به نام خداوند بخشنده و مهربان

درس مقدماتی بر بیوانفورماتیک

تمرین چهارم

استاد درس: دکتر زینعلی

نام دانشجو:

روزبه قاسمی ۹۵۳۱۴۲۴

تیر ۱۳۹۹

سوال اول

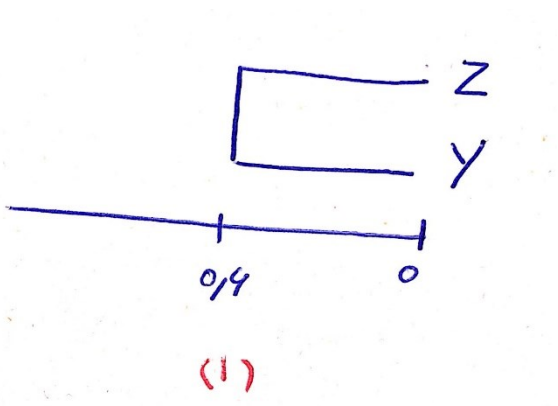
ابتدا جدول را بازنمایی می‌کنیم(برای سادگی ضربدر ۱۰۰ می‌کنیم) و طبق فرضی که استاد در کانال فرمودند قطر را در نظر نمی‌گیریم که برای راحتی صفر می‌کنیم:

	Q	Z	W	Y	C
Q	0				
Z	1.5	0			
W	1.8	1.4	0		
Y	1.6	1.2	1.5	0	
C	44.9	43.1	42.5	43.5	0

الف)

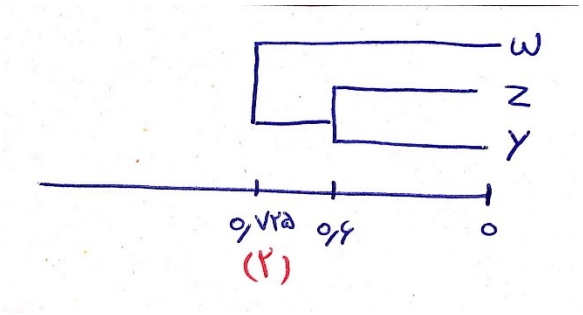
برای انجام UPGMA باید کوچکترین عدد جدول بالا را در نظر بگیریم و سپس سطر و ستون مربوطه را ادغام کنیم.

	Q	Z	W	Y	C
Q	0				
Z	1.5	0			
W	1.8	1.4	0		
Y	1.6	1.2	1.5	0	
C	44.9	43.1	42.5	43.5	0



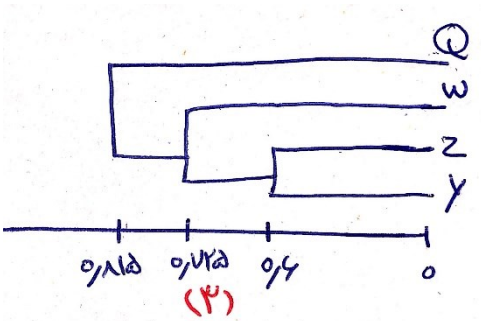
پس باید Z و Y را باهم ادغام شوند:

	Q	YZ	W	C
Q	0			
YZ	1.55	0		
W	1.8	1.45	0	
C	44.9	43.3	42.5	0



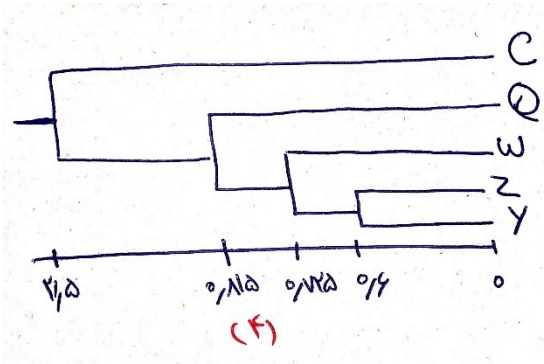
حال پس باید W و YZ را باهم ادغام شوند:

	Q	YZW	C
Q	0		
YZW	1.63	0	
C	44.9	43.04	0



حال پس باید YZW و Q را باهم ادغام شوند:

	YZWQ	C
YZWQ	0	
C	43.5	0



ب)

برای حل این مسئله به روش Nj داریم که(ابتدا یکبار دیگر جدول اولیه را بازنویسی می‌کنیم):

	Q	Z	W	Y	C
Q	0				
Z	1.5	0			
W	1.8	1.4	0		
Y	1.6	1.2	1.5	0	
C	44.9	43.1	42.5	43.5	0

ابتدا باید R را برای هر تاکسون حساب کرد:

for taxon Q:

$$r_Q = QZ + QW + QY + QC = 1.5 + 1.8 + 1.6 + 44.9 = 49.8$$

$$r'_Q = \frac{r_Q}{5-2} = \frac{49.8}{3} = 16.6$$

for taxon Z:

$$r_Z = ZQ + ZW + ZY + ZC = 1.5 + 1.4 + 1.2 + 43.1 = 47.2$$

$$r'_Z = \frac{r_Z}{5-2} = \frac{47.2}{3} = 15.733$$

for taxon W:

$$r_W = WQ + WZ + WY + WC = 1.8 + 1.4 + 1.5 + 42.5 = 47.2$$

$$r'_W = \frac{r_W}{5-2} = \frac{47.2}{3} = 15.733$$

for taxon Y:

$$r_Y = YQ + YZ + YW + YC = 1.6 + 1.2 + 1.5 + 43.5 = 47.8$$

$$r'_Y = \frac{r_Y}{5-2} = \frac{47.8}{3} = 15.933$$

for taxon C:

$$r_C = QC + CZ + CW + CY = 44.9 + 43.1 + 42.5 + 43.5 = 174$$

$$r'_C = \frac{r_C}{5-2} = \frac{174}{3} = 58$$

حال باید فاصله تغییر یافته هر جفت را بیابیم:

$$d'_{QZ} = d_{QZ} - \frac{r_Q + r_Z}{2} = 1.5 - \frac{49.8 + 47.2}{2} = 1.5 - 48.5 = -47$$

$$d'_{QW} = d_{QW} - \frac{r_Q + r_W}{2} = 1.8 - \frac{49.8 + 47.2}{2} = 1.8 - 48.5 = -46.7$$

$$d'_{QY} = d_{QY} - \frac{r_Q + r_Y}{2} = 1.6 - \frac{49.8 + 47.8}{2} = 1.6 - 48.8 = -47.2$$

$$d'_{QC} = d_{QC} - \frac{r_Q + r_C}{2} = 44.9 - \frac{49.8 + 174}{2} = 44.9 - 111.9 = -67$$

$$d'_{ZW} = d_{ZW} - \frac{r_Z + r_W}{2} = 1.4 - \frac{47.2 + 47.2}{2} = 44.9 - 47.2 = -45.8$$

$$d'_{ZY} = d_{ZY} - \frac{r_Z + r_Y}{2} = 1.2 - \frac{47.2 + 47.8}{2} = 1.2 - 47.5 = -46.3$$

$$d'_{ZC} = d_{ZC} - \frac{r_Z + r_C}{2} = 43.1 - \frac{47.2 + 174}{2} = 44.9 - 110.6 = -67.5$$

$$d'_{WY} = d_{WY} - \frac{r_W + r_Y}{2} = 1.5 - \frac{47.2 + 47.8}{2} = 1.5 - 47.5 = -46$$

$$d'_{WC} = d_{WC} - \frac{r_W + r_C}{2} = 42.5 - \frac{47.2 + 174}{2} = 42.5 - 110.6 = -68.1$$

$$d'_{YC} = d_{YC} - \frac{r_Y + r_C}{2} = 43.5 - \frac{47.8 + 174}{2} = 43.5 - 110.4 = -67.4$$

	Q	Z	W	Y	C
Q	0				
Z	-47	0			
W	-46.7	-45.8	0		
Y	-47.2	-46.3	-46		
C	-67	-67.5	-68.1	-67.4	0

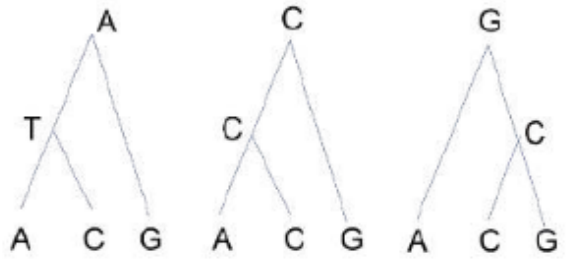
به دلیل حجم محاسبات زیاد در ادامه مطلب، این سوال را نصفه انجام دادم اما تا اینجا صحیح است.

ج)

همانطور که اشاره کردم، قسمت ب به طور کامل انجام نشد و درخت به دست نیامد اما بد نیست پاسخ این سوال را بر اساس شهود و قواعدی که در درس خواندیم پاسخ دهم. می‌دانیم که روش UPGMA از فرضیه ساعت مولکولی استفاده می‌کند و همانطور که خواندیم این که یک فرض غلط است به این علت که سرعت انتشار شاخه‌ها در دنیای واقعی یکسان نیست و در نتیجه درختی که ایجاد می‌شود در بسیاری از اوقات اشتباه است. این مشکل در الگوریتم NJ بر طرف شده است، به گونه‌ای که NJ با استفاده از یک گام تبدیل، نرخ‌های تکاملی نابرابر بین توالی‌ها را اصلاح می‌کند. پس در نتیجه درخت حاصل از الگوریتم UPGMA با درخت حاصل از الگوریتم NJ به احتمال زیاد نباید یکسان باشد.

سوال دوم

From\ To	A	C	G	T
A	0.6	0.1	0.1	0.2
C	0.2	0.6	0.1	0.1
G	0.1	0.1	0.7	0.1
T	0.1	0.2	0.1	0.6



اگر از سمت راست به چپ، درخت هارو مرتب کنیم. امتیاز درخت اول برابر است با:

$$L_1 = Pr(G \to C) * Pr(G \to A) * Pr(C \to G) * Pr(C \to C)$$

حال برای سادگی محاسبه از طرفین Ln می‌گیریم.

$$lnL_1 = ln\,Pr(G \to C) + lnPr(G \to A) + lnPr(C \to G) + ln\,Pr(C \to C)$$

$$lnL_1 = (-2.302) + (-2.302) + (-2.302) + (-0.510) = -7.416$$

امتیاز درخت دوم(درخت وسطی) برابر است با:

$$L_2 = Pr(C \to C) * Pr(C \to G) * Pr(C \to A) * Pr(C \to C)$$

حال برای سادگی محاسبه از طرفین Ln می‌گیریم.

$$lnL_2 = ln\,Pr(C \to C) + lnPr(C \to G) + lnPr(C \to A) + ln\,Pr(C \to C)$$

$$lnL_2 = (-0.510) + (-2.302) + (-2.302) + (-0.510) = -5.624$$

امتیاز درخت سوم(درخت سمت چپ) برابر است با:

$$L_3 = Pr(C \to C) * Pr(C \to G) * Pr(C \to A) * Pr(C \to C)$$

حال برای سادگی محاسبه از طرفین Ln می‌گیریم.

$$lnL_3 = ln\,Pr(A \to T) + lnPr(A \to G) + lnPr(T \to A) + ln\,Pr(T \to C)$$

$$lnL_3 = (-2.302) + (-2.302) + (-1.609) + (-2.302) = -8.515$$

طبق نتایج بدست‌آمده، درخت وسط، درخت ML مسئله است.

سوال سوم

بعد از ساخت درخت فیلوژنتیکی، باید توسط یک روش صحت آن مورد ارزیابی قرار گیرد. برای اینکه قابل اعتماد بودن درخت را بررسی کنیم، از استراتژی‌های تحلیلی نمونه‌برداری مجدد مانند خود راه‌اندازی[1] استفاده می‌کنیم.

خود راه‌اندازی، تکنیکی آماری برای سنجش خطاهای نمونه‌برداری درخت فیلوژنتیکی است.

در این سوال نیز فرض شده است که یک هم‌ترازی مولکولی شامل ۳۰۰ سایت برای ۷ گونه مختلف داریم. و درخت حداکثر شانس[2] یک کلاد از گونه‌های ۱ تا ۳ دارد. در روش خود راه‌اندازی یک سری توالی داریم که درخت فیلوژنتیکی از آن ساخته می‌شود. ما برای اینکه پی ببریم درخت درست ساخته شده است یا نه، ما از ۱۰۰ تا ۱۰۰۰ بار، همانند سازی به صورت تصادفی انجام می‌دهیم. در واقع یک نمونه‌گیری تصادفی از بعضی ستون‌ها می‌گیریم و یک همترازی جدید مصنوعی ایجاد می‌کنیم. به ازای هر هم‌ترازی یک درخت ایجاد می‌کنیم و ۱۰۰ تا ۱۰۰۰ درخت ایجاد می‌شود. سپس بر اساس قاعده اکثریت درخت‌مان را مشخص می‌کنیم و به یک درخت توافقی می‌رسیم که یک سری درصد را شامل می‌شود که مقدارهای خود راه‌انداز است. اگر این درصدها بالای ۷۰ درصد باشد، درخت ترسیم شده معنی دار خواهد بود.

[1] Bootstraping
[2] Maximum Likelihood

سوال چهارم

الف)

طبق فرمول Jukes-cantor داریم که:

$$d_{S1S2} = -\frac{3}{4}\, ln\left(1-\frac{4}{3}\, P_{S1S2}\right) = -\frac{3}{4}\, ln\left(1-\frac{4}{3} * 0.1\right) = 0.10732$$

ب)

مانند الف، طبق فرمول Jukes-cantor داریم که:

$$d_{S2S3} = -\frac{3}{4}\, ln\left(1-\frac{4}{3}\, P_{S1S2}\right) = -\frac{3}{4}\, ln\left(1-\frac{4}{3} * 0.1\right) = 0.10732$$

ج)

اگر بخواهیم بدون محاسبه فرمول ژوکز-کانتور در نظر بگیریم، ما انتظار داریم که حاصل جمع d_{S1S2} و d_{S2S3} چون هرکدام ۱۰ درصد تغییر می‌کنند، جمع آنها باید برای تغییر از S1 به S3 باید ۰.۲۱۴۶۴ باشد. حال اگر بخواهیم با فرمول بدست آوریم، طبق فرمول Jukes-cantor داریم که:

$$d_{S1S3} = -\frac{3}{4}\, ln\left(1-\frac{4}{3}\, P_{S1S3}\right) = -\frac{3}{4}\, ln\left(1-\frac{4}{3} * 0.2\right) = 0.2326$$

همانطور که مشخص است، بیشتر از انتظار ماست. این به این علت است که فرمول ژوکز-کانتور غیرخطی است!

پایان