای ستون اول دنباله‌ها



سوال ۳)

الف) Profile

Pseudocount = 1, background = overall freq, تا ۳رقم اعشار

AT - G - CCG
AA - G - CTT
T - ACT - CA
CTGACGGA

	1	2	3	4	5	6	7	8	Overall freq
A	2/4	1/4	1/4	1/4	-	-	-	2/4	7/32
C	1/4	-	-	1/4	1/4	2/4	2/4	-	7/32
G	-	-	1/4	2/4	-	1/4	1/4	1/4	6/32
T	1/4	2/4	-	-	1/4	-	1/4	1/4	6/32
-	-	1/4	2/4	-	2/4	1/4	-	-	6/32

Update with pseudocount:

(Values+pseudocount)/(N + B*pseudocount), N=4, B=5=4+1(gap)

	1	2	3	4	5	6	7	8	Overall freq
A	2+1/4+5	1+1/4+5	1+1/4+5	1+1/4+5	0+1/4+5	0+1/4+5	0+1/4+5	2+1/4+5	7+8/32+5*40 = 15/72
C	1+1/4+5	0+1/4+5	0+1/4+5	1+1/4+5	1+1/4+5	2+1/4+5	2+1/4+5	0+1/4+5	7+8/32+5*8 = 15/72
G	0+1/4+5	0+1/4+5	1+1/4+5	2+1/4+5	0+1/4+5	1+1/4+5	1+1/4+5	1+1/4+5	6+8/32+5*8 = 14/72
T	1+1/4+5	2+1/4+5	0+1/4+5	0+1/4+5	1+1/4+5	0+1/4+5	1+1/4+5	1+1/4+5	6+8/32+5*8 = 14/72
-	0+1/4+5	1+1/4+5	2+1/4+5	0+1/4+5	2+1/4+5	1+1/4+5	0+1/4+5	0+1/4+5	6+8/32+5*8 = 14/72

=

	1	2	3	4	5	6	7	8	Overall freq
A	0.333	0.222	0.222	0.222	0.111	0.111	0.111	0.333	0.208
C	0.222	0.111	0.111	0.222	0.222	0.333	0.333	0.111	0.208
G	0.111	0.111	0.222	0.333	0.111	0.222	0.222	0.222	0.194
T	0.222	0.333	0.111	0.111	0.222	0.111	0.222	0.222	0.194
-	0.111	0.222	0.333	0.111	0.333	0.222	0.111	0.111	0.194

Normalize with overall freq:

	1	2	3	4	5	6	7	8	Overall freq
A	1.601	0.534	0.534	0.534	1.067	1.067	1.067	1.601	0.208
C	0.534	1.601	1.601	1.067	1.067	0.534	0.534	1.067	0.208
G	1.144	1.144	1.144	0.572	1.716	1.144	0.572	0.572	0.194
T	1.144	1.144	0.572	1.144	0.572	0.572	1.716	1.144	0.194
-	0.572	0.572	1.144	1.716	0.572	1.716	1.144	0.572	0.194

Log2

	1	2	3	4	5	6	7	8	Overall freq
A	0.679	-0.905	-0.905	-0.905	0.094	0.094	0.094	0.679	
C	-0.905	0.679	0.679	0.094	0.094	-0.905	-0.905	0.094	
G	0.194	0.194	0.194	-0.805	0.779	0.194	-0.805	-0.805	
T	0.194	0.194	-0.805	0.194	-0.805	-0.805	0.779	0.194	
-	-0.805	-0.805	0.194	0.779	-0.805	0.779	0.194	-0.805	

ب) احتمال دنباله AA-CTCTG

0.679+0.094+0.779+0.094+0.194+0.679+0.194+0.194 = 2.907

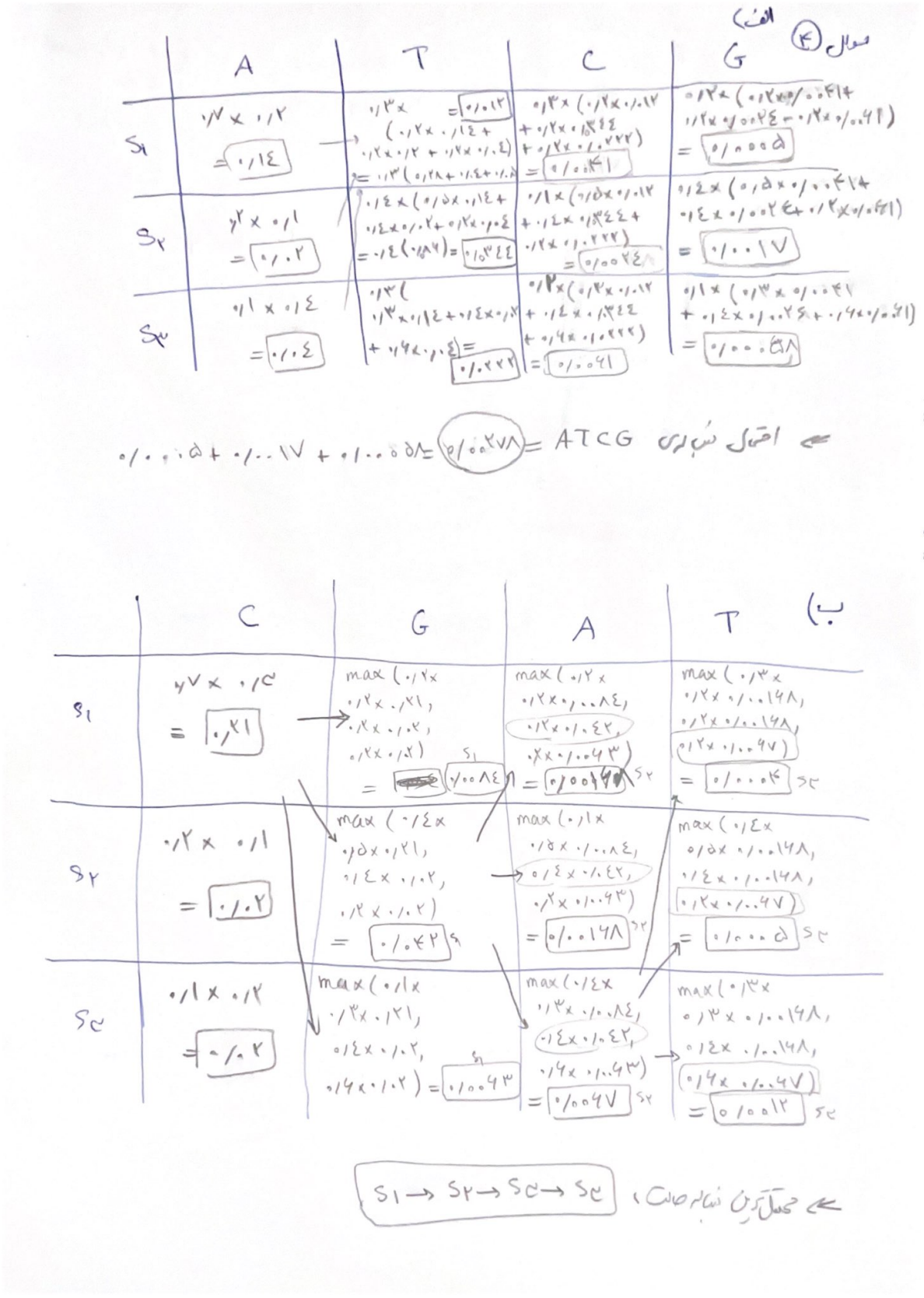
یعنی احتمال این دنباله، 2^2.907 = 7.5 تقریبا ۷ و نیم برابر رندم چنس است.

ج) محتمل‌ترین دنباله

در هر ستون(جایگاه در دنباله) کاراکتری انتخاب می‌شود که بیش‌ترین مقدار امتیاز را دارد.

AT-G-CCA

سوال ۴)



ماکسمم مقدار در ستون آخر مربوط به s3 است(0.0012) که اگر مسیرش به عقب را دنبال کنیم محتمل‌ترین دنباله حالت را می‌یابیم.(مسیرها توسط فلش مشخص شده‌اند همچنین استیت‌های مربوط به مقادیر ماکسیمم در جلویشان آمده‌است)