پاسخ سوالات تیوری پروژه‌ی مبانی بیوانفورماتیک

میلاد آقاجوهری ، احسان سلطان‌آقایی

۸ بهمن ۱۳۹۶

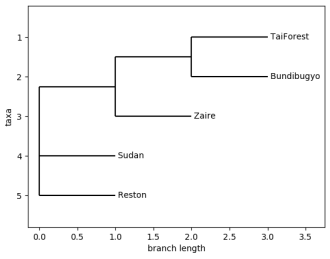
۱ سوال اول

۲ قسمت سوم

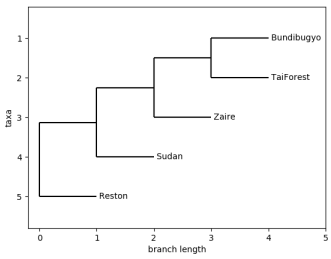
۱.۲ تشکیل درخت زندگی برای هر ژن

۱.۱.۲ مقایسه‌ی نتایج الگوریتم برای ۵ ژن

در ابتدا تمامی نتایج حاصله را روبروی هم مشاهده می‌کنیم:

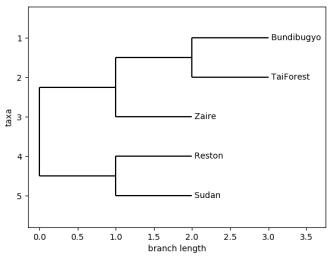


NP_NJ.png (آ)

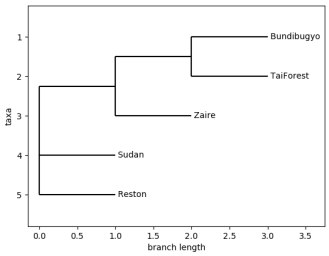


NP_UPGMA.png (ب)

برای این ژن درخت‌های حاصل یکی هستند.

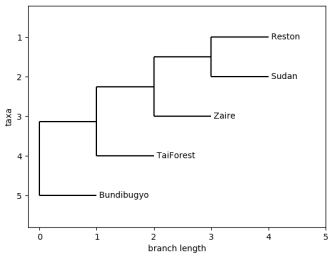


VP۳۵_UPGMA.png (ب)

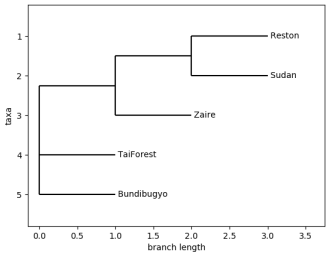


Vp۳۵_NJ.png (آ)

برای این ژن درخت‌های حاصل یکی هستند(در واقع یک مفهوم را میرسانند).

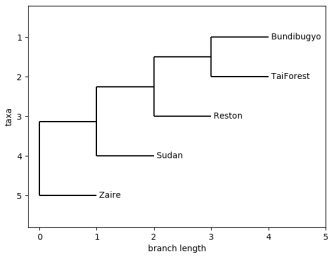


VP۴۰_UPGMA.png (ب)

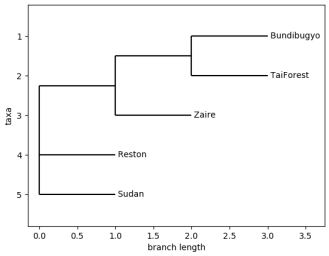


Vp۴۰_NJ.png (آ)

برای این ژن درخت‌های حاصل تفاوت جزیی دارند، البته معنایشان تقریبا یکسان است، صرفا UPGMA اطلاعات بیشتری داده است.

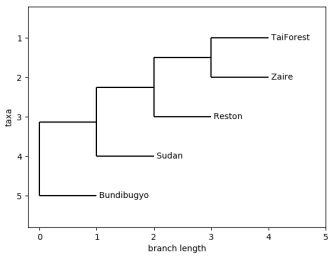


GP_UPGMA.png (ب)

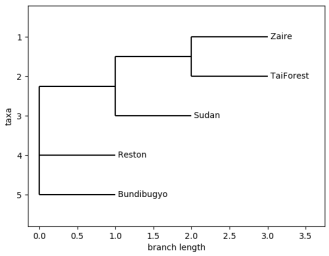


GP_NJ.png (آ)

درخت‌های حاصل در این جا بسیار متفاوت هستند و تنها در مورد وTaiForest Bundibugyo یکسان هستند.

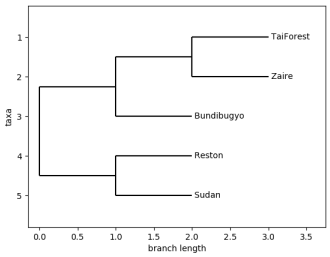


VP۳۰_UPGMA.png (ب)

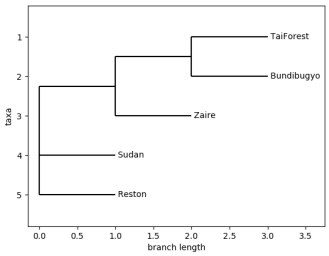


VP۳۰_NJ.png (آ)

در این‌جا درخت‌های حاصل تفاوت زیادی دارند.

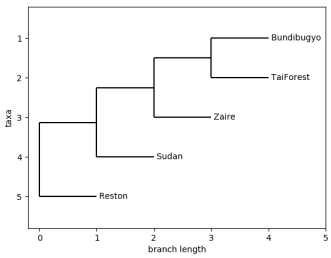


VP۲۴_UPGMA.png (ب)

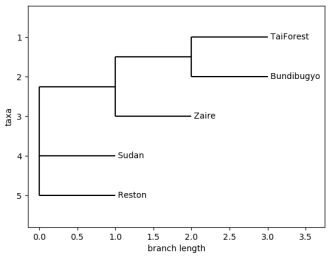


VP۲۴_NJ.png (آ)

تنها در قسمت Zaire و Bundibugyo تفاوت در درخت‌ها هست.



L_UPGMA.png (ب)



L_NJ.png (آ)

در اینجا هر دو درخت یک مفهوم را میرسانند اگرچه ظاهرا تفاوت جزیی مشاهده می‌شود.

۲.۲ مقایسه‌ی نتایج

نمی‌توان گفت نتایج تفاوت‌هاي چشم‌گیر دارند. مخصوصا اینکه در NJ در چهار ژن به یک درخت رسیده‌ایم که در سه عدد از UPGMA ها هم همین درخت حاصل شده‌است اما به هر حال نتایج در مورد بعضی ژن‌ها تفاوت دارند.

۱.۲.۲ بیان دلیل تفاوت‌ها

بستگی دارد منظور کدام نوع از تفاوت‌ها باشد. اگر منظور تفاوت‌های بین نتایج دو الگوریتم است، که این دو الگوریتم فرض‌های متفاوتی در مورد درخت اصلی دارند که به آن‌ها در قسمت‌های قبلی اشاره شده و منطقی است که نتایج متفاوت باشد. مثلا UPGMA تاکید دارد که فاصله‌ي تمام برگ‌ها از درخت ریشه یکسان باشد.

اگر منظور تفاوت‌های بین درخت‌های حاصل برای ژن‌هاست، الزاما جهش‌های ژنی روند درختی دقیق ایجاد ویروس‌های جدید را دنبال نمی‌کنند، مثلا ممکن است یک نوع از ویروس در یک جایگاه مثلا جایگاه پنجاهم که در ژن سوم است جهش کند اما انواع دیگر نکنند، در این‌جا روش‌های ما این ویروس را جدا از بقیه ترسیم می‌کند و حتی او را از تمام برادران و پدرانش در یک دسته‌ی جدا میگذارد. پس الزاما بر حسب یک ژن نمی‌توان تصمیم گرفت که روند جهش ها و تولید گونه های جدید ویروس چه بوده است اما با ترکیب آن‌ها می‌توان دقت بیشتری در این امر به دست آورد.

۲.۲.۲ مقایسه‌ی دو الگوریتم NJ ، UPGMA

مزیت بزرگ الگوریتم ،UPGMA محاسبه‌ی درخت‌های ریشه‌دار است. این الگوریتم با فرض یکسان بودن فاصله‌ی تمام برگ‌ها از ریشه درخت را محاسبه می‌کند، این فرضی است که در بسیاری از موارد نادرست است. فرض دیگر این الگوریتم ثابت بودن نرخ تکامل است(که من متوجه نمی‌شوم این فرض در کجای این الگوریتم هست ناشی از عدم تسلط من به ریاضی پشت منطق این الگوریتم است). ثابت فرض کردن نرخ تکامل را نمی‌توان یک نقطه‌ی ضعف در نظر گرفت، زیرا طبق مطالبی که در ویکیپدیا خواندم اکثر کاربردهای این الگوریتم در مواقعی است که می‌خواهند بدون در نظر گرفتن نرخ تکاملی بین توالی‌ها و تنها با دقت به شباهت آن‌ها، آن‌ها را گروه‌بندی کند که در شاخه‌ای به نام Phenetics که هدف آن دسته‌بندی میان موجودات بر حسب مشاهدات کلی است، روشی مناسب است.

به نظر من حتی در همین پروژه دیدیم که این الگوریتم در زمینه‌ی دسته‌بندی بر حسب ژن‌ها چندان هم خوب عمل نکرد و درخت‌های نسبتا متفاوتی را برای هر ژن نتیجه می‌داد البته اجماع تمام آن‌ها به درختی صحیح منتهی شد که نتیجه منطبق با نتیجه‌ي ترکیب‌شده‌ی NJ بود و البته طبق بررسی‌های *یواشکی :)* ما در اینترنت با نتیجه‌ي درست هم منطبق است.

نکته‌ی منفی الگوریتم NJ تولید درخت‌های بدون ریشه است، اما نکته‌ي مثبت آن عدم فرض یکسان بودن نرخ تکامل است(البته من متوجه نمی‌شوم که این فرض در کجای الگوریتم هست که ناشی از عدم تسلط من به ریاضی پشت منطق این الگوریتم است).همین فرض مثبت این الگوریتم آن را برای بررسی داده‌های توالی بسیار مناسب کرده است. نقطه‌ی مثبت دیگر این الگوریتم سریع بودن آن در قیاس با روش‌های دیگر است که آن را برای بررسی داده‌های در مقیاس بزرگ و استفاده

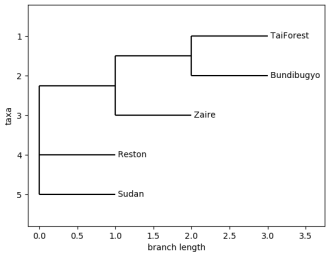
از روش‌های آماری که بارها از داده نمونه‌برداری می‌کنند(مانند (bootstrap مناسب کرده‌است. این روش در صورتی که ماتریس فاصله‌ها درست باشد، پاسخ صحیح را تولید می‌کند، اما حتی اگر ماتریس فاصله‌ها درست نباشد و حدودا درست باشد، این الگوریتم به طرز معجزه‌آسایی به احتمال خوبی باز هم درخت درست را تولید می‌کند(شهودی در مورد این که این اتفاق چرا می‌افتد پیدا نکردم، حتی ویکیپدیای انگلیسی از لفظ but neighbor joining often constructs the correct tree topology anyway) استفاده کرده‌است و به یک مقاله (اینجا) لینک داده است که فرصت نشد مطالعه کنم. این الگوریتم به بعضی از شاخه‌ها وزن منفی می‌دهد و این نامطلوب است(این در ویکیپدیا گفته شده، اما ما در این تمرین مشاهده نکردیم).

۳.۲ ترکیب درخت‌ها و ارائه‌ی درخت نهایی

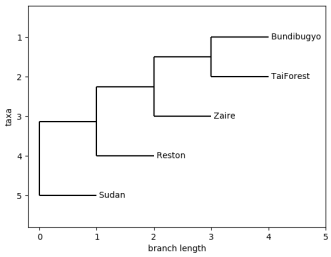
ما در اینجا از پکیج ape متد consensus استفاده کردیم. در این روش ما از احتمال برابر با ۰,۵ استفاده کردیم به این صورت است که یال‌هایی را که در ۵۰ درصد یا بیش‌تر از درخت‌ها آمده است با همدیگر ترکیب می‌کند و یک majority rule consensus tree حاصل می‌کند. این روش روش مناسبی به نظر می‌رسد چرا که به هر یال به اندازه‌ای که تکرار شده وزن و اهمیت می‌دهد و در واقع روش مبتنی بر اهمیت دادن به یال‌های آمده در روش‌هاست.(البته واژه‌ی یال در این‌جا بار معنایی مناسب و دقیق را ندارد و منظور ما در اینجا از یال clade است.)گویا این الگوریتم یال‌ها را بر حسب تعداد تکرار مرتب می‌کند و سعی می‌کند درختی تولید کند که با اکثر آن‌ها که بالای ۵۰ هستند بخواند.

۴.۲ مقایسه‌ی درخت ترکیبی و درخت حاصل از همترازی سراسری

ابتدا درخت حاصل از همترازی سراسری را مشاهده می‌کنیم.

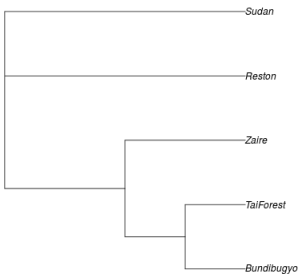


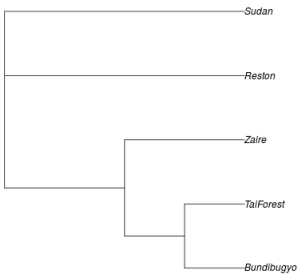
(آ) genome_NJ.png



(ب) genome_UPGMA.png

و سپس درخت‌های ترکیبی را مشاهده می‌کنیم:



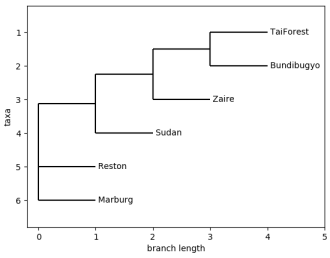


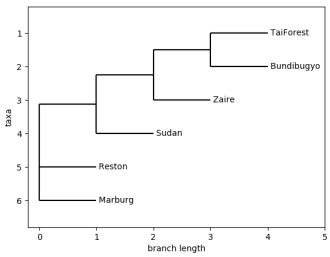
mix_UPGMA.png (ب)　　mix_NJ.png (آ)

همانطور که مشاهده می‌شود تمام این نتایج یکسان هستند و چندان مقایسه‌ای نداریم:).

۵.۲ تعیین نقطه‌ی شروع

روش ما در این بخش همان روش همترازی سراسری است. کافی است ژن تمام ۵ گونه‌ی ویروس ابولا را با ویروس ماربرگ همترازی سراسری کنیم و از فاصله‌ی ویرایش آن استفاده کرده و درخت را تشکیل دهیم،خواهیم داشت:





(ب)
all_and_marburg_UPGMA.png　　all_and_marburg_NJ.png (آ)

پس نتیجه می‌گیریم که در ابتدا(یا طبق عبارت صورت سوال نقطه‌ی شروع) ویروسی بوده که به سه گونه‌ی marburg و Restone و گونه‌ی پدر تمام ویروس‌ها Sudan و Zaire و Taiforest و Bundibugyo جهش پیدا کرده. آن گونه‌ی پدر آخری به Sudan و گونه‌ی پدر گونه‌های Zaire و Bundibugyo و TaiForest جهش پیدا کرده. آن گونه‌ی پدر گونه‌های Zaire و Bundibugyo و TaiForest به Zaire و گونه‌ی پدر Bundibugyo و TaiForest جهش پیدا کرده. در نهایت این گونه‌ی پدر آخری به TaiForest و Bundybugyo جهش پیدا کرده.

۳ بخش چهارم

۱.۳ چه زمانی از هم جدا شده‌اند؟

در این‌جا ما از مدل Jukes_Cantor استفاده کرده‌ایم. علت استفاده از این روش این است که تنها روشی است که بلد بودیم و البته ساده است و قدرت محاسباتی زیادی به ما می‌دهد(با فرض بازگشت‌پذیر بودن مارکوف‌ها در زمان و غیره). حقیقت این است که چون با روش‌های دیگر آشنا نیستم نمیتوانم بگویم که انتخابی داشته‌ام. اما این روش در کل ساده و منطقی به نظر می‌رسد. مثلا دارد نرخ خطای دی ان ای پلیمراز را مدل می‌کند و منطقی است که این نرخ خطا چندان تغییر نکند. مخصوصا این که ویروس و گونه‌هایش در چیزی حدود چند صد سال اخیر به وجود آمده‌اند. در این روش یک ماتریس احتمال به صورت

$$P(t)=\begin{array}{c} \begin{matrix} A & T & G & C \end{matrix} \\ \begin{pmatrix} 1-3f(t) & f(t) & f(t) & f(t) \\ f(t) & 1-3f(t) & f(t) & f(t) \\ f(t) & f(t) & 1-3f(t) & f(t) \\ f(t) & f(t) & f(t) & 1-3f(t) \end{pmatrix} \end{array} \begin{matrix} A \\ T \\ G \\ C \end{matrix}$$

شکل ۱۱: ماتریس احتمال مارکوف

داریم که احتمال تبدیل یک نوکلئوتید به دیگر نوکلئوتیدها را مدل می‌کند. با این مدل‌سازی و استفاده از تقریب‌هایی که در کلاس استاد مطهری به آن‌ها اشاره شده است می‌توان به این دو نتیجه رسید:

$$\mathbf{pij(t) = f(t) =} \frac{1}{4} - \frac{e^{-4\alpha t}}{4}$$

$$\mathbf{pii(t) = 1-3f(t)=} \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$

شکل ۱۲: نتایج حاصله که در آن‌ها آلفا برابر یک سوم است
$\alpha = \frac{۱}{۳}$

پس با توجه به تعداد تعداد جایگاه‌هایی که دو ژن در آن‌ها به هم شبیه هستند(آن را b بنامید)و تعداد جایگاه‌هایی که در دو ژن متفاوت هستند(آن‌ها را a بنامید). پس احتمال رسیدن دو ویروس به هم را می‌توان در صورت دانستن فاصله‌ی بین آن دو با فرمول:

$$(\frac{۱}{۴}(۱ - e^{-\alpha t}))^a \times (\frac{۱}{۴}(۱ + ۳e^{-\alpha t}))^b$$

حساب می‌شود. حال باید tای را انتخاب کنیم که احتمال بالا را بیشینه کند. با گرفتن لوگاریتم برای ساده‌تر شدن و انجام محاسبات خواهیم داشت:

$$t = -\frac{3}{4} \ln(1 - \frac{4}{3} p)$$

شکل ۱۳: نتیجه‌ی نهایی
که در آن $p = \frac{a}{a+b}$

البته باید دقت کنیم که در این تصاویر در واقع $t = \lambda \times time$ است که در آن λ برابر با نرخ تحول است. پس باید t حاصل در فرمول بالا را بر λ که اطلاعات داده شده در صورت پروژه برابر با $۱/۹ \times ۱۰^{-۳}$ در نظر گرفته شده، تقسیم کنیم. در این قسمت البته از این منبع استفاده کردیم. حال ما از distance edit به عنوان یک تقریب برای a و از طول یکی از ژنوم‌ها(با توجه به نزدیک بودن طول تمامی ژنوم‌ها) به عنوان یک تقریب برای $a + b$ یا در واقع طول ژنوم استفاده کردیم پس داریم:

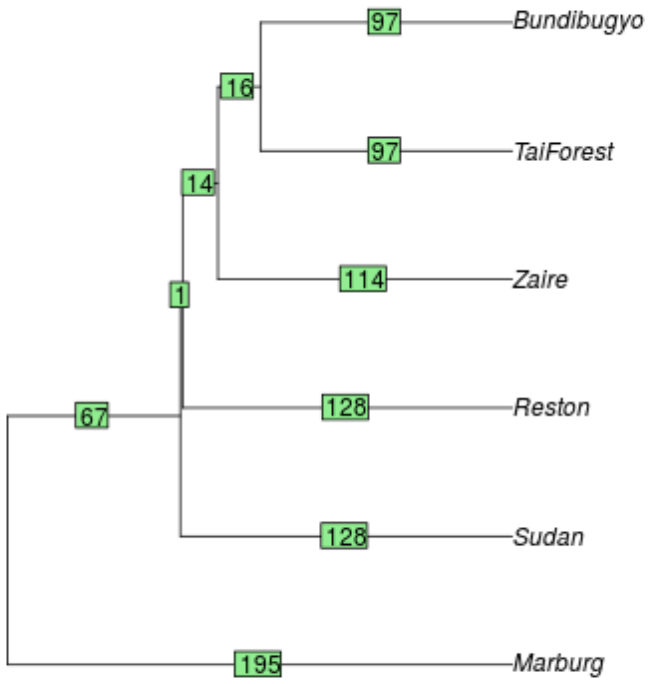
$$p \approx \frac{edit\_distance}{genome\_length}$$

و سپس یک ماتریس محاسبه کردیم که شامل این فاصله‌های زمانی بود

	Reston	Sudan	TaiForest	Zaire	Bundibugyo	Marburg
Reston	0	258	257	253	258	393
Sudan	258	0	259	255	260	398
TaiForest	257	258	0	227	195	400
Zaire	252	254	227	0	229	395
Bundibugyo	257	258	195	229	0	401
Marburg	385	390	394	389	395	0

شکل ۱۴: فاصله‌های زمانی بین گونه‌ها

و آن را به الگوریتم UPGMA دادیم و طول یال‌ها را برای ما حساب کرد و این نتیجه خروجی شد:



شکل ۱۵: درخت زمانی حاصله

که در آن این اعداد نوشته شده در جداول سبز در واقع سال‌های تخمین زده شده هستند، مثلا طبق این جدول marburg ۱۹۵ سال با پدر این ویروس‌ها فاصله دارد و sudan ۱۲۸ سال با پدر شاخه‌ی ابولا ویروس فاصله دارد و پدر ابولاویروس‌ها ۶۷ سال با پدر مشترکش با ماربرگ ویروس فاصله دارد. پس در کل حدس این است که جد مشترک ابولا ویروس‌ها ۱۲۸ سال قبل میزیسته و پدر مشترک تمام آن‌ها با ماربرگ در ۱۹۵ سال پیش میزیسته(خصوصیت جالب UPGMA که فرض می‌کند تمام برگ‌ها در یک زمان هستند را نیز می‌توانید در تصویر مشاهده کنید).