فاز اول پروژه بیوانفورماتیک

نگارعسکری ۹۹۱۰۵۶۱۲ سوگندصالحی ۹۹۱۰۵۵۵۹

۱. Microarray یک تکنولوژی با تعداد زیادی چیپ می باشد و بر روی هر چیپ تعداد زیادی پیکسل قرار دارد و بر روی هر کدام از پیکسل ها یک توالی از دی ان ای تک رشته ای. این به ما این امکان را می دهد که از یک سمپل، RNA و از روی آن cDNA را به صورت قطعات کوچک تولید کنیم. که این پروسه ی تبدیل به قطعات کوچک، با shaking یا sonification انجام می گیرد. به دلیل تعداد زیاد دی ان ای، آن‌ها هر بار از جاهای مختلف (به صورت رندوم) خرد می شود. نهایتا آن cDNA را با یک رنگ فلورسنت رنگ کرده و آن را روی چیپ قرار می دهیم (البته بعد از گرم شدن اولیه). هر قطعه ای مکمل خودش را روی چیپ می یابد و به آن متصل می شود. آن را wash می کنیم تا قطعاتی که به مکمل درست خود نچسبیده اند دور ریخته شوند. در نهایت آن چیپ را به دستگاه مختص خودش می دهیم تا آن را بخواند و به ما نشان دهد که هر پیکسل چه مقدار رنگ فلورسنت در خود دارد. دستگاه اسکنر از ورودی های microarray یک سری تصاویر درست می کند و خود پردازش اولیه روی آنها انجام می دهد و خروجی نهایی آن می تواند یک ماتریس عددی باشد. در این ماتریس عددی به تعداد سمپل ها ستون داریم و به تعداد پروب ها سطر. این ماتریس در این پروژه عملا همان gene expression matrix می باشد.

۲. برای بررسی کیفیت داده، نمودار boxplot ماتریس را رسم می کنیم تا ببینیم که پراکندگی داده ها در کل به چه صورت می باشد. با تحلیل نمودار، می توانیم ببینیم که بازه ی کلی داده ها حدودا بین ۱ تا ۱۴ می باشد و همچنین مدیان آن ها تقریبا در یک بازه ی کوچک و محدود یکسان می باشد. اگر بجای بازه ی ۱ تا ۱۴، اعداد بزرگتری همچون ۴۰۰۰ داشتیم، نیاز بود که داده را log-scale، و اگر مدیان ها پراکنده بودند، آن را نرمالایز کنیم. اما در شرایط فعلی داده از پیش log-scaled و normalized می باشد و در نتیجه لازم به انجام دوباره ی این کار نیست. به هر حال اگر می خواستیم اینکارها را انجام دهیم به چند نکته باید توجه می کردیم. یک اینکه برای log-scale کردن لازم بود که $\log_2(ex+1)$ را محاسبه کنیم تا از log(0) پرهیز شود و همچنین برای normalize کردن می توانستیم از normalizeQuantiles(ex) استفاده کنیم. در هر حالت، از آنجایی که ما در کل با خود ابجکت به کار ادامه می دهیم، پس از تغییرات لازم است که مقدار جدید را در exprs(gset) قرار دهیم. بدیهتا کنترل ما باعث می

شود که در یک بازه ی مشخص قرار داشته باشیم و تغییرات بیش از اندازه که تاثیر اریب در خروجی نهایی ما دارند از بین ببریم. در کل اگر داده ای از قبل نرمالایز نشده باشد، ممکن است که الگوریتم ما، اشتباها بجای اهمیت دادن به ویژگی هایی که مهم تر می باشند، به تغییرات absolute مقادیر توجه کند. و این می تواند در visualization و analysis خود ما نیز گمراه کننده باشد. (مراجعه به boxplot.pdf در فولدر results)

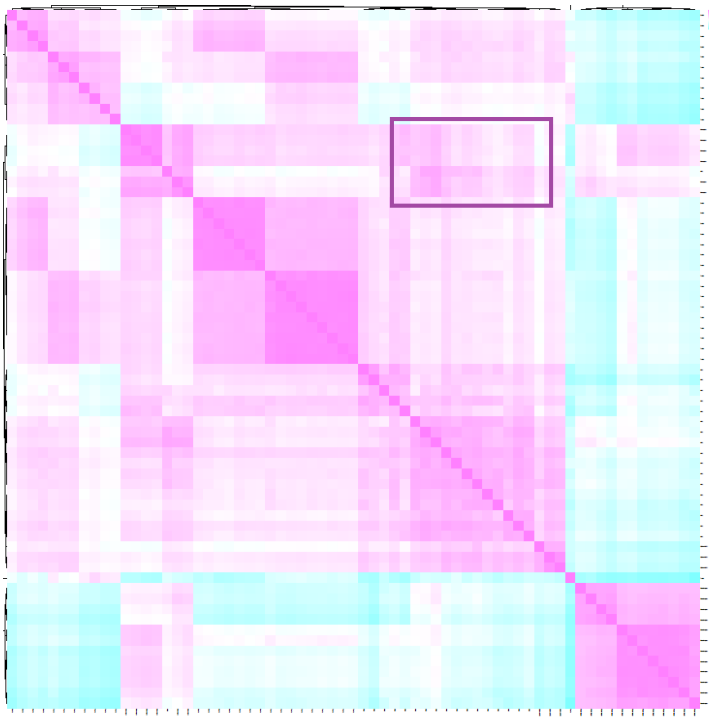
۳. کاهش بعد در کل کمک های بسیاری در فهم و تحلیل داده می کند. عملا با کاهش بعد، ما تعداد ویژگی های مهم تری از فضا را حفظ می کنیم و سایر ویژگی هایی که آنقدر اطلاعات زیادی به ما نمی دهند (و گاهی حتی حاوی نویز هستند) حذف می کنیم. با این کار، پیچیدگی داده کمتر می شود و می تواند ما را نهایتا به داده ی جدیدی برساند که دسته بندی و بررسی در آن راحت تر می باشد. این قدم باعث می شود که visualization به مراتب ساده تر شود.

با این سه روش کاهش بعد را انجام داده و نمودارهای نتیجه را بررسی کردیم. نمودار ها در فولدر results به نام های، MDS.pdf و PCA_samples.pdf و tSNE.pdf قرار دارند.

همانطور که می بینیم، دسته ها در tSNE.pdf تفکیک پذیر تر بوده و ما را به طبقه بندی مناسب تری می رسانند که به ما دید بهتری از تفاوت های این گروه ها با توجه به ابعاد پراهمیت‌ترشان می دهد. در بقیه ی روش ها، داده ها از گروه های مختلف در هم پخش هستند و یافتن الگوریتمی که آن ها را از هم تفکیک کند، سخت تر می باشد.

۴. source name به ما نشان می دهد که سمپل از چه نوع سلولی گرفته شده است. می بینیم که آپشن های B cells، T cells، Granulocytes، CD34+ و Monocytes وجود دارند.

با رسم نمودار heatmap که در فولدر Results در فایل CorHeatmap_second.pdf قرار دارد می‌توانیم بررسی روابط هم بستگی را انجام دهیم. همانطور که در نمودار مشخص است، کمترین هم بستگی AMLها با دسته ی Granulocytes می باشد. سایر سلول ها هم بستگی تقریبا یک شکلی با AMLها (با توجه به رنگ نمودار) دارند، ولی تا حدی مشخص است که Monocytes هم بستگی بیشتری دارند.



سلول های سالمی را که شباهت بیشتری به AML دارند انتخاب می کنیم زیرا می خواهیم که تفاوت اصلی شان همان داشتن یا نداشتن بیماری باشد. این بررسی می تواند insight بهتری نسبت به آنالیز کلی مان بدهد و باعث می شود ژن هایی که تفاوت معنی دار expression در بین این دو گروه دارند به احتمال بیشتری مربوط به بیماری باشند.

برای درک بهتر و نمایش، یک tSNE دیگر نیز برای حالتی که گروه های source name اهمیت دارند، رسم شده است. (فایل tSNE_second.pdf در فولدر results)