فاز اول

پروژه مقدمه‌ای بر بیوانفورماتیک - دکتر علی شریفی‌زارچی و دکتر سمیه کوهی

دانشکده مهندسی کامپیوتر

دانشگاه صنعتی شریف

نیم‌سال اول ۰۱-۰۲

امیرحسین باقری - ۹۸۱۰۵۶۲۱
مهدی مستانی - ۹۷۱۰۰۵۱۳
محمدرضا مفیضی - ۹۸۱۰۶۰۵۹

۱ ریزآرایه چیست؟

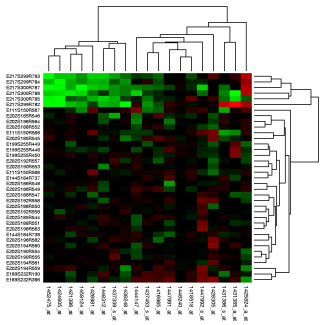
ریزآرایه [۱]، ابزاری آزمایشگاهی است که برای تشخیص بیان هزاران ژن به طور همزمان استفاده می‌شود. ریزآرایه‌های DNA لام‌های میکروسکوپی هستند که با هزاران نقطه کوچک در موقعیت‌های مشخص چاپ می‌شوند و هر نقطه حاوی یک توالی DNA یا ژن شناخته‌شده است.

روش کار

برای انجام تحلیل ریزآرایه، مولکول‌های mRNA معمولاً از هر دو نمونه آزمایشی و نمونه مرجع جمع‌آوری می‌شوند. به عنوان مثال، نمونه مرجع را می‌توان از یک فرد سالم، و نمونه آزمایشی را می‌توان از یک فرد مبتلا به بیماری مانند سرطان جمع‌آوری کرد. سپس دو نمونه mRNA به DNA مکمل (cDNA) تبدیل می‌شوند و هر نمونه با یک ترکیب فلورسنت [۲] با رنگ متفاوت برچسب‌گذاری می‌شود. مثلا، نمونه آزمایشی cDNA ممکن است با رنگ فلورسنت قرمز برچسب‌گذاری شود، در حالی که cDNA مرجع با رنگ فلورسنت سبز برچسب‌گذاری می‌شود.
سپس دو نمونه با هم مخلوط شده و اجازه داده می‌شود تا به لام ریزآرایه متصل شوند. فرآیندی که در آن مولکول‌های cDNA به ترکیب‌های DNA روی لام متصل می‌شوند، هیبریداسیون [۳] نامیده می‌شود. پس از هیبریداسیون، ریزآرایه برای اندازه‌گیری میزان بیان هر ژن چاپ‌شده روی لام اسکن می‌شود. اگر بیان یک ژن خاص در نمونه آزمایشی بیشتر از نمونه مرجع باشد، نقطه مربوطه روی ریزآرایه قرمز به نظر می‌رسد. از طرفی، اگر بیان در نمونه آزمایشی کمتر از نمونه مرجع باشد، آن نقطه سبز به نظر می‌رسد. در نهایت، اگر میزان بیان در دو نمونه یکسان باشد، نقطه زرد خواهد بود. داده‌های جمع‌آوری‌شده از طریق ریزآرایه‌ها را می‌توان برای ایجاد پروفایل‌های بیان ژن، که تغییرات همزمان در بیان بسیاری از ژن‌ها در پاسخ به یک بیماری یا درمان خاص را نشان می‌دهد، استفاده کرد. [۱]

فرمت داده‌های خروجی

مجموعه داده‌های ریزآرایه معمولاً بسیار بزرگ هستند و فرمت داده‌های خروجی به صورت یک فایل خام (Raw Matrix) در قالب یک متن tab-seperated حاوی داده‌های بیش از یک سنجش ترکیبی (ترکیب‌ها در سطر و نتایج آزمایش‌ها در ستون) است. در تصویر ۱ میزان بیان هر ژن به صورت heatmap نمایش داده شده است.



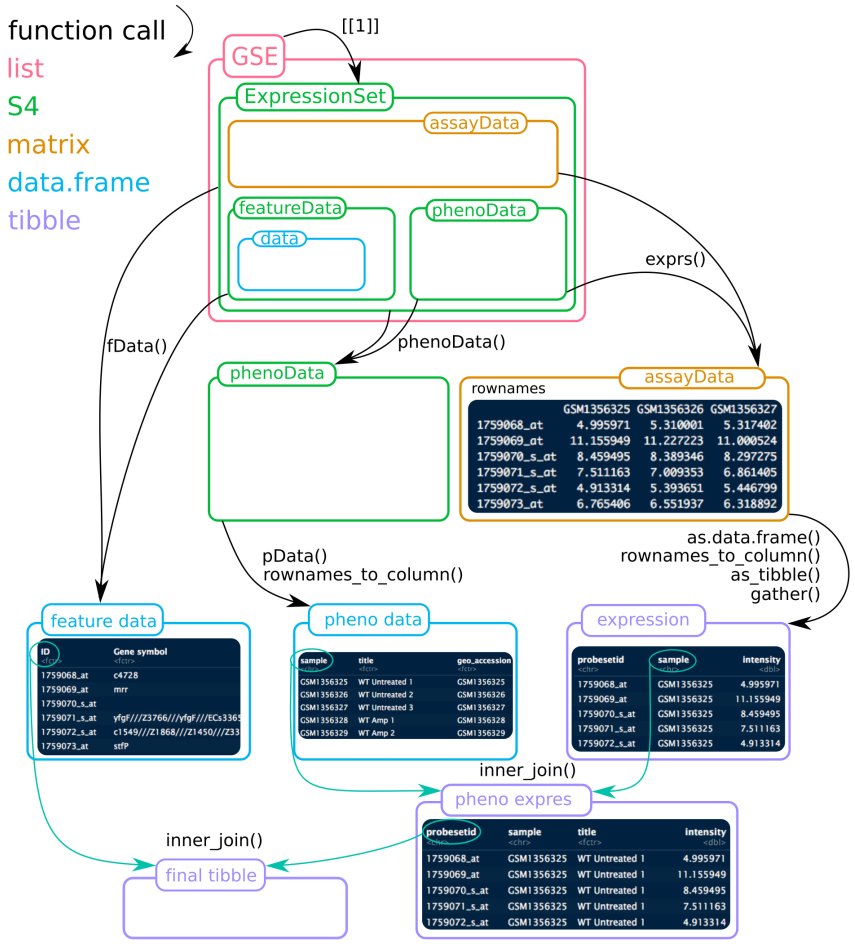
شکل ۱: میزان بیان ژن در ریزآرایه [۲]

همچنین داده‌های ریزآرایه در دیتاست با توجه به شکل ۲ قابل دسترسی خواهد بود.

۲ کیفیت داده‌ها

در ابتدا بررسی می‌کنیم که بیشترین و کمترین مقدار داده‌ها در ماتریس بیان چقدر است. مشاهده می‌شود که بیشترین، عددی برابر با 13.76154 و کمترین برابر با 1.611473 است که نتیجه گیری این است که داده‌ها نرمال است(اگر که بیشترین مقدار عددی بیشتر از ۱۰۰ بود از لگاریتم داده‌ها استفاده می‌کردیم تا نتیجه‌گیری‌های بهتری داشته باشیم). در ادامه نیز کد لازم در صورتی که داده‌ها نیاز به تبدیل به مقیاس

[1] microarray
[2] fluorescent
[3] hybridization



شکل ۲: ساختار داده‌ها در دیتاست [۳]

لگاریتمی داشته باشیم، قرار داده‌شده‌است تا صورت نیاز، این کار انجام شود(برای مثال اگر اختلاف ابتدا و انتهای بازه داده‌ها بیشتر از ۵۰ باشد یا ۱ درصد انتهای داده‌های عددی بیشتر از ۱۰۰ باشد).

برای بررسی‌های بیشتر در رابطه با آنالیز کیفیت داده‌ها، نمودار Adjusted P-value را بررسی می‌کنیم. همان‌طور که در نمودار مشاهده می‌شود، توزیع Adjusted P-value ها بدین صورت است که بخش قابل توجهی از آن‌ها در بازه 0 تا 0.05 قرار دارند که بیان‌گر این موضوع است که در بخش قابل از توجهی از داده‌ها تفاوت معنی‌داری وجود دارد که یعنی نمونه‌های انتخاب شده این ویژگی را دارند که تفاوت‌های لازم را نمایش دهند. از طرفی نیز برای مقادیر بزرگتر از 0.05 نیز توزیع تقریبا یکنواختی را مشاهده می‌کنیم که این موضوع کیفیت خوب داده‌ها را تایید می‌کند؛ زیرا که بسیاری از ژن‌ها تفاوت بیان قابل توجهی در نمونه‌های مریض و سالم ندارند و در هر دو حالت بیان تقریبا یکسانی دارند.

سپس نمودار چندک‌های نمونه با چندک‌های تئوری که از توزیع t-Student حاصل می‌شود را رسم می‌کنیم. مشاهده می‌شود که این نمودار تقریبا یک خط را توصیف می‌کند که یک حالت بسیار مناسب است که بیان‌گر این موضوع است که مقادیر مورد استفاده در نمونه از توزیع نظری پیش‌بینی شده تقریبا پیروی می‌کند.

در مرحله بعد، نمودار آتشفشانی [۴] را بررسی می‌کنیم که این نمودار بدین صورت است که میزان اهمیت آماره (که در این آزمایش p-value

[4] volcano plot

است که البته برای نمایش بهتر آن از log p-value- استفاده می‌شود) در مقابل میزان تغییر دیتا (که در این قسمت نیز برای نمایش بهتر از lg fold change بهره می‌گیریم) مورد بررسی قرار می‌گیرد. در این نمودار، داده‌هایی که p-value آن‌ها مناسب است (کمتر 0.05) ، آبی شده‌اند و از طرفی ژن‌های آبی‌رنگ در سمت راست نمودار، میزان بیان آن‌ها در نمونه سالم بیشتر از بیمار است و ژن‌های آبی‌رنگ در سمت چپ نمودار، میزان بیان آن‌ها در نمونه بیمار بیشتر از سالم است و هر چه قدر به سمت راست و چپ نمودار حرکت کنیم، این تفاوت بسیار بیشتر و قابل توجه‌تر می‌شود . این نمودار بیان می‌کند که بخشی از ژن‌ها شرط p-value را برقرار نمی‌کنند و برخی از دیتا ها که این شرط را برقرار می‌کند، fold change آن‌ها چندان قابل توجه نیست ولی تعدادی از داده‌ها که مقادیر p-value آن‌ها بسیار پایین است و نیز fold change آن‌ها مقادیر مثبت نسبتا زیاد دارند (نقاط بالا و سمت راست نمودار) و یا مقادیر منفی نسبتا زیاد دارند (نقاط بالا و سمت چپ نمودار) را می‌توان به‌طور ویژه بررسی نمود و در مجموع نمودار بیان می‌کند که کیفیت داده‌ها در مرحله قابل قبولی قرار دارد.

در قسمت بعدی نمودار اختلاف میانگین [5] را بررسی می‌کنیم. این نمودار تا حدودی مشابه نمودار آتشفشانی عمل می‌کند با این تفاوت که $\log_2$ میانگین بیان ژن‌ها را در آن در نظر می‌گیریم و با lg fold change مقایسه می‌کنیم. در این نمودار مقادیری که p-value آن‌ها مناسب است (کمتر 0.05) است و نیز $\log FC$ آن‌ها مثبت است را قرمز و آن‌هایی که منفی است را آبی کرده‌ایم.

در مرحله بعد، نمودار جعبه‌ای را مورد بررسی قرار می‌دهیم. نمودار حاصل تقریبا نشان می‌دهد که میانه‌ی نمونه‌ها با یکدیگر برابر است که نشان‌گر این است که داده‌ها نرمال و قابل مقایسه هستند و از طرفی طول جعبه‌ها نیز این ویژگی را دارد . چارک‌ها و کمترین و بیشترین مقدار در هر یک از نمونه‌ها نیز تا حدودی با یکدیگر برابر است و از طرفی نیز بیشترین مقدار نمونه‌ها، عددی حداکثر ۱۵ است که یعنی داده‌ها در مقیاس لگاریتمی می‌باشد و نیازی به تبدیل آن‌ها نیست.

در بخش بعدی، نمودار توزیع نمونه‌ها را بررسی می‌کنیم که در آن توزیع همه نمونه‌ها را با رنگ‌های متفاوتی در یک نمودار رسم می‌کنیم. مشاهده می‌کنیم که خم تمام نمونه‌ها تقریبا بر یکدیگر منطبق است که بیان‌گر این است که نمونه‌ها نرمال و قابل مقایسه با هم هستند.

همچنین نمودار Mean-Variance برای بررسی نسبت واریانس به میانگین میزان بیان ژن رسم شده که می‌تواند نشان دهد داده‌ها چقدر از هم پراکنده هستند.

در انتها نیز از UMAP [6] استفاده می‌کنیم که یک روش کاهش بعد داده است که برای نحوه ارتباط داده‌ها با یکدیگر استفاده می‌شود. در نمودار مشخص است که نمونه‌های سالم به چند دسته تقسیم شده‌اند که البته در هر دسته نمونه‌ها بسیار شبیه به یکدیگرند (تقسیم آن‌ها به چند دسته نیز به این علت است که در نمونه‌های سالم نیز چند دسته‌ی متفاوت داشتیم) و از طرفی نمونه‌های بیمار نیز در یک دسته قرار دارند که موکد این است که نمونه‌ها از کیفیت مناسبی برخوردار هستند. البته اگر که توجه نماییم، برخی از نمونه‌های سالم در دسته‌ی بیمار قرار گرفته‌اند که در واقع بیان‌گر این است که این نمونه‌ها شباهت زیادی به نمونه‌های بیمار دارند که این نکته می‌تواند یکی از روش‌های یافتن نمونه‌های شبیه به بیمار در قسمت ۴ام باشد.

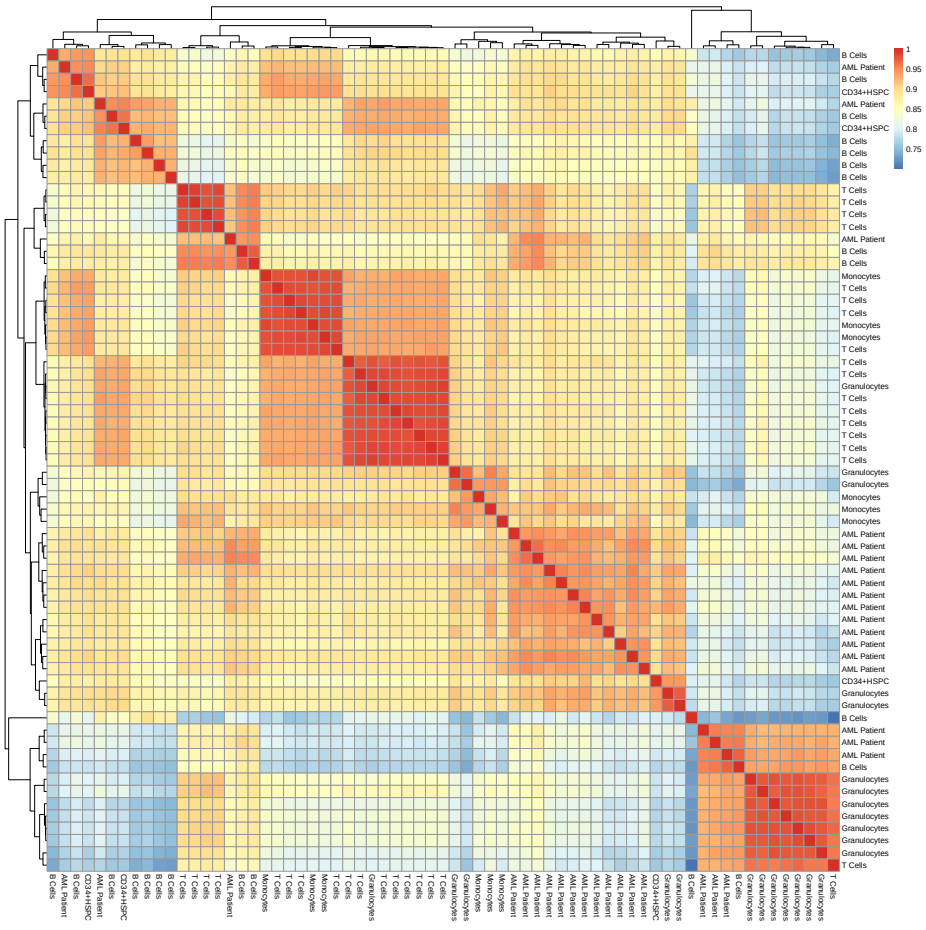
۳ کاهش ابعاد داده‌ها

پس

۴ همبستگی بین گروه‌ها

نمودار هم‌بستگی بین گروه‌های مختلف در شکل ۳ رسم شده است. با توجه به نمودار هم‌بستگی، گروه Granulocytes هم‌بستگی نسبتا پایینی با گروه نمونه‌های بیمار دارد. و از طرفی نمونه‌های T Cells و B Cells هم‌بستگی قابل توجهی با نمونه‌های بیمار دارند.

[5] Mean Difference (MD) plot
[6] Uniform Manifold Approximation and Projection



شکل ۳: میزان هم‌بستگی نمونه‌ها با یک‌دیگر

مراجع

[1] Nature Defenition Microarray (2014), Nature Education, https://www.nature.com/scitable/definition/microarray-202

[2] Microarray Hitmap (2006), Wikipedia, https://commons.wikimedia.org/wiki/File:Heatmap.png#/media/File:Heatmap.png

[3] GEO study, R tidyverse, https://lsru.github.io/tv_course/TD_project_solution.html