سارا حسنلو نیکفر

۹۹۴۱۳

گزارش پروژه

استخراج الگوهای پرتکرار

دکتر نریمانی

داده کاوی پیشرفته

۲۸ اسفند ۱۴۰۰

فهرست

مقدمه:

در این پروژه از دیتا ست LIHC برای یافتن frequent pattern های مربوط به جهش های متناظر با بیماران استفاده شده است. هر سطر این دیتاست شامل بیمار و یک جهش مربوط به آن بیمار می باشد. به این معنی که امکان تکرار هر بیمار و جهش متناظر در سطرها وجود دارد زیرا بیماران بیش از یک جهش را دارند. در هر سطر هر بیمار و جهش مربوط به او را به صـــورت متنی با یک جداکننده -(می توان مشـــاهده کرد. اما علاوه بر این دیتاست حاوی سطور تکراری نیز می باشد که بر اساس بررسی های انجام شده حدود ۲۰۰ هزار سطر تکراری وجود دارد که حذف آنها به خودی خود در کاهش سرعت و دقت محاسبات تاثیر به سزایی دارد.

قسمتی از فایل مقایسه موارد تکراری در این قسمت آورده شده است فایل اصلی با اسم comparison در فولدر files ذخیره شده است. با مقایسه دو ستون اعداد متوجه اختلافات و تکرارهای موجود میشویم.

138	TCGA-2V-...	['FOXG1-14-...	15	['GTF2IRD2-7-7/...	15
139	TCGA-G3-...	['ZFR-5-3240...	15	['ALPP-2-233246...	15
140	TCGA-DD-...	['AC008103.5...	15	['TANC2-17-614...	16
141	TCGA-DD-...	['NPIPB5-16-...	15	['ADAM21P1-14...	16
142	TCGA-ZP-...	['NBPF1-1-1...	17	['DSPP-1-885364...	17
143	TCGA-UB-...	['POTEG-14-...	17	['ZC3H11A-1-20...	17
144	TCGA-BD-...	['RANBP5-9-...	13	['PMS2CL-7-677...	18
145	TCGA-DD-...	['EEF1A1-6-7...	25	['RP11-417I8.6-1...	21
146	TCGA-WQ-...	['ZFR-5-3240...	35	['PTPRF-1-1107...	35
147	TCGA-DD-...	['RP11-156P...	51	['PKN2-1-692705...	45
148	TCGA-G3-...	['KRTAP4-11...	45	['PRL-6-2229458...	46
149	TCGA-BC-...	['DNAH7-2-...	51	['TRIM54-2-2752...	51
150	TCGA-DD-...	['PTPRB-12-...	87	['RP11-31M7.1-...	52
151	TCGA-DD-...	['RP11-262H...	55	['PKN2-1-692705...	55
152	TCGA-4F-...	['SRSF8-11-9...	59	['CTNNB1-3-412...	59
153	TCGA-DD-...	['LPIN2-18-2...	121	['UBXN2A-2-247...	93
154	TCGA-DD-...	['ATP2B2-3-...	113	['LINC00842-10-...	113
155	TCGA-DD-...	['VPS4B-16-6...	491	['UBXN2A-2-247...	267
156	TCGA-BC-...	['LINC00115-...	1182	['ULK4P3-15-304...	1182
157	TCGA-BC-...	['FAM86C2P...	1455	['DNAH14-1-225...	1455
158	TCGA-DD-...	['ZNF733P-7...	50405	['CAMSAP2-1-2...	50406
159	TCGA-DD-...	['INTS4L2-7-...	170459	['CPS1-2-21-507...	87068
160	TCGA-DD-...	['MUC2-11-1...	90415	['CPS1-2-21-507...	90416

راهکار انجام شده:

دیتاست توسط پکیج pandas از فایل csv لود شده و با فرمت دیتا فریم ذخیره شده و سطور تکراری آن تو سط drop_duplicates حذف می شود. طوریکه داده ها از ۹۸۹۰۵۰ سطر به ۷۸۷۴۵۷ سطر کاهش پیدا می کند.

بر اساس دو کاراکتر مذکور بیمار و جهش از هم جدا شده و پرانتز اضافی انتهای شناسه بیمار نیز حذف می شود. که در فایل Splited_LIHC ذخیره شده اند. Minimum support را ۲ در نظر گرفته و داده ها بر اساس جهش ها گروه بندی شده و تعداد تکرار هر جهش ذخیره میشود و جهش هایی با تعداد تکرار کمتر از ۲ از دیتافریم حذف میشوند و به این صورت آیتم ست های با طول ۱ در مرحله اول مشخص میشوند. جهش ها را با تعداد آنها در فایل freq__one__item ذخیره نموده ام. علاوه بر آن frequent pattern ها در دیکشنری freqPatternDict با کلید نام جهش و مقدار support مربوط به هر جهش ذخیره می شوند تا در ادامه الگوهای با طول بیشتر به آن اضافه شده و یا الگوهایی در طول فرآیند از آن حذف شود.

برای یافتن آیتم ست های با طول بیش از ۱، فرمت داده های فیلتر شده بر ا ساس Minimum support را این بار بر اساس بیمار گروه بندی کرده و جهش های مربوط به هر بیمار در ستونی به صورت لیست ذخیره میشوند. در مجموع ۱۶۲ بیمار پس از پیش پردازش های بیان شده وجود دارد. چند بیمار آخرتقریبا ۱۰ بیمار آخر دیتاست، تعداد جهش های بسیار زیادی دارند که از حدود ۹۳ شروع شده و به حدودا ۹۰ هزار میرسد. بقیه بیماران جهش هایی با تعداد بین ۱ تا ۶۰ دارند. به طوریکه بیمار ۱۴۷ ام، ۱۸ جهش دارد. بنابراین طول یک آیتم ست میتواند به ۹۰ هزار برسد که قطعا مجموعه هایی با این تعداد آیتم جز frequent pattern به حساب نمی آیند چون فقط برای یک بیمار تکرار شده اند. و اکثر مجموهای آیتم های جهش طول کمی دارند یعنی آیتم ها یا جهش های کمتری را شامل میشوند.

برای تو ضیحات ا ضافی، قابل ذکر ا ست ۱۲۰ بیمار تعداد جهش کمتر م ساوی ۱۱ دارند که ن شان دهنده نحوه پراکندگی تعداد جهش ها در بیماران است. پس تقریبا یک دیدی از فراوانی جهش ها تا این مرحله به دست آورده ایم. که در فایل patient_mutations برای مشاهده ذخیره شده است.

در دیتافریم patient_mutations تعداد سطور مربوط به بیماران حداکثر ۱۶۲ بود که با iloc، slice کرده و همچنین تعداد زیرمجموعه هایی که باید ساخته شوند حداکثر به اندازه تعداد آخرین سطر slice یا همان آخرین بیمار بر اساس slice در نظر گرفته میشود زیرا بیمار و جهش های مربوط به او به صورت صعودی بر اساس تعداد جهش های هر بیمار مرتب شده اند.

```
for subsetLen in range(2, maxlen + 1, 1):
```

ساختن زیرمجموعه و شمردن آنها در دیتاست از اصول اصلی frequent pattern analysisمی باشد.

```
dataSet = patient_mutations.iloc[0:161, :]
maxlen = 90416
```

slice نمودن جهت تسریع در اجرا و تحلیل الگوریتم و راستی آزمایی آن صورت گرفت. به این صورت که پردازش روی سطرهای کمتری انجام شده و امکان بررسی بهتر الگوریتم وجود دارد. در نهایت با تعداد واقعی دیتاست بدون اعمال slice و محدودیت در تعداد زیرمجموعه ها، الگوریتم اجرا می شود و frequent patternها به ادامه دیک شنری ا ضافه شده و در نهایت در فایل freqPatternDict نو شته شدند. جزییات مربوط به مواردی که اینجا بیان شد به صورت شفاف در فرآیند کار تو ضیح داده خواهد شد. قابل ذکر ا ست با توجه به اینکه شمارش روی تمام سطرهای پایانی هم صورت میگیرد حذف فرایند ایجاد زیرمجموعه از چند سطر آخر و شمارش سطرهای بعدی سطر مربوطه که دنباله خیلی زیادی از جهش ها را دارند خلل زیادی در الگوهای پرتکرار ایجاد نمیکند چون تعداد این سطرها محدود است و لوپ روی جهش هایی با تعداد بالاتر از هزار بر نا مه را بسیار تحت تاثیر قرار می د هد. در شکل این تعداد آ مده است و کا مل آن در فا یل patient_mutations موجود است.

	Patient	Mutation	length
137	TCGA-2Y-A9H1-01A-11D-A382-10	['HSD17B7-1-162769603-Missense_Mutation', 'WASH4P-16-67407-Silent', 'HLX-1-221057861-Missense_Mutation', 'ADAM21P1-14-7...	15
138	TCGA-2Y-A9H5-01A-11D-A382-10	['FOXG1-14-29236691-Missense_Mutation', 'NBPF10-1-145293269-Splice_Site', 'SMG1-16-18937327-Missense_Mutation', 'C14orf39-...	15
139	TCGA-G3-AAV5-01A-11D-A36X-10	['ZFR-5-32407029-Silent', 'ADAM21P1-14-70714144-RNA', 'GTF2IRD2P1-7-72663998-RNA', 'NOX4-11-89106611-Missense_Mutation'...	15
140	TCGA-DD-A118-01A-11D-A12Z-10	['AC008103.5-22-18846232-RNA', 'ZNF860-3-32032046-Missense_Mutation', 'GNAS-20-57484420-Missense_Mutation', 'NFATC4-14-...	16
141	TCGA-DD-AAE7-01A-11D-A40R-10	['NPIPB5-16-22545897-Missense_Mutation', 'NBPF10-1-145323656-Missense_Mutation', 'ADAM21P1-14-70713742-RNA', 'NCOA6-2...	16
142	TCGA-ZP-A9CY-01A-11D-A382-10	['NBPF1-1-16918653-Splice_Site', 'DPY19L2P2-7-102825947-RNA', 'DSPP-4-88536460-Silent', 'SMG1-16-18937327-Missense_Mutatio...	17
143	TCGA-UB-A7ME-01A-11D-A33K-10	['POTEG-14-19553528-Missense_Mutation', 'KRTAP4-11-17-39274087-Missense_Mutation', 'LRRCC1-8-86019547-Missense_Mutatio...	17
144	TCGA-BD-A2L6-01A-11D-A20W-10	['RANBP6-9-6012658-Missense_Mutation', 'UNC93B1-11-67763107-Silent', 'NBPF10-1-145293512-Missense_Mutation', 'LGALS9B-17...	18
145	TCGA-DD-A11B-01A-11D-A12Z-10	['EEF1A1-6-74227627-Missense_Mutation', 'AC024560.3-3-197348739-RNA', 'ZXDB-X-57619097-Missense_Mutation', 'EEF1A1-6-742...	21
146	TCGA-WQ-A9G7-01A-11D-A36X-10	['ZFR-5-32407029-Silent', 'LRRCC1-8-86019547-Missense_Mutation', 'MN1-22-28194936-Silent', 'PTPRF-1-44071948-Missense_Muta...	35
147	TCGA-DD-A11A-01A-11D-A12Z-10	['RP11-156P1.3-17-45127107-RNA', 'ACTR3C-7-149983565-Missense_Mutation', 'ZNF181-19-35232117-Silent', 'RP11-146E13.4-14-19...	45
148	TCGA-G3-A25S-01A-11D-A16V-10	['KRTAP4-11-17-39274087-Missense_Mutation', 'PRG4-1-186276981-Silent', 'WBP2-17-73851333-Missense_Mutation', 'SEMA3A-7-8...	46
149	TCGA-BC-A217-01A-11D-A152-10	['DNAH7-2-196825086-Missense_Mutation', 'GSDMC-8-130777987-Missense_Mutation', 'TRIM54-2-27528584-Missense_Mutation', '...	51
150	TCGA-DD-A115-01A-11D-A12Z-10	['PTPRB-12-70960239-Missense_Mutation', 'TRIM71-3-32915309-Splice_Site', 'PLXND1-3-129275500-Missense_Mutation', 'CPAMD8...	52
151	TCGA-DD-A1EJ-01A-11D-A152-10	['RP11-262H14.1-9-66459820-RNA', 'PTPRZ1-7-121684588-Missense_Mutation', 'GNB5-15-52476791-Missense_Mutation', 'TMEM13...	55
152	TCGA-4R-AA8I-01A-11D-A382-10	['SRSF8-11-94800490-RNA', 'ARHGAP11A-15-32915726-Missense_Mutation', 'MESP2-15-90320173-Silent', 'IRAK4-12-44161948-Miss...	59
153	TCGA-DD-A1EF-01A-11D-A12Z-10	['LPIN2-18-2925247-Missense_Mutation', 'FRG1B-20-29625935-Missense_Mutation', 'CDHR5-11-618833-Missense_Mutation', 'LYAR...	93
154	TCGA-DD-A1EA-01A-11D-A12Z-10	['ATP2B2-3-10443888-Missense_Mutation', 'DPF3-14-73159816-Missense_Mutation', 'EIF4A2-3-186504304-Missense_Mutation', 'AP...	113
155	TCGA-DD-A1EE-01A-11D-A12Z-10	['VPS4B-18-61067824-Missense_Mutation', 'AQR-15-35198872-Missense_Mutation', 'ARHGEF17-11-73022229-Missense_Mutation', '...	267
156	TCGA-BC-A112-01A-11D-A12Z-10	['LINC00115-1-762070-RNA', 'LINC00115-1-762136-RNA', 'LINC00115-1-762154-RNA', 'ATAD3B-1-1416247-Splice_Site', 'ATAD3B-1-...	1182
157	TCGA-BC-A3KG-01A-11D-A20W-10	['FAM86C2P-11-67560590-RNA', 'NOC2L-1-888554-Splice_Site', 'CDK11A-1-1653065-Frame_Shift_Del', 'KCNAB2-1-6158547-Frame_...	1455
158	TCGA-DD-A39Y-01A-11D-A20W-10	['ZNF733P-7-62752443-RNA', 'INTS4L2-7-65150699-RNA', 'INTS4L2-7-65150719-RNA', 'GTF2IRD2P1-7-72664010-RNA', 'GTF2IRD2P1...	50406
159	TCGA-DD-A1EG-01A-11D-A20W-10	['INTS4L2-7-65150719-RNA', 'GTF2IRD2P1-7-72664021-RNA', 'INTS4L2-7-65150699-RNA', 'GTF2IRD2P1-7-72664010-RNA', 'NBPF8-1-...	87068
160	TCGA-DD-A3A0-01A-11D-A20W-10	['MUC2-11-1093299-Silent', 'INTS4L2-7-65150719-RNA', 'GTF2IRD2P1-7-72664021-RNA', 'NBPF8-1-144220785-Splice_Site', 'MT-RNR...	90416

برای درک بی شتر فرایند مثالی با توجه به شکل بیان میکنم. مثلا برای سطر ۱۴۵ که حاوی ۲۱ جهش ه ست علاوه برای شمارش زیرمجموعه های با تعداد کمتر از ۲۱ در سطور بعدی – که از ۲ شروع میشود و تا ۲۱ ادامه دارد – در نهایت زیرمجموعه ۲۱ تایی این سطر ایجاد شده و سطور ۱۴۵ تا ۱۶۰ برای وجود زیرمجموعه ۲۱ تایی بررسی شده و شمارش انجام میشود. قابل ذکر است زیرمجموعه های کمتر از ۲۱ قبلا در سطور قبلی ایجاد

و شمارش شده اند و روی این سطر موارد تکراری که قبلا شمارش شده اند، شمارش نخواهد شد و تنها موارد جدید احتمالی شمارش خواهند شد. یعنی بررسی انجام میشود تا آیتم مربوطه قبلا شمارش نشده باشد.

```
if item not in freqPatternDict.keys()
```

می دانیم در دیتافریم patient_mutations بیماران و جهش های متناظر با آنها و تعداد جهش ها ذخیره شده و به صورت صعودی بر اساس تعداد جهش ها مرتب شده اند.

بیشـــترین تعداد جهش ها که در دیتافریم مربوطه به ســـطر آخر تعلق دارد به عنوان حداکثر تعداد برای ایجاد آیتم ســـت ها یعنی maxlen در نظر گرفته میشـــوند. به این معنی که، ابتدا زیرمجموعه های ۲ تایی از تمام جهش های دیتافریم ایجاد میکنیم برای مقابله با مشـــکل ram برای حجم بالای اطلاعات ســـطر به ســـطر روی دیتافریم این زیرمجموعه ها ایجاد میشـــود. برای مثال برای زیرمجموعه ۲ تایی ابتدا زیرمجموعه های ســـطر اول استخراج شده و با همه سطر های بعدی این سطر مقایسه میشود و شمارش صورت میگیرد. سپس زیرمجموعه دوتایی سطر دوم استخراج شده و با تمام سطرهای بعدی این سطر مقایسه میشود و این فرایند برای تمامی سطر ها و برای تعداد متفاوت زیرمجموعه ها به صـــورت دو حلقه لوپ تودرتو تکرار میشـــود. قابل ذکر اســـت تکه کدها فقط برای شفافیت آورده شده و بین حلقه ها کدهای دیگری هم وجود دارد.

```
for index in range(lenDf):
        for i in range(index + 1, len(patient_mutations)):
```

انجام شمارش به صورت زیر است که در آن r سطری می باشد که شمارش روی آن در حال انجام است.

```
if set(item).issubset(set(r['Mutation'])):
    cntItem += 1
```

برای هر آیتم ۲ تایی در مجموعه این زیرمجموعه ها تمای سطور بعدی دیتاست چک میشوند (چون به صورت اتوماتیک هنگام انتخاب زیرمجموعه سطر جاری برای وجود زیرمجموعه خاص شمرده میشود) و در صورت وجود یک زیرمجموعه دو تایی در یک ســـطر تعداد متغیر cntItem یکی افزایش می یابد که به منزله support آن آیتم می با شد. در تکرار بعدی زیرمجموعه سه تایی ایجاد شده و روی آیتم های سه تایی این مجموعه تمامی ســـطرهای دیتافریم بررســـی و در صـــورت وجود آیتم شـــمارش انجام میگیرد. این فرایند ادامه می یابد تا به maxlen برسیم و زیرمجموعه هایی با طول maxlen ایجاد کرده و آنها را بشماریم.

پس از شمارش cntItem ، در صورتی که مقدار شمارش از minSup کمتر با شد آیتم مربوطه با support صفر در دیکشنری الگوهای پرتکرار ذخیره میشود. در واقع پرتکرار درنظر گرفته نمی شود. و من برای ا ستفاده های بعدی ذخیره این آیتم را انجام میدهم.

```
freqPatternDict.update({item: 0})
```

در صورتی که مقدار شمارش از minSup بیشتر یا مساوی آن باشد، دو موضوع اصلی باید در فرآیند چک شود:

اول اینکه closed items باید پیدا شوند برای این کار دیکشنری مربوط به الگوهای پرتکرار چک شده و مقدار support آیتم هایی که زیرمجموعه آیتم جاری بوده و support یکسان با آیتم جاری دارند به ۱- تغییر میکند تا قابل شناسایی باشند و در نهایت پس از اتمام فرآیند از لیست الگوها حذف شوند.

```
notClosedSet = [dictItem for dictItem, sup in freqPatternDict.items() if
                set(item).issuperset({dictItem}) and sup == cntItem]
dictOfItems = dict.fromkeys(notClosedSet, -1)
```

دوم اینکه اگر زیرمجموعه ای از این آیتم در دیکشنری ، پرتکرار نباشد، این آیتم هم در دیکشنری پرتکرار نخواهد بود یعنی با مقدار support صفر ذخیره خواهد شد.

```
wrongSubsets = [dictItem for dictItem, sup in freqPatternDict.items() if
                set(item).issuperset({dictItem}) and sup == 0]
if any(wrongSubsets):
    cntItem = 0
freqPatternDict.update({item: cntItem})
```

پس از ایجاد زیرمجموعه های هر سطر و شمارش آنها در تمام سطورهای بعدی، و اتمام فرایند برای هر سطر – row

```
item_sets = set(itertools.combinations(row['Mutation'], subsetLen))
```

الگوهای ایجاد شده در پایان شمارش آیتمها ست های با طول یکسان در تمامی سطور، در فایلی ذخیره میشوند.

```
with open("freqPatternDict.txt", 'w') as f:
    for key, value in freqPatternDict.items():
        f.write('%s:%s\n' % (key, value))
```

و در نهایت تمامی موارد صفر و منفی یک از این فایل حذف میشود و الگوهای پرتکرار در فایل freqPatterns ذخیره می شوند.

نکاتی را برای محدود نمودن جستجو و بهبود زمان اجرا بیان میکنم:

11	TCGA-HP-A5N0-01A-11D-A28X-10	['AC034193.5-3-10035779-RNA', 'FNBP4-11-47788664-In_Frame_Del']	2
12	TCGA-KR-A7K7-01A-11D-A33K-10	['KRTAP10-6-21-46011400-Silent', 'CSMD3-8-113347573-Frame_Shift_Ins']	2
13	TCGA-RC-A7SF-01A-11D-A34Z-10	['RP11-403I13.7-1-149285546-RNA', 'SULF1-8-70514026-Frame_Shift_Del']	2
14	TCGA-DD-AADU-01A-11D-A40R-10	['CDC7-1-91967356-Nonsense_Mutation', 'CTNNB1-3-41266124-Missense_Mutation', 'OR6K6-1-158725536-Missense_Mutation']	3
15	TCGA-K7-A5RG-01A-11D-A28X-10	['KRTAP1-1-17-39197393-Missense_Mutation', 'TP53-17-7576928-Splice_Site', 'CCT6P1-7-65222986-RNA']	3
16	TCGA-G3-AAV1-01A-11D-A382-10	['MYEOV-11-69063836-Missense_Mutation', 'KANK1-9-732477-Silent', 'CTNNB1-3-41266098-Missense_Mutation']	3
17	TCGA-FV-A3R3-01A-11D-A22F-10	['NSUN5P1-7-75045350-RNA', 'GNAS-20-57484420-Missense_Mutation', 'RRN3P2-16-29110438-RNA']	3
18	TCGA-DD-A4NL-01A-11D-A28X-10	['UPF3A-13-115047559-Silent', 'FRG1B-20-29625971-Missense_Mutation', 'CTNNB1-3-41268766-Missense_Mutation']	3

برای زیرمجموعه سه تایی سطور دو تایی شمارش نمیشوند مثلا شمارش زیرمجموعه های سه تایی از سطر ۱۴ شروع میشود.

```
if subsetLen <= row['length']:
```

سطر ۱۵ سطر ۱۴ را برای شمارش چک نخواهد کرد چون این بررسی در هنگام شمارش زیرمجموعه سه تایی سطر ۱۴ به سطرهای بعدی انجام شده است و بنابراین برگشت به عقب نداریم.

نتایج به دست آمده:

```
dataSet = patient_mutations.iloc[0:124, :]
maxlen = 11
```

در نهایت برای ۱۲۳ بیمار الگو مشخص شده و شمارش روی جهش های تمامی بیماران انجام شد و نتایج به دست آمده به صورت دستی با فایل اصلی مقایسه شده و برای موارد مقایسه شده درستی الگوریتم مشهود بود.

نمونه ای از فایل الگوهای پرتکرار با موارد صفر و منفی ۱ آورده شده است. و در نهایت از ۱۴۶۹۱۰ الگوی بررسی شده ۱۰۵۶۸۲ الگو پرتکرارند. که به ترتیب در دو فایل freqPatternDict و freqpatterns قابل مشاهده هستند.

```
C2orf76-2-120078774-Frame_Shift_Del:2
MIR1302-3-2-114340544-RNA:2
NBPF20-1-148341911-Frame_Shift_Del:2
NBPF20-1-148341885-Frame_Shift_Del:2
AC027612.3-2-91888487-RNA:2
NUDT3-6-34256629-Frame_Shift_Del:2
NTRK1-1-156851279-Frame_Shift_Del:2
APH1A-1-150240391-Frame_Shift_Del:2
NAV1-1-201618151-Frame_Shift_Del:2
NEB-2-152348643-Frame_Shift_Del:2
AC027612.3-2-91893901-RNA:2
APOB-2-21234005-Frame_Shift_Del:2
('HIVEP3-1-42041241-Silent', 'ZFPL1-11-64854466-Frame_Shift_Del'):0
('TGFB1-19-41858864-Missense_Mutation', 'RP11-156P1.3-17-45127218-RNA'):0
('OR8H3-11-55890080-Missense_Mutation', 'GTF2IRD2P1-7-72667574-RNA'):0
('IDH1-2-209113113-Missense_Mutation', 'ZNF860-3-32032046-Missense_Mutation'):0
('MT-RNR2-MT-2690-RNA', 'RANBP6-9-6012658-Missense_Mutation'):0
('AC034193.5-3-10035779-RNA', 'FNBP4-11-47788664-In_Frame_Del'):0
('KRTAP10-6-21-46011400-Silent', 'CSMD3-8-113347573-Frame_Shift_Ins'):0
('RP11-403I13.7-1-149285546-RNA', 'SULF1-8-70514026-Frame_Shift_Del'):0
('CTNNB1-3-41266124-Missense_Mutation', 'OR6K6-1-158725536-Missense_Mutation'):0
('CDC7-1-91967356-Nonsense_Mutation', 'OR6K6-1-158725536-Missense_Mutation'):0
('CDC7-1-91967356-Nonsense_Mutation', 'CTNNB1-3-41266124-Missense_Mutation'):0
('KRTAP1-1-17-39197393-Missense_Mutation', 'TP53-17-7576928-Splice_Site'):0
('KRTAP1-1-17-39197393-Missense_Mutation', 'CCT6P1-7-65222986-RNA'):0
('TP53-17-7576928-Splice_Site', 'CCT6P1-7-65222986-RNA'):2
('MYEOV-11-69063836-Missense_Mutation', 'CTNNB1-3-41266098-Missense_Mutation'):0
('KANK1-9-732477-Silent', 'CTNNB1-3-41266098-Missense_Mutation'):0
('MYEOV-11-69063836-Missense_Mutation', 'KANK1-9-732477-Silent'):0
('NSUN5P1-7-75045350-RNA', 'RRN3P2-16-29110438-RNA'):0
('NSUN5P1-7-75045350-RNA', 'GNAS-20-57484420-Missense_Mutation'):0
('GNAS-20-57484420-Missense_Mutation', 'RRN3P2-16-29110438-RNA'):0
('UPF3A-13-115047559-Silent', 'CTNNB1-3-41268766-Missense_Mutation'):0
('FRG1B-20-29625971-Missense_Mutation', 'CTNNB1-3-41268766-Missense_Mutation'):0
('UPF3A-13-115047559-Silent', 'FRG1B-20-29625971-Missense_Mutation'):0
('GPX1-3-49395482-Missense_Mutation', 'ZNF727-7-63538568-Missense_Mutation'):0
```

```
('NBPF22P-5-85581569-RNA', 'CTNNB1-3-41266098-Missense_Mutation'):0
('NBPF22P-5-85581569-RNA', 'SPTA1-1-158614175-Frame_Shift_Del'):0
('NBPF22P-5-85581569-RNA', 'KRT8-12-53298675-Missense_Mutation'):0
('KRT8-12-53298675-Missense_Mutation', 'CTNNB1-3-41266098-Missense_Mutation'):0
('C1orf68-1-152692472-Missense_Mutation', 'CTNNB1-3-41266098-Missense_Mutation'):0
('MN1-22-28194936-Silent', 'LCN9-9-138556579-Missense_Mutation'):0
('MN1-22-28194933-Silent', 'NCOA6-20-33345744-Silent'):2
('MN1-22-28194936-Silent', 'NCOA6-20-33345744-Silent'):2
('LCN9-9-138556579-Missense_Mutation', 'NCOA6-20-33345744-Silent'):0
('MN1-22-28194933-Silent', 'POTEC-18-14513675-Missense_Mutation'):0
('MN1-22-28194936-Silent', 'POTEC-18-14513675-Missense_Mutation'):0
('MN1-22-28194936-Silent', 'MN1-22-28194933-Silent'):3
('LCN9-9-138556579-Missense_Mutation', 'MN1-22-28194933-Silent'):0
('LCN9-9-138556579-Missense_Mutation', 'POTEC-18-14513675-Missense_Mutation'):0
('NCOA6-20-33345744-Silent', 'POTEC-18-14513675-Missense_Mutation'):2
('CCT6P1-7-65226641-RNA', 'KRTAP4-8-17-39254054-Missense_Mutation'):0
('CTNNB1-3-41266097-Missense_Mutation', 'ANKRD36C-2-96643871-Splice_Site'):0
('KRTAP4-8-17-39254054-Missense_Mutation', 'ANKRD36C-2-96643871-Splice_Site'):0
('CTNNB1-3-41266097-Missense_Mutation', 'KRTAP4-8-17-39254054-Missense_Mutation'):0
('CCT6P1-7-65226641-RNA', 'CTNNB1-3-41266097-Missense_Mutation'):3
('CCT6P1-7-65226641-RNA', 'ANKRD36C-2-96643871-Splice_Site'):0
('IL9R-X-155239804-Missense_Mutation', 'ANKRD36C-2-96643871-Splice_Site'):0
('IL9R-X-155239804-Missense_Mutation', 'CTNNB1-3-41266097-Missense_Mutation'):0
('IL9R-X-155239804-Missense_Mutation', 'CCT6P1-7-65226641-RNA'):0
('IL9R-X-155239804-Missense_Mutation', 'KRTAP4-8-17-39254054-Missense_Mutation'):0
('AC068057.1-2-105320362-RNA', 'UBBP4-17-21731261-Frame_Shift_Ins'):0
('RP11-156P1.3-17-45127107-RNA', 'AC068057.1-2-105320362-RNA'):0
('FAM182B-20-25755549-Missense_Mutation', 'AC068057.1-2-105320362-RNA'):0
('UBBP4-17-21731261-Frame_Shift_Ins', 'RP11-509A17.3-15-20559589-RNA'):0
('RP11-156P1.3-17-45127107-RNA', 'FAM182B-20-25755549-Missense_Mutation'):0
('RP11-156P1.3-17-45127107-RNA', 'UBBP4-17-21731261-Frame_Shift_Ins'):0
('RP11-156P1.3-17-45127107-RNA', 'RP11-509A17.3-15-20559589-RNA'):0
```

این شکل ها نشان میدهند الگوهای ۳ آیتمی به ندرت پرتکرارند.

```
('CROCCP2-1-16956410-RNA', 'CCT6P3-7-64528850-RNA', 'OR8H3-11-55890080-Missense_Mutation'):0
('BTN2A3P-6-26422353-RNA', 'POTEG-14-19553528-Missense_Mutation', 'CROCCP2-1-16956410-RNA'):0
('BTN2A3P-6-26422353-RNA', 'POTEG-14-19553528-Missense_Mutation', 'CCT6P3-7-64528850-RNA'):0
('BTN2A3P-6-26422353-RNA', 'POTEG-14-19553528-Missense_Mutation', 'MTHFD1-14-64914955-Silent'):0
('BTN2A3P-6-26422353-RNA', 'POTEG-14-19553528-Missense_Mutation', 'IGKV1D-16-2-90139477-RNA'):0
('C9orf152-9-112963511-Missense_Mutation', 'DSC2-18-28648117-Missense_Mutation', 'ZNF814-19-58385748-Missense_Mutation'):0
('C9orf152-9-112963511-Missense_Mutation', 'MTMR8-X-63445128-Missense_Mutation', 'EEF1A1-6-74227628-Missense_Mutation'):0
('C9orf152-9-112963511-Missense_Mutation', 'KIF1B-1-10381913-Missense_Mutation', 'EEF1A1-6-74227628-Missense_Mutation'):0
('C9orf152-9-112963511-Missense_Mutation', 'MTMR8-X-63445128-Missense_Mutation', 'ZNF814-19-58385748-Missense_Mutation'):0
('C9orf152-9-112963511-Missense_Mutation', 'KIF1B-1-10381913-Missense_Mutation', 'ZNF814-19-58385748-Missense_Mutation'):0
('C9orf152-9-112963511-Missense_Mutation', 'DSC2-18-28648117-Missense_Mutation', 'MTMR8-X-63445128-Missense_Mutation'):2
('EEF1A1-6-74227627-Missense_Mutation', 'C9orf152-9-112963511-Missense_Mutation', 'DSC2-18-28648117-Missense_Mutation'):0
('EEF1A1-6-74227627-Missense_Mutation', 'DSC2-18-28648117-Missense_Mutation', 'KIF1B-1-10381913-Missense_Mutation'):0
('MTMR8-X-63445128-Missense_Mutation', 'EEF1A1-6-74227628-Missense_Mutation', 'ZNF814-19-58385748-Missense_Mutation'):0
('EEF1A1-6-74227627-Missense_Mutation', 'C9orf152-9-112963511-Missense_Mutation', 'KIF1B-1-10381913-Missense_Mutation'):0
('DSC2-18-28648117-Missense_Mutation', 'EEF1A1-6-74227628-Missense_Mutation', 'ZNF814-19-58385748-Missense_Mutation'):0
('EEF1A1-6-74227627-Missense_Mutation', 'EEF1A1-6-74227628-Missense_Mutation', 'ZNF814-19-58385748-Missense_Mutation'):0
('DSC2-18-28648117-Missense_Mutation', 'MTMR8-X-63445128-Missense_Mutation', 'KIF1B-1-10381913-Missense_Mutation'):2
('EEF1A1-6-74227627-Missense_Mutation', 'DSC2-18-28648117-Missense_Mutation', 'EEF1A1-6-74227628-Missense_Mutation'):0
('DSC2-18-28648117-Missense_Mutation', 'KIF1B-1-10381913-Missense_Mutation', 'ZNF814-19-58385748-Missense_Mutation'):0
('EEF1A1-6-74227627-Missense_Mutation', 'MTMR8-X-63445128-Missense_Mutation', 'KIF1B-1-10381913-Missense_Mutation'):0
('EEF1A1-6-74227627-Missense_Mutation', 'DSC2-18-28648117-Missense_Mutation', 'ZNF814-19-58385748-Missense_Mutation'):0
('EEF1A1-6-74227627-Missense_Mutation', 'C9orf152-9-112963511-Missense_Mutation', 'EEF1A1-6-74227628-Missense_Mutation'):0
('MTMR8-X-63445128-Missense_Mutation', 'KIF1B-1-10381913-Missense_Mutation', 'EEF1A1-6-74227628-Missense_Mutation'):0
('KIF1B-1-10381913-Missense_Mutation', 'EEF1A1-6-74227628-Missense_Mutation', 'ZNF814-19-58385748-Missense_Mutation'):0
('DSC2-18-28648117-Missense_Mutation', 'MTMR8-X-63445128-Missense_Mutation', 'EEF1A1-6-74227628-Missense_Mutation'):0
('EEF1A1-6-74227627-Missense_Mutation', 'MTMR8-X-63445128-Missense_Mutation', 'EEF1A1-6-74227628-Missense_Mutation'):0
('EEF1A1-6-74227627-Missense_Mutation', 'C9orf152-9-112963511-Missense_Mutation', 'ZNF814-19-58385748-Missense_Mutation'):0
('MTMR8-X-63445128-Missense Mutation', 'KIF1B-1-10381913-Missense Mutation', 'ZNF814-19-58385748-Missense Mutation'):0
```

الگوهایی با طول۳ آیتم یا بیش از ۳ آیتم به ندرت پرتکرارند و بعضا پرتکرار نمیباشند شکل بالا الگوهای ۷ و ۸ آیتمی را نمایش میدهد که در آن صفر به معنای support کمتر از minsup میباشد که برای راحتی در پردازش با صفر نمایش داده شده اند.

نتیجه گیری:

این پروژه برای من بسیار جذاب بود و باعث شد موارد زیادی یاد بگیرم البته با تغییر کد زنی و بهینه سازی آن امکان بهبود سرعت برای الگوریتم مربوطه وجود دارد و می توان با پیاده سازی OOP و ایجاد متودهای مربوط به هر بخش این برنامه را به طور موثر در موارد مشابه به کار برد. در صورت کاربردی بودن امکان انجام آن در آینده وجود دارد.

در آینده امکان این وجود دارد که برای تست هم الگوریتمی طراحی شود تا بررسی دقت الگوریتم به صورت دستی نباشد.