بسمه تعالی

دانشگاه صنعتی شریف

دانشکده مهندسی کامپیوتر

# تحلیل داده‌های ریزآرایه لوکمی حاد مغز استخوان

مدرسین: دکتر شریفی زارچی، دکتر کوهی

گردآورنده: ترلان بهادری

ترم پاییز 00ـ99

# فهرست مطالب

# ۱ مقدمه

لوسمی حاد میلوئیدی (AML) ، یکی از چهار نوع اصلی سرطان خون است که بر سلول‌های مغز استخوان اثر می‌گذارد. در این بیماری مغز استخوان میلوبلاست‌ها(نوعی گلبول سفید)، گلبول‌های قرمز و یا پلاکت‌های غیرعادی می‌سازد. عامل اصلی بیماری سرطان خون ناشناخته است، اما پزشکان و دانشمندان معتقدند تلفیقی از عوامل ژنتیکی و عوامل محیطی در ایجاد این لوسمی نقش دارند. این بیماری روندی حاد دارد و معمولاً در صورتی که درمان نشود در عرض چند هفته تا چند ماه منجر به مرگ بیمار می‌شود. یکی از روش‌های تشخیص (AML) استفاده از آزمایش‌های ژنتیکی است، زیرا برخی جهش‌های ژنتیکی می‌توانند باعث ایجاد این بیماری شوند. هدف این پروژه شناخت ژن‌های موثر در به وجود آمدن این بیماری است و نتایج با بررسی داده‌های ریزآرایه به دست آمده از تعدادی نمونه سالم و تعدادی نمونه مبتلا به (AML) به دست آمده است.

# ۲ داده‌ها و روش‌ها

در این پروژه از داده‌های GSE48558 استفاده شده و داده‌ها با استفاده از زبان Python مورد بررسی قرار گرفته‌اند. از کتابخانه‌های زیر استفاده شده است:
GEOparse, pandas, MatPlotLib, numpy, scikit-learn, seaborn
برای تحلیل داده‌ها ابتدا لازم است دسته بندی شوند. بدین منظور، از میان ۱۷۰ نمونه موجود در سری داده GSE48558 نمونه‌هایی که Phenotype آنها Normal است (۴۹ داده) به عنوان گروه کنترل، و نمونه‌هایی که Source Name آنها AML Patient است (۱۸ نمونه) به عنوان گروه تست در نظر گرفته شدند. ابتدا فایل سری داده با فرمت SOFT از این لینک دریافت شد و سپس با استفاده از کتابخانه GEOparse و دستور زیر خوانده شد:

```
gse = GEOparse.get_GEO(filepath="GSE48558_family.soft.gz", destdir="GSE48558")
```

و با استفاده از دستورات زیر دسته بندی شد:

```
def read_expressions(gsm, exprs):
  if len(gsm.table) > 0:
    tmp = gsm.table['VALUE']
    tmp.index = gsm.table['ID_REF']
    gsmNames.append(name)
    if len(exprs) == 0:
      exprs = tmp.to_frame()
    else:
    exprs = pd.concat([exprs, tmp.to_frame()], axis=1)
  return exprs

for name, gsm in gse.gsms.items():
  name = name.strip()
  sample = gse.gsms[name]
  if str(sample.metadata['source_name_ch1']) ==
    str(['AML Patient']):
    test_samples.append(sample)
    exprs = read_expressions(gsm, exprs)

  if str(sample.metadata['characteristics_ch1']) ==
    str(['phenotype: Normal']):
    control_samples.append(sample)
    exprs = read_expressions(gsm, exprs)

```

در بخش ۳، نحوه کنترل کیفیت داده‌ها با استفاده از نمودار جعبه‌ای و استانداردسازی، نحوه کاهش حجم داده‌ها با استفاده از روش PCA و همچنین بررسی همبستگی بین نمونه‌ها با محاسبه correlation بین هر جفت نمونه و رسم نمودار همبستگی و خوشه بندی توضیح داده شده است. همچنین توضیح نحوه محاسبه p-value و پیدا کردن ژن‌هایی که

بیان متمایز در نمونه‌های مبتلا به بیماری و نمونه‌های سالم داشتند آورده شده است. در بخش ۴ ژن‌های به‌دست آمده در بخش ۳، با استفاده از Enrichr تحلیل شده‌اند و ontology و pathway های مرتبط با آن‌ها به‌دست آمده است. در بخش ۵ نیز با توجه به pathway های به‌دست آمده در بخش قبل، یک روش درمان جدید برای AML به اختصار توضیح داده شده است.

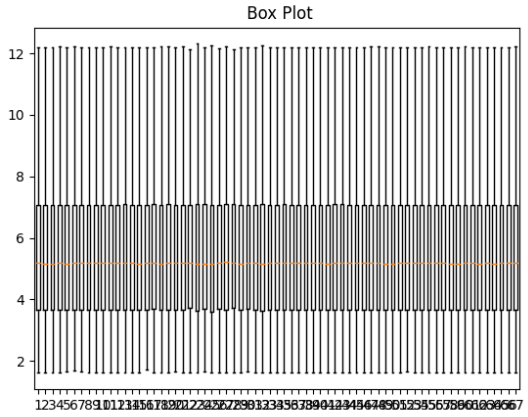
# ۳ تحلیل داده‌ها

## ۱.۳ کنترل کیفیت داده

در بخش Metadata داده‌ها ذکر شده است که این داده‌ها نرمال‌سازی شده‌اند. برای اطمینان می‌توانیم نمودار جعبه‌ای بیان ژن‌ها برای نمونه‌های مورد بررسی را رسم کنیم:

```
exprs.columns = gsmNames

# Plot boxplot of expression data
with PdfPages('GSE_boxplot.pdf') as pdf:
  plt.boxplot(exprs, showfliers=False)
  plt.title('Box Plot')
  pdf.savefig()
  # plt.savefig('boxplot.png')
  plt.close()

```

نمودار جعبه‌ای به صورت زیر رسم می‌شود که فایل PDF آن نیز به این گزارش پیوست شده است.

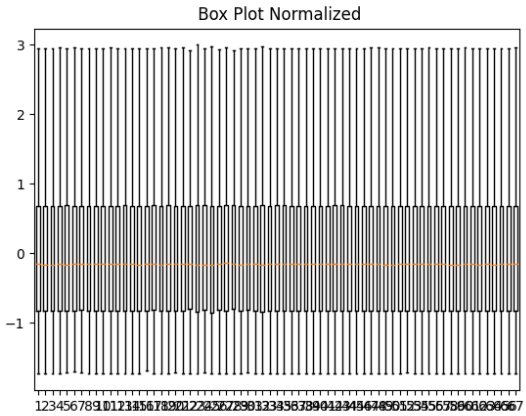



شکل ۱: نمودار جعبه‌ای برای کنترل کیفیت داده

طبق نمودار مشخص است که میانه و چارک اول و سوم برای همه نمونه‌ها تقریباً برابر است، پس نیازی به نرمال‌سازی نیست و همچنین با توجه به بازه‌ای که داده‌ها در آن قرار دارند مشخص است که به صورت لگاریتمی هستند. کیفیت داده‌ها برای تحلیل مناسب است اما در قسمت بعد برای کاهش ابعاد لازم است یک استانداردسازی روی داده‌ها انجام شود به طوریکه اختلاف میزان بیان هر ژن از میانگین آن در کل داده‌ها را برای هر نمونه به دست آوریم. این کار باعث می‌شود اثر ژن‌هایی که در تمام نمونه‌ها به میزان تقریباً یکسانی بیان شده‌اند خنثی شود.برای این کار از دستور زیر استفاده می‌شود:

```
exprs_norm = (exprs - exprs.mean()) / exprs.std()

```

و نمودار جعبه‌ای بر اساس نمونه‌ها به شکل زیر در می‌آید:




شکل ۲: نمودار جعبه‌ای پس از استانداردسازی

البته در اینجا هدف استانداردسازی بر اساس بیان هر ژن در همه نمونه‌ها بوده و نه بر اساس بیان همه ژن‌های هر نمونه (چون از قبل هم بیان ژن‌های هر نمونه نرمال‌سازی شده بود).

## ۲.۳ کاهش ابعاد داده

پس از حذف تاثیر ژن‌هایی که تقریباً در همه نمونه‌ها به یک میزان بیان می‌شوند، گام بعدی کاهش ابعاد داده است که در اینجا با استفاده از روش Principal Component Analysis و دستورات زیر انجام شده است:

```
trans = exprs_norm.transpose()

n_components = 50
pca = PCA(n_components)
fitted = pca.fit(trans)
explained_variance = pca.explained_variance_ratio_
pca_exprs = pca.transform(trans)
X_selected_df = pd.DataFrame(pca_exprs, columns=[pca_exprs.columns[i] for i in range(len(
    pca_exprs.columns)) if pca.get_support()[i]])
print("VARS", explained_variance)
print(pca.components_)
print(pca_exprs)

print(sum(explained_variance[0:40]))

```

با استفاده از explained_variance می‌توانیم سهم هر ژن در کل واریانس داده‌ها را به دست آوریم. پس از اجرای این دستورات می‌بینیم که سهم ۴۰ ژنی که بیشترین واریانس را دارند برابر 0.951 کل واریانس است و این نشان می‌دهد که بخش زیادی از ژن‌ها داده‌های بی‌تاثیری هستند. برای انتخاب ۴۰ ژن که بیشترین واریانس را دارند از دستور زیر استفاده شده است:

```
selector = VarianceThreshold() #default threshold = 0
selector.fit_transform(trans)
vars = selector.variances_
```

```
  vars = list(vars)
  vars.sort()
  vars.reverse()
  print(vars)

```

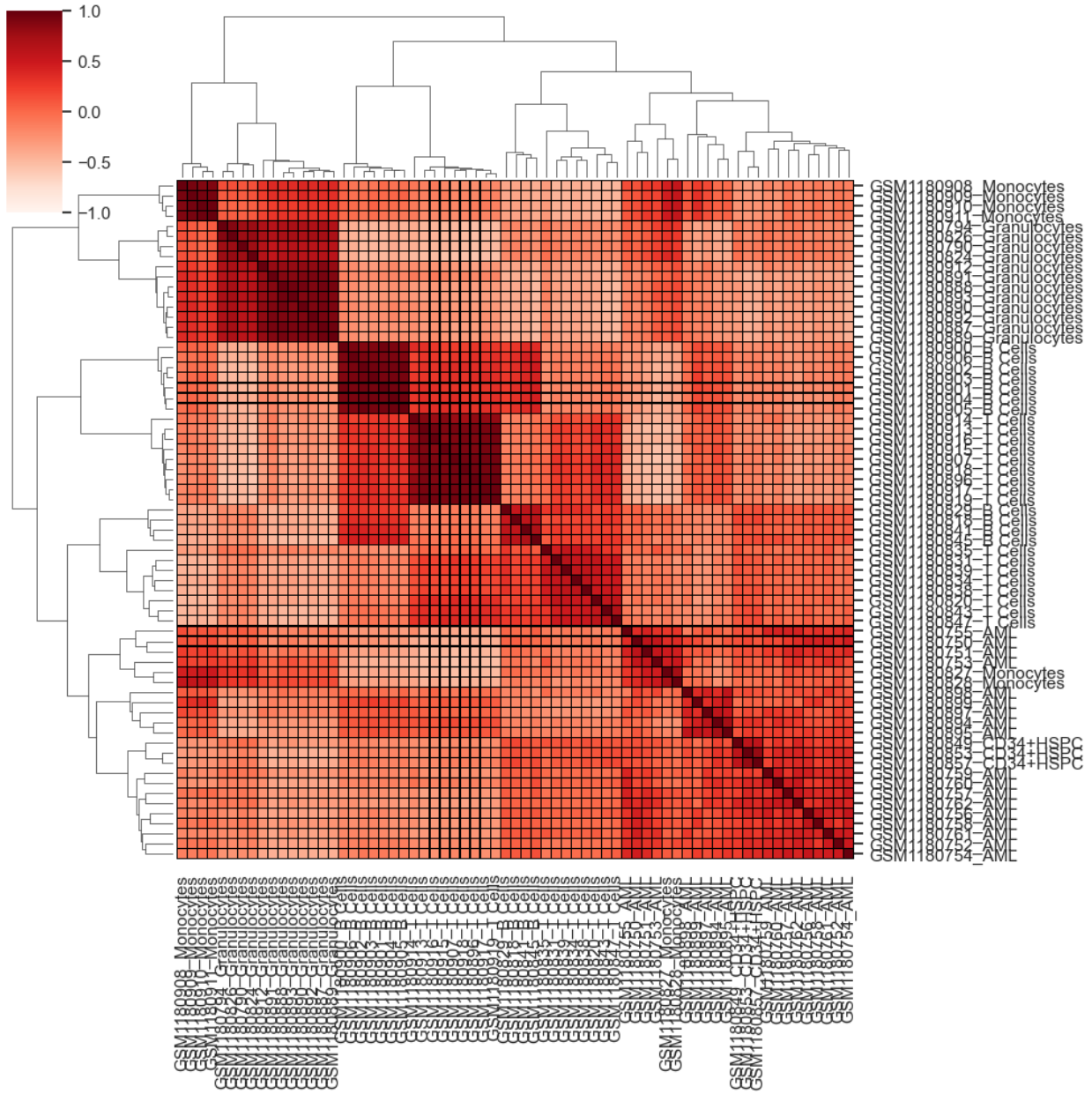
## ۳.۳ بررسی همبستگی بین نمونه‌ها

برای بررسی همبستگی بین نمونه‌ها، ابتدا correlation بین آنها با استفاده از دستور زیر محاسبه شد:

```
heat_exprs = pd.DataFrame()
pca_trans = pca_exprs.transpose()

for name, gsm in gse.gsms.items():
  name = name.strip()
  sample = gse.gsms[name]
  if sample in control_samples:
    col_id = name + '_' + str(sample.metadata['source_name_ch1'])[2:-2]
    heat_exprs[col_id] = pca_trans[name]
  elif sample in test_samples:
    col_id = name + '_' + str('AML')
    heat_exprs[col_id] = pca_trans[name]

corr = heat_exprs.corr()

```

و سپس Cluster Map آنها با استفاده از پکیج seaborn با این دستور رسم شد:

```
ax = sns.clustermap(data=corr, xticklabels=corr.columns, yticklabels=corr.columns,
vmax=1, vmin = -1,
cmap = 'Reds', linewidths=0.1, linecolor='Black')

```




شکل ۳: نمودار بررسی همبستگی و خوشه‌بندی نمونه‌ها

طبق این نمودار دیده می‌شود که همبستگی بین نمونه‌هایی که از یک نوع سلول به دست آمده‌اند با هم بیشتر است. همچنین بیشترین همبستگی بین نمونه‌هایی است که از منشا سلول‌های B Cell, T Cell, Granulocyte به دست آمده‌اند. این سلول‌ها انواعی از گلبول‌های سفید خون هستند که در سیستم ایمنی و مقابله با ویروس‌ها، باکتری‌ها و انواع پاتوژن‌ها نقش دارند و همچنین مشخص شده است که در بروز و درمان انواع سرطان‌های خونی مانند لوکمی و لنفوم می‌توانند موثر باشند. در مجموع نمونه‌های مبتلا به AML همبستگی کمتری با هم دارند و علت آن این است که در هرکدام از آن‌ها میزان بیان ژن‌های متفاوتی با نمونه‌های سالم تفاوت دارد. نمونه‌های مبتلا به AML بیشترین همبستگی را به ترتیب با نمونه‌های سالم به دست آمده از CD34 و Monocytes دارند. CD34 نوعی پروتئین است که با ژنی به همین نام کد شده

است و وظیفه آن کامل شناخته نشده است، اما در ورود سلول‌های T Cell به غدد لنفی نقش دارد. Monocytes هم نوعی گلبول سفید خونی هستند که و یکی از انواع سلول‌های میلوئیدی (مغز استخوانی) محسوب می‌شوند.

## ۴.۳ بررسی تمایز در بیان ژن‌ها

برای پیدا کردن تاثیرگذارترین ژن (ژنی که ضریب آن بیشتر است) در هر PC از این کد استفاده شد:

```
most_important = [np.abs(pca.components_[i]).argmax() for i in range(n_pcs)]

initial_feature_names = gene_idx

# get the names
most_important_names = [initial_feature_names[most_important[i]] for i in range(n_pcs)]
most_important_gene_names = [genes_dict[int(i)].Gene_symbol for i in most_important_names]
print(most_important_gene_names)
# using LIST COMPREHENSION HERE AGAIN
dic = {'PC{}'.format(i + 1): most_important_names[i] for i in range(n_pcs)}
```

با توجه به اینکه نمونه‌های مبتلا به AML بیشترین همبستگی را با نمونه‌های سالمی که منشا آن‌ها CD34+HSPC است داشتند، میزان p-value برای این نمونه‌ها محاسبه شد. برای بررسی تمایز در بیان ژن‌ها، مقدار p-value آنها با استفاده از دستور زیر محاسبه شد:

```
pv_df = pd.DataFrame()
group = list()
i = 0
for name, gsm in gse.gsms.items():
name = name.strip()
sample = gse.gsms[name]
if sample in control_samples and str(sample.metadata['source_name_ch1'])[2:-2] == 'CD34+HSPC':
col_id = name + '_' + str(sample.metadata['source_name_ch1'])[2:-2]
pv_df[col_id] = {'exprs': pca_trans[name]}
group.append('Normal')
elif sample in test_samples:
col_id = name + '_' + str('AML')
pv_df[col_id] = {'exprs': pca_trans[name]}
group.append('AML')
pv_df['group'] = group

def calculate_pvalues(df):
  df = df.dropna()._get_numeric_data()
  dfcols = pd.DataFrame(columns=df.columns)
  pvalues = dfcols.transpose().join(dfcols, how='outer')
  for r in df.columns:
    for c in df.columns:
      pvalues[r][c] = round(pearsonr(df[r], df[c])[1], 4)
  return pvalues

pv_df = (calculate_pvalues(pv_df))

```

سرانجام تاثیرگذارترین ژن‌هایی که میزان p-value آن‌ها از 0.05 کمتر بود به دست آوردیم، به این صورت که ژن‌هایی که میزان logFC آن‌ها از ۱ بیشتر بود یعنی در نمونه مبتلا به AML بیشتر بیان شده‌اند و ژن‌هایی که logFC آن‌ها از ۱- کمتر بود یعنی در نمونه‌های مبتلا به AML بیشتر بیان شده‌اند. در واقع logFC برای هر ژن، حاصل تقسیم میزان بیان آن در نمونه‌های مبتلا به AML بر نمونه‌های سالم مورد بررسی (CD34) است.

```
gene_up = list()
gene_down = list()
for gene in genes:
if gene.adj_P_val < 0.05 and gene.logFC < -1:
gene_down.append(gene.Gene_symbol)
elif gene.adj_P_val < 0.05 and gene.logFC > 1:
```

```
gene_up.append(gene.Gene_symbol)
```

در نهایت ژن‌هایی که میزان بیان آنها در نمونه‌های AML افزایش یافته بود در فایل aml_up_gene.txt و ژن‌هایی که میزان بیان آن‌ها کاهش یافته بود در فایل aml_down_gene.txt ذخیره شده‌اند و این فایل‌ها در مرحله بعد برای بررسی pathway ها مورد استفاده قرار گرفته‌اند.

# ۴ بررسی در Enrichr

پس از بررسی ژن‌هایی که بیان آن‌ها در نمونه‌های AML افزایش یافته، تعدادی از pathway هایی که کمترین p-value را دارند (حدود 0.0008 )، مربوط به kinase هستند. مانند: CASK human kinase ARCHS4 coexpression
PAK3 human kinase ARCHS4 coexpression
EPHA6 human kinase ARCHS4 coexpression
SBK1 human kinase ARCHS4 coexpression
در واقع kinase یک آنزیم است که کاتالیزگر فرآیند انتقال فسفات مولکول‌های فسفردار است. این پروتئین نقش کلیدی در آغاز ترجمه RNA دارد که میزان آن در بیماران AML بیشتر از حد نرمال است. Ontology مرتبط با این دسته از ژن‌ها، negative regulation of developmental process (GO:0051093) با cyclic nucleotide-dependent p-value = 0.0001 protein kinase activity (GO:0004690) است با p-vaue=0.0003 . از بررسی ژن‌های با میزان بیان کمتر نیز اطلاعات زیر به‌دست آمد: در این مورد نیز تعدادی از pathway هایی که کمترین p-value را داشتند، مربوط به Kinase بودند، مانند: MAPKAPK2 human kinase ARCHS4 coexpression (p-value = 1.084e-8)
CAMK1 human kinase ARCHS4 coexpression (p-value = 1.389e-6)
و سایر pathway ها: Nuclear Lamina Cleavage (p-value = 5.086e-5)
Neutrophil Degranulation via FPR1 Signaling (p-value = 7.763e-5)
که nuclear lamina یک شبکه فیبری در هسته بیشتر سلول‌هاست و در فرآیند رونویسی DNA نقش دارد. Neutrophil ها فراوانترین نوع گلبول سفید خونی هستند که در فرآیند degranulation مولکول‌هایی شامل پروتئین‌های سمی برای سلول‌ها، برای مقابله با باکتری‌ها از خود آزاد می‌کنند. در بیماران مبتلا به AML میزان بیان ژن مربوط به این کار کمتر از حالت عادی است و به طور کلی ژن تنظیم کننده آنزیم kinase به میزان غیرمتعادل (زیاد یا کم) بیان می‌شود.

# ۵ تاثیر Kinase در درمان AML

همانطور که در قسمت قبل دیدیم، در بیماران مبتلا به AML میزان آنزیم kinase از حد نرمال خارج می‌شود. این آنزیم در انتقال سیگنال‌ها در بدن طی فرآیند phosphorylation نقش دارد. یکی از روش درمان AML که اکنون به صورت آزمایشی به کار می‌رود، استفاده از داروهای بازدارنده lrkinase است که جلوی فعالیت این آنزیم را می‌گیرند و می‌توانند در مورد بیمارانی که میزان kinase در بدنشان بیش از حد نرمال است به کار روند. این روش در دهه اخیر میزان بهبود مبتلایان به این بیماری را به اندازه قابل توجهی افزایش داده است. در شیمی‌درمانی معمول برای درمان AML مشکلی که وجود دارد این است که علاوه بر سلول‌های هدف، سایر سلول‌های سیستم ایمنی نیز آسیب می‌بینند، اما در روش استفاده از بازدارنده‌های kinase مناطق هدف مشخص است و سایر سلول‌ها آسیبی نمی‌بینند. مهمترین انواع پروتئین‌های kinase که در pathway مربوط به AML نقش دارند، عبارت‌اند از: PIK3/AKT, MAPK/ERK, STAT5 بازدارنده‌هایی که بهرتین عملکرد را در درمان این بیماری داشته‌اند، مربوط به سایت‌های پروتئینی آمینواسید های serine, threonine, tyrosine هستند. پیش‌بینی می‌شود در آینده این نوع روش درمان با هدف‌گیری دقیق تومورها گسترش یابد و نرخ بهبود از بیماری AML بیشتر شود.

# ۶ منابع


1. Acute Myeloid Leukemia
2. AML statistics
3. Dimension Reduction Techniques with Python
4. B-Cells and T-Cells
5. Enrichr
6. Nuclear Lamina
7. Neutrophil Degranulation
8. Protein Kinase Inhibitors as Therapeutic Drugs in AML: Advances and Challenges