پروژه‌ی مقدمه‌ای بر بیوانفورماتیک
فاز اول

مهدی قیدی - ۹۸۱۰۰۳۲۳
امید آزادی - ۹۸۰۰۰۰۰۰
نوید اسلامی - ۹۸۰۰۰۰۰۰

۱۶ آذر ۱۴۰۱

در این فایل به پاسخ دادن هر کدام از سؤالات مطرح شده در دستور کار می‌پردازیم، و نیز روند کلی صورت گرفته برای انجام این فاز از پروژه را توصیف می‌کنیم. نمودارهای کشیده شده همگی در پوشه‌ی Results هستند، کد زده شده در پوشه‌ی Source، و نیز داده‌های دریافت شده در پوشه‌ی Data هستند.

ابتدا، برای این که با زبان R و نیز دستگاه Microarray آشنا شویم، تمام اعضای گروه ویدیوهای لینک شده از کلاس بیوانفورماتیک پیشرفته‌ی دکتر شریفی را نگاه کردیم. همراه با نگاه کردن این ویدیوها، خود R، RStudio و نیز پیشنیازهای مورد نیاز برای کارکرد کتابخانه‌های خود را نصب کردیم. سپس، به آماده کردن داده‌ها، نوشتن کدهای لازم و نیز نوشتن این گزارش پرداختیم. در تمام مراحل انجام این پروژه، افراد گروه مشارکت برابر داشتند.

۱ توصیف و خروجی Microarrayها

این ابزار، درواقع یک Chip از سطوح کوچک و پیکسل مانندی است که هر کدام یک Probe نام دارند. از نظر تئوری، این Probeها مشابه توالی‌های تک رشته‌ای DNA هستند که به سیلیکون متصل شده‌اند و در نتیجه، می‌توانند با رشته‌های cDNA پیوند برقرار کنند. اما با توجه به این که صرفاً پیوند برقرار می‌کنند و نیز چون طول این توالی‌ها کم است، هر Probe دقیقاً مربوط به یک ژن و یک توالی خاص نیست، بلکه می‌تواند به یک تعدادی ژن وصل شود.

این دستگاه با هدف خاص بررسی میزان بیان ژن ساخته نشده است، اما از آن در این زمینه به خوبی می‌توان بهره برد. برای بررسی میزان بیان ژن، رویکرد این است که اول از DNA اولیه‌ی خود در سلول‌های مد نظر، RNA بسازیم و سپس از روی این RNAها که می‌توانیم آن‌ها را به دست آوریم، cDNA بسازیم. از این cDNAها، بعد از این که خردشان کردیم، می‌توانیم استفاده کنیم و آن‌ها را به Probeهای خود وصل کنیم. (دو روش برای خرد کردن این cDNAها داریم، که یکی با تکان دادن معمولی و شدید رخ می‌دهد، و دیگری با استفاده از Sonication است. در هر دوی این روش‌ها، چون تعداد قطعات زیاد است، از جاهای تصادفی خرد می‌شوند.)

اما قبل از این که این cDNA را به چیپ اضافه کنیم، آن را باید با استفاده از یک رنگ فلورسانت رنگ کنیم. سپس، این مخلوط را رنگ می‌کنیم تا دو رشته‌ی cDNA از هم جدا شوند و سپس آن را روی چیپ می‌ریزیم. در این حالت، هر رشته مکمل خود را روی Probeها پیدا می‌کند و به آن‌ها می‌چسبد. در آخر، چیپ را می‌شوریم تا اجزای cDNAای که جفت نشده بودند جدا شوند و چیپ تمیز شده را در یک اسکنر لیزری می‌گذاریم تا مشخص کند که کدام Probeها بیشتر از این cDNAها را به خود جذب کرده است، که با استفاده از میزان نوری که دیده می‌شود مشخص می‌شود. به این صورت، یک تخمین از میزان Expression هر کدام از این Probeها خواهیم داشت، که همانطور که گفتیم، لزوماً به معنی زیاد یا کم بودن یک تک ژن نیست، بله به معنی زیاد یا کم بودن یک زیرمجموعه‌ای از ژن‌های ممکن است. (ژن‌های Isomorph با هر Probe.)

در نتیجه، خروجی این دستگاه صرفاً یک آرایه از میزان Expressionها یا درواقع میزان پیوندی که هر Probe ایجاد می‌کند خواهد بود. تبدیل این میزان نور دیده شده به عدد اما با استفاده از دستگاه‌های مخصوص آن صورت می‌گیرد. پس چون که چند نمونه را امتحان می‌کنیم، به ازای هر داده یک ستون از Expressionها می‌گیریم و در آخر به یک ماتریس عددی می‌رسیم که باید آن را تحلیل کنیم.

۲ Quality Control of Data

قبل از این که به توصیف فرایندهای طی شده در این بخش بپردازیم، لازم است ذکر کنیم که داده‌ها را با استفاده از Accession Number آن‌ها از GEO دریافت کردیم و نیز دسته‌بندی گفته شده را نیز روی آن‌ها در ابزار GEO2R اجرا کردیم. این دسته بندی تمام نمونه‌ها را در بر نمی‌گرفت اما، که باعث شد تصمیم بگیریم که تمام نمونه‌های دیگر را دور بریزیم و فقط از آن‌ها استفاده کنیم تا نمودارها و تحلیل‌های ما ساده شود.

سپس، با کمک گرفتن از کد موجود در GEO2R، رشته‌ی مربوط به دسته‌بندی داده‌ها را کپی کردیم و با استفاده از Levelها، نام‌های مناسب را به هر کدام دادیم و داده‌های نامربوط را دور ریختیم. جزئیات انجام این فرایند را در متن کد ما می‌توانید مشاهده کنید، که بلافاصله بعد از لود کردن داده‌ها صورت می‌گیرد.

حال به بررسی کیفیت داده‌ها می‌پردازیم. اولین نکته‌ای که بررسی کردیم، مقدار بیشینه و کمینه‌ی داده‌ها بود، که بفهمیم در اسکیل لگاریتمی هستند یا نه. چون که مقدار بیشینه‌ی آن کمتر از ۱۵ بود و کمینه‌ی آن بیشتر از ۱، می‌فهمیم که در اسکیل لگاریتمی هستیم و نیازی به لگاریتمی کردن اسکیل داده‌ها نداریم. (این مقادیر را با چاپ کردن آن‌ها در خود کد R بررسی کردیم. لزوم انجام این کار نیز در این بود که دوست داریم اعداد ما بیان‌گر مرتبه‌ی بزرگی باشند تا واقعاً تفاوت‌های غلطت را بتوانیم تشخیص دهیم، و نیز بتوانیم محاسبات بهتر و Stableتری را در الگوریتم‌هایی می‌خواهیم برای تحلیل آن‌ها بزنیم داشته باشیم و دچار Overflow نشویم. (به طور مثال، محاسبه‌ی Correlation به مراتب دقت بالاتری خواهد داشت اگر اعداد ما خیلی بزرگ نباشند.)

در مرحله‌ی بعد، چیزی که باید بررسی می‌کردیم توزیع چارکی هر کدام از Probeها یا ژن‌ها در تمام سمپل‌ها ما بود. لازم است که این توزیع نرمالایز شده باشد، چون در غیر این صورت، در الگوریتم‌هایی که در ادامه می‌زنیم نمی‌توانیم مقایسه بین مقادیر Probeها داشته باشیم. همچنین، هر Probe چون بایاس‌هایی در اندازه‌گیری دارد، دوست داریم که آن‌ها را هم کنار بگذاریم تا بتوانیم باز مقایسه‌های بهتری انجام دهیم بین مقادیر چند Probe. پس با کشیدن یک Box Plot از داده‌های هر Probe، به نمودار موجود در فایل Boxplot.pdf رسیدیم، که نشان می‌داد داده‌های ما از توزیع چارکی خیلی یکنواخت و خوبی پیروی می‌کنند و لازم نیست که دستی آن‌ها را Normalize کنیم.

سپس، چون که یک سری از الگوریتم‌هایی که در ادامه می‌زنیم وابسته به صفر بودن میانگین کلی هر کدام از پارامترهای حاصل از Probe هستند، (به خصوص PCA) تصمیم گرفتیم که میانگین مقادیر هر کدام از این Probeها را بررسی کنیم و در صورتی که مقادیر بسیار متفاوتی داشتند، آن‌ها را صفر کنیم. توجه کنید که این صفر کردن الگوریتم‌های آینده‌ی ما را Stableتر می‌کند و به این قابلیت را می‌دهد که خروجی‌های بهتری داشته باشیم و مثلاً PCهایی که در PCA به دست می‌آوریم به فیچرهای نامفید داده‌ها وابسته نشده باشند. در همین راستا، از هر سطر که مربوط به هر کدام از Probeهای ما بود میانگین گرفتیم و حاصل را در فایل Gene_Means.pdf نمایش دادیم. همانطور که مشاهده می‌شود، واریانس به شدت زیادی برای این میانگین‌ها داریم، که باعث شد که تصمیم بگیریم همه‌ی آن‌ها را صفر کنیم. در نتیجه، با استفاده از تابع scale این کار را انجام دادیم و حاصل را در یک جای دیگری ذخیره کردیم. نمودار میانگین‌های نهایی را نیز می‌توانید در Gene_Means_Zeroed.pdf مشاهده کنید، که نشان می‌دهد که این مقادیر به شدت نزدیک به صفر شده‌اند.

به طور خاص، می‌توانیم تأثیر این فرایند را در دو فایل PCA/PCA_Genes_Not_Zeroed.pdf و PCA/PCA_Genes.pdf ببینیم. این دو فایل روی خود مقادیر Probeها PCA زده‌اند و ابعاد آن‌ها را کاهش داده‌اند تا بتوانیم آن‌ها را در صفحه رسم کنیم. در فایل اول مشاهده می‌شود که PC1 یک

واریانس به شدت بیشتری نسبت به سایر PCها دارد، در حالی که در فایل دوم یک توزیع واریانس نرم‌تری داریم. این به خاطر همین موضوع است که این واریانس زیاد به دلیل دور بودن میانگین‌ها و بیشینه و کمینه‌ی مقادیر Probeها مختلف است، که ناخواسته مقدار واریانس را زیاد می‌کند. همچنین، اگر Scatter Plot این دو فایل را مشاهده کنیم، می‌بینیم که در فایل اول داده‌های خیلی Skewed هستند و توزیع نرمالی که برای ما ایده‌آل هست را ندارند. اما این موضوع در فایل دوم حل شده است و تا حد خوبی یک توزیع نرمال مشاهده می‌کنیم. با این اوصاف، می‌فهمیم که این کم کردن میانگین کار ما را بهتر کرده است.

در آخر نیز، به نظر ما، لازم است که Correlation Matrix را نیز بررسی کرد خود نمونه‌ها تا مطمئن شویم که مجموعه‌ی نمونه‌های ما Outlier ندارد. با رسم این نمودار، متوجه می‌شویم که نمونه‌ی GSM1180835 که تقریباً با همه Correlation منفی دارد و عملاً از همه‌ی نمونه‌های دیگر مجزا است. پس این نمونه را از داده‌های خود حذف می‌کنیم، و از Corr_Heatmap.pdf به نمودار Correlation Heatmap جدید Corr_Heatmap_No_Outliers.pdf می‌رسیم.

همچنین، در Corr_Heatmap.pdf یک دسته از نمونه‌های سالم را پیدا می‌کنیم که فقط به هم شبیه هستند و از همه‌ی نمونه‌های دیگر متفاوت هستند، به دلیل مقادیر مشهود در Correlationهای آن‌ها. اما از آن‌جایی که این بخش را باید در سؤال آخر بررسی کنیم، توضیحات بیشتر آن را به بخش آخر موکول می‌کنیم.

۳ Dimension Reduction

کاهش دادن ابعاد داده‌های ما دو دلیل عمده دارد:

- در آخر دوست داریم که یک مدل و یا متر برای تشخیص سلول‌های سرطانی از سلول‌های سالم به دست آوریم. اما چون مجموعه‌ی داده‌ی ما در مقابل مجموعه‌ی Featureهای ما خیلی کوچک است، لازم است که تعداد فیچرها را کم کنیم که پیچیدگی مدل نهایی ما نیز کم شود و از خطر رخ دادن مشکلاتی همچون Overfitting دوری کنیم. همیچنن، با کم کردن این تعداد فیچرها، می‌توانیم واقعاً متر مهم را به دست آوریم و مترها و پارامترهایی که اهمیت کمی دارند را دور بریزیم.
- همچنین، دوست داریم که بتوانیم داده‌های خود را نمایش دهیم و از روی شکل آن‌ها در مورد صحت نمونه برداری و نیز نتیجه‌گیری‌ها، نظر دهیم. به طور مثال، انتظار داریم که نمونه‌های سرطانی تا حد خوبی پخش باشند، چون این موضوع از نظر مباحث Epigenics برقرار باید باشد.

در نتیجه، با این دو هدف، سراغ آزمایش کردن روش‌های مختلف کاهش بعد می‌رویم. هر کدام از این روش‌ها را در زیر توصیف می‌کنیم و نتایج آن‌ها را نقد می‌کنیم.

۱.۳ PCA

این روش سعی می‌کند به صورت خطی داده‌ها را در یک فضای کوچک‌تر طوی تصویر کند که واریانس‌های داده‌های Preserve شوند. در این روش، اگر بعد نمونه‌ها را کم کنیم و نمودار توزیع

آن‌ها را در فایل PCA/PCA_Samples.pdf رسم کنیم، می‌بینیم که یک سری از داده‌های سالم به خوبی از داده‌های سرطانی تفکیک می‌شوند، اما یک سری نیز در خیلی نزدیکی آن‌ها قرار می‌گیرند. این Margin کم می‌تواند باعث شود که مدلی که در آخر برای Classification استفاده کنیم خوب عمل نکند و دوست داریم که چنین حالتی نداشته باشیم و واقعا Clusterهایی دور از هم داشته باشیم.

MDS ۲.۳

این روش سعی می‌کند که داده‌ها را در فضای با بعد کمتر طوری بچیند که ماتریس فواصل دو به دوی نمونه‌های ما تا جای ممکن مشابه حالت اولیه‌ی High-Dimensional باشد. این الگوریتم دو نوع Metric و Non-Metric دارد، که در این مجموعه‌ی نمونه‌ی خاص، دقیقاً مشابه هم عمل می‌کنند و یک خروجی را به ما می‌دهند. همچنین، اگر کمی بررسی کنیم، می‌بینیم که خروجی حاصل از این الگوریتم عملاً همان خروجی حاصل از PCA است که صرفاً Flip شده است. در نتیجه، همه‌ی بدی و خوبی‌های آن را نیز دارد. نمودارهای مربوط به خروجی‌های این الگوریتم فایل‌های MDS/Metric_MDS_Samples.pdf و MDS/Non-Metric_MDS_Samples.pdf هستند.

t-SNE ۳.۳

این الگوریتم اما با ورودی گرفتن یک Perplexity، مکان مناسب داده‌ها را در دو بعد برای ما مشخص می‌کند. همانطور که در فایل خروجی t-SNE/t-SNE_Samples.pdf مشاهده می‌شود، این روش کاهش بعد بهترین عملکرد را دارد و در کنار داشتن خوبی‌های دو روش قبل، Margin بزرگ‌تری برای داده‌ها ایجاد می‌کند و می‌توان از آن برای Classification بهتر استفاده کرد.

پس چون که t-SNE هم تفکیک‌های دسته‌های دور را به خوبی انجام می‌دهد، و چون Margin بیشتری بین داده‌های سالم و سرطانی ایجاد می‌کند، بهتر است که از آن برای کاهش ابعاد خود استفاده کنیم. توجه کنید که اینجا نیز یکنواخت پخش شدن نمونه‌های سرطانی را به خوبی می‌توانیم مشاهده کنیم، که یکی دیگر از مترهایی بود که طبق تئوری اپیژنیک انتظار داشتیم. این در حالی است که همین پخش شدن در دو روش دیگر کمتر مشهود است.

Correlation Heatmap ۴

فیلد Source Name مشخص می‌کند که نمونه‌ی مورد نظر از کدام سلول‌ها، محیط‌ها و یا افراد گرفته شده است. به طور مثال، سلول‌ها بیمار همگی از Source با نام AML Patient گرفته شده‌اند، که به این معنی است که از یک فرد بیمار آمده است. اما سلول‌های سالم از دسته‌هایی مثل T Cells، B Cells و غیره آمده است، که به معنی این است که از سلول‌های مختلف از افراد سالم تهیه شده‌اند.
حال، برای بررسی همبستگی بین داده‌ها، به نمودار مرسوم در فایل Corr_Heatmap_No_Outliers.pdf نگاه می‌کنیم. همانطور که در این نمودار دیده می‌شود، نمونه‌های سالمی که از منبع Granulocytes هستند، همگی با هم Correlation خوبی دارند، اما با اکثر نمونه‌های دیگر، به خصوص با نمونه‌های سرطانی، Correlation منفی دارند. این به این معنی است که این نمونه‌ها به احتمال خوبی از یک توزیع نامرتبطی پیوری می‌کنند و می‌توانیم آن‌ها را به طور کلی از داده‌های خود حذف کنیم، چون داده‌های دیگر ما آن‌هایی هستند که واقعاً مهم است در آن‌ها Differentiator پیدا کنیم.

سایر گروه‌ها را که بررسی کنیم، متوجه می‌شویم که منبع CD34+HSPC از بقیه‌ی نمونه‌ها کمی همبستگی بیشتری با داده‌های سرطانی دارند، خصوصاً چون که در درختی که در این نمودار کشیده شده‌اند در فاصله‌ی کمتری قرار دارند. اما می‌توان دید که عملاً سایر گروه‌ها نیز Correlation نابدیهی‌ای که با نمونه‌های سرطانی نیز دارند، و بهتر است که در تحلیل‌های خود آن‌ها را نیز در نظر داشته باشیم و صرفاً دسته‌ای که بالا حذف کردیم را واقعاً حذف کنیم.

لزوم انتخاب این دسته‌ی با همبستگی بالا در این است که هدف ما پیدا کردن یک متر Non-Trivial برای جدا کردن نمونه‌های سالم و سرطانی شبیه به هم است. در نتیجه، نمونه‌هایی که دارای همبستگی بالایی هستند شباهت شایانی نیز دارند و خوب است که در تحلیل‌های خود برای پیدا کردن چنین متری، از این داده‌ها استفاده کنیم تا واقعاً بتوانیم یک Differentiator خوب و مفید برای تشخیص سرطان پیدا کنیم. (درواقع وقت خود را برای پیدا کردن یک جداساز برای داده‌های واضحاً متفاوت دور نریزیم.)