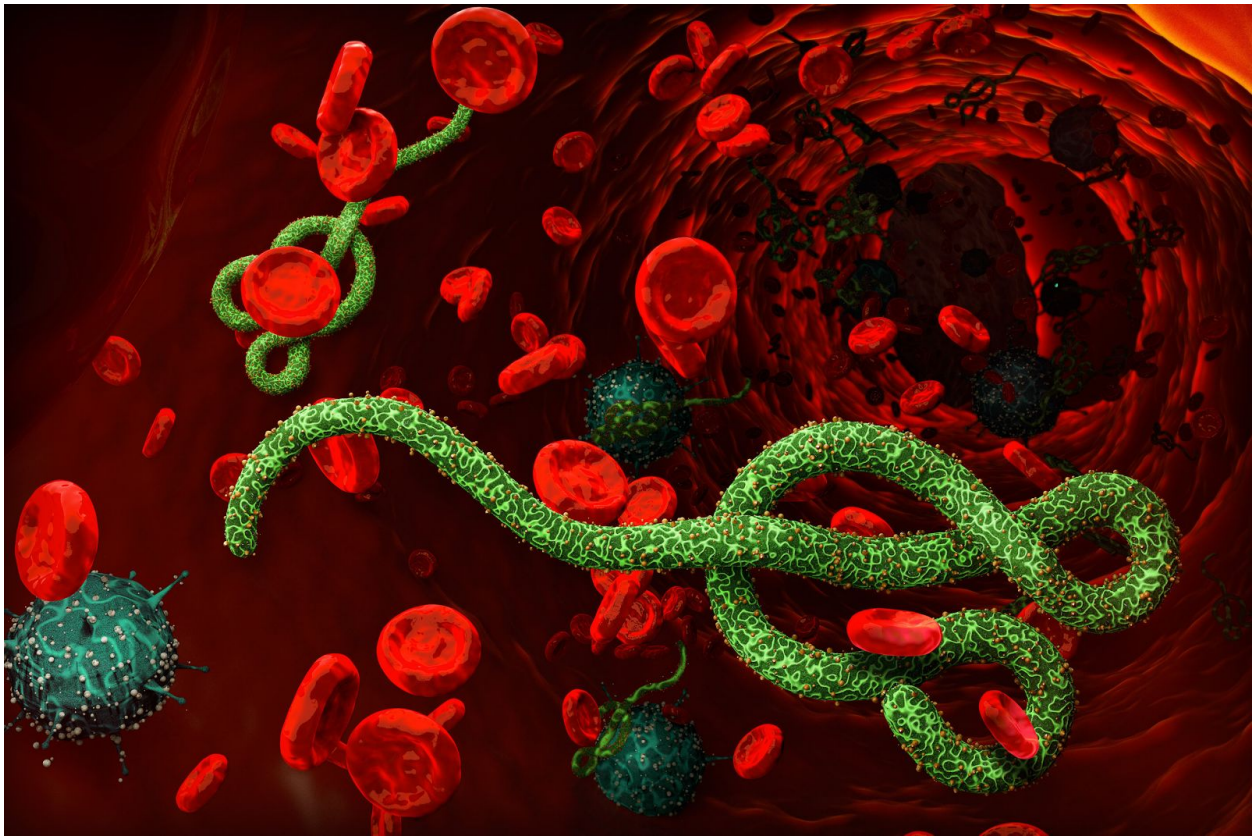
درخت زندگی ابولا ویروس



۱. مقدمه

۱.۲ :

ژنوم ابولاویروس شامل ۷ ژن است. نام و توضیح مختصری درباره هر کدام از آن‌ها در زیر آمده است:

NP: از این ژن یک پروتئین هسته‌ای(nucleoprotein) تولید می‌شود که نقش ساختاری دارد. چون این ژن در ابتدای ژنوم قرار دارد بیشتر از سایر ژن‌ها به پروتئین ترجمه می‌شود.

VP35: از این ژن یک پروتئین ساختاری تولید می‌شود که جزئی از complex مضاعف‌سازی و رونویسی از ژن می‌باشد.(replication/transcription complex). پروتئین حاصله نقشی شبیه پروتئین P در پارامیکسو ویروس و رابدوویروس دارد.

VP40: این ژن به پروتئین ساختاری که در ماتریکس جای می‌گیرد ترجمه می‌شود.

GP: این ژن به یک گلیکوپروتئین بزرگ منفرد ساختاری ترجمه می‌شود.

VP30: از این ژن یک پروتئین ساختاری تولید می‌شود که جزئی از complex مضاعف‌سازی و رونویسی از ژن می‌باشد.(replication/transcription complex). پروتئین حاصله یک فسفوپروتئین است.

VP24: پروتئین حاصل از این ژن همراه envelope است. همچنین این پروتئین یک پروتئین ساختاری مینور در ماتریکس می‌باشد.

L: این ژن به پروتئین ساختاری RNA Polymerase وابسته به RNA ترجمه می‌شود. [1]

پروتئین‌هایی که در replication این ویروس نقش دارند به دنبال تکثیر ویروس باعث گسترش این ویروس در بدن موجودات زنده می‌شوند. بنابراین پروتئین‌های حاصله از VP35, VP30 و L احتمالا نقش مهی در بیماری‌زایی ایفا کنند.

همچنین پروتئین حاصله از ژن VP24 به یک پروتئین ناقل بر روی سطح سلول‌های ایمنی که نقش مهمی در مسیر اینترفرون دارد متصل شده وآن را مهار می‌کند. اینترفرون از سلول‌های ایمنی بدن ترشح شده و مانع از تکثیر بیشتر ویروس در بدن می‌شود. بنابراین VP24 در بیماری زایی نقش بسیار مهمی دارد.

پروتئین حاصل از GP نیز به receptor سلولی متصل شده و باعث ورودی ویروس به درون سلول می‌گردد. بنابراین این پروتئین نیز در بیماری زایی نقش دارد.[2]

[1] Characterization of Ebolavirus regulatory genomic regions, Virus Research Journal
[2] Sequence analysis of the Ebola virus genome: organization, genetic elements, and comparison with the genome of Marburg virus, Virus Research Journal

۱.۳:

بعد از ورود ویروس به بدن انسان و پریمات‌ها یک دوره تکثیر اولیه ویروس اتفاق میفتد. در این مرحله پاسخ ایمنی وجود ندارد. گلیکوپروتئین ویروس یک پروتئین ترشحی غیرساختاری است که به مولکول CDb16 در نوتروفیل‌ها می‌چسبد و آن‌ها را مهار می‌کند. پس GP در مهار سیتم ایمنی میزبان نقش دارد. ۴ تا ۶ روز پس از عفونت اولیه سیستم ایمنی با تولید اینترلوکین ۱ بتا، اینترلوکین ۶و فاکتور نکروز دهنده تومور(TNF) به عفونت پاسخ می‌دهد.

یک GP دوم به سلول‌های اندوتلیال میزبان می‌چسبد. ویروس ابولا به این سلول‌ها هجوم برده، در آن‌ها تکثیر می‌شود و این سلول‌ها را تخریب می‌کند. تکثیر ویروس و تخریب سلول باعث نکروز موضعی بافت می‌شود که شدیدترین آن در کبد رخ می‌دهد. در انواع کشنده بافت میزبان حاوی مقادیر فراوان ویروس ابولا می‌باشد و با تخریب بافت انعقاد داخل عروقی رخ می‌دهد. این مسئله به همراه نکروز بافت کبد، موجب خونریزی از قسمت‌های مختلف بدن و شوک هایپوولمیک می‌شود و در نهایت موجب مرگ میزبان می‌گردد. [3]

۲. توالی

۲.۱:

داده‌های ژنوم در فایل BioProjectFiles موجود است.

۲.۲:

در مرحله اول نیاز اطلاعات را از فایل‌های مربوط به ژنوم گونه‌های مختلف خواندم. برای این‌کار ابتدا فرمت تمامی فایل‌ها را به .txt تغییر دادم. سپس با دستور open در پایتون تمامی فایل‌ها را باز کردم. سپس با توجه به فرمت fasta نیاز بود خط اول هر فایل را نادیده بگیرم. سپس خط به خط فایل‌ها را خواندم و اطلاعات ژنوم هر گونه را در آرایه‌ای مربوط به آن گونه ریختم. سپس این آرایه‌ها را به صورت string در آوردم که قابل استفاده در کتابخانه biopython شود.

[3] Ebola Virus Infection: Practice Essentials, Background, Pathophysiology and Etiology, MedScape

چون alignment به کل ژنوم هر موجود بسیار طول می‌کشید و این که ما ترتیب ژن‌ها و طول تقریبی آن ها را می‌دانیم نیازی نبود alignment را بر روس کل ژنوم هر موجود انجام دهم.

برای هر ژنوم محدوده‌ای از ژنوم را در نظر گرفتم و از الگوریتم local alignment استفاده کردم. بازه الاینمنت را از ۵۰۰ تا قبل از جای پیشبینی شده تا ۲ برابر طول آن قرار دادم. در local alignment اگر امتیازها را به صورت ۱ برای match و ۱- برای سایر موارد(mismatch و gap) قرار دهیم هزینه شکاف‌های ابتدایی و انتهایی صفر می‌شود. تنها مشکل این روش زمان طولانی اجرای آن برای رشته‌های بلند بود. به همین دلیل برای ژن L ناچار به دوقسمت کردن این ژن شدم.(به علت طول زیاد L)

:۲.۳

در این مرحله برای هر ژن یک ماتریس ۵ در ۵ تشکیل دادم. هر بعد هر ماتریس نشان‌دهنده ژن یک گونه خاص می‌باشد. الگوریتم نیدلمن وانچک همترازی سراسری بر دو ترتیب متوالی (مانند A و B) انجام می‌دهد. معمولاً در بیوانفورماتیک برای همترازی توالی‌های پروتئینی یا نوکلئوتیدی کاربرد دارد.

برای به دست آوردن فاصله بین دو موجود در هر ژن از الگوریتم global alignment با امتیازهای 0 برای match و ۱- برای سایر موارد استفاده شده‌است. چون distance matrix باید شامل اعدادی مثبت شود در نهایت هر خانه ماتریس در ۱- ضرب شده‌است. در آخر به ذخیرسازی این ماتریس در فایل csv پرداختم که این فایل‌ها در پوشه csv قرار دارند.

۳. درخت زندگی

:۳.۱

در این قسمت برای ساخت درخت از کتابخانه Bio.Phylo.TreeConstruction استفاده کردم که هر دو الگوریتم neighbor joining و upgma را دارد. برای رسم درخت نیز از کتابخانه Phylo استفاده کردم.

دو الگوریتم neighbor joining و upgma روش‌های مختلفی برای ساخت درخت استفاده می‌کنند. در الگوریتم NJ ماتریس داده شده به آن بهتر است به صورت additive باشد. در UPGMA درخت در نهایت ریشه‌دار و در NJ در نهایت بدون ریشه می‌شود.

از دیگر تفاوت‌های این دو الگوریتم این می‌باشد که در NJ درخت خروجی rate در شاخه‌های مختلف امکان دارد متفاوت باشد ولی در upgma این rateها یکسان می‌باشند.

در مسائلی که cleck modecular می‌باشند خروجی upgma دقیق است اما در مجموع الگوریتم nj خروجی دقیق‌تری می‌دهد.[4] [5]

در الگوریتم NJ باید مراحل زیر طی شود:

ابتدا آرایه ۱ بعدی total_distance را تشکیل می‌دهیم که در واقع مجموع سطرهای مختلف distance matrix می‌باشد.

سپس ماتریس matrix_star را به صورت زیر تعریف می‌کنیم:

matrix_star[i][j] = (n - 2) * matrix[i][j] - total_distance[i] - total_distance[j]

سپس کمترین خانه ماتریس matrix_star را انتخاب کرده و آن دو را با هم مرج می‌کنیم.

در نهایت ماتریسی که ابعاد آن یکی کاهش پیدا کرده را در مرحله ۱ قرار داده و این کار را به صورت بازگشتی انجام می‌دهیم تا اندازه ماتریس ۱ شود. [6]

در الگوریتم UPGMA باید مراحل زیر طی شود:

بیشترین خانه ماتریس distance matrix را انتخاب کرده و آن دو را با هم مرج می‌کنیم.

سپس ماتریسی که ابعاد آن یکی کاهش پیدا کرده را در مرحله ۱ قرار داده و این کار را به صورت بازگشتی انجام می‌دهیم تا اندازه ماتریس ۱ شود. [7]

:۳.۲

برای این کار از کتابخانه Bio.Phylo.Consensus استفاده می‌کنیم.

[4] https://www.mun.ca/biology/scarr/Panda_UPGMA_&_NJ.html
[5] https://www.researchgate.net/post/What_is_the_difference_between_UPGMA_and_NEJ_method_while_constructing_a_tree_using_a_MEGA_4_software
[6] https://en.wikipedia.org/wiki/Neighbor_joining
[7] https://en.wikipedia.org/wiki/UPGMA

consensus tree یک تخمینی از درخت نهایی را به ما می‌دهد. این الگوریتم ۲ نوع درخت Majority rule consensus tree و Strict consensus tree پیادسازی می‌شود. هر دوی این‌ها بر اساس فرکانس تکرار cluster کار می‌کنند.

strict consensus tree تنها cluster هایی را در بردارد که در تمامی درخت‌ها یافت می‌شوند. این در خالی است که majority rule consensus tree دارای cluster هایی است که حداقل در نصف درخت‌ها تکرار شده‌اند. این نصف بودن را می‌توان با cutoff تغییر داد.

Strict consensus tree تنها زمانی کاربرد دارد که ما به دنبال cluster هایی هستیم که همه جا تکرار شده‌اند.

در اینجا از روش Majority rule consensus tree استفاده کردیم زیرا این روش در واقع median consensus tree نیز می‌باشد و برای ساختن confidence intervals نیز مناسب است. این روش به خاطر تغییر پذیری مقدار cutoff انعطاف پذیر نیز است.[8] [9]

:۳.۳

در اینجا از الگوریتم myers's bit vector algorithm که در global alignment پیاده سازی شده‌است استفاده کردم.

:۳.۴

این قسمت مشابه قسمت قبل می‌باشد، فقط فواصل تا marburg را نیز به ماتریس فاصله اضافه کردم.

[8] https://assets.geneious.com/manual/11.0/GeneiousManualsu111.html
[9] http://taxonomy.zoology.gla.ac.uk/rod/cplite/ch4.pdf

۴. تخمین گذشته، پیش‌بینی آینده

۴.۱:

مدل Jukes_Cantor یک مدل سیر تکاملی DNA می‌باشد.در این مدل p به صورت تقسیم تعداد اشتراکات دو رشته بر طول مشترک ۲ رشته به دست می‌آید. سپس برای به دست آوردن t از فرمول زیر استفاده می‌شود:

$$t = -0.75 * \ln(1 - p/0.75)$$

در نهایت با تقسیم این t بر لاندا که در سوال داده شده زمان‌ها به دست می‌آیند.

لازم به ذکر است که این مدل ساده‌ترین مدل جانشینی موجود است و. همچنین در این مدل فرض می‌شود که substitution rates در بین سایت‌های مختلف برابر است.[10] [11]

بعد از تعریف تابع jukes cantor آن را برای هر دو گونه موجودات فراخوابی کردن و خروجی این فراخوانی‌ها را در time matrix قرار دادم. در این ماتریس فاصله زمانی هر دو گونه آمده است.

سپس اگر درخت ریشه دارد حاصل از الگوریتم upgma بر روی این ماتریس را در نظر گرفتم و فاصله جد مشترک تا marburg را به دست آوردم که همان زمان جدا شدن گونه‌ها از جد مشترک است.

۴.۲:

میانگین edit distance بین موجودات برابر 6541 می‌باشد. بنابراین می‌توان فرض کرد به طور متوسط یه این تعداد جهش نیاز است تا گونه جدید به وجود بیاید. همچنین طول متوسط این ۵ گونه 18920 می‌باشد. طبق مدل پیاده سازی شده با فراخوانی تابع jukes cantor عدد 243 به دست می‌آید(به سال) که زمان متوسط برای جهش بعدی است.

[10] https://en.wikipedia.org/wiki/Models_of_DNA_evolution
[11] http://treethinkers.org/jukes-cantor-model-of-dna-substitution/