به نام خدا

فاز اول پروژه بیوانفورماتیک

اعضای گروه:

محمدرضا دولتی 97110411

محمد حیدری 97110071

فرید فتوحی 98110073

سوال اول

در دهه گذشته، با توانایی به مطالعه در اطلاعات ژنتیکی ژنوم در دامنه گسترده، ریزآرایه ها روشی با توان عملیاتی بالا به منظور تجزیه و تحلیل های بیان ژن به حساب آمدند. ظهور فن آوری نوین ریز آرایه را می توان در سال ۱۹۹۵ دانست. با استفاده از این تکنیک میتوان به طور همزمان بیان هزاران ژن را در حداقل زمان ممکن انجام داد و در سالهای اخیر موجب تولید حجم انبوهی از دادههای بیان ژنی شده است. این روش هر توالی ژنی شناخته شده به عنوان یک کاوشگر روی یک آرایه شیشه ای یا نایلونی تبت می شود. mRNA استخراج شده از بافت یا نمونه خون بارنگ های فلورسنت علامت گذاری می شود و کاوشگرها با RNA مکمل بر روی یک آرایه هیبرید می شوند.

انواع microarray

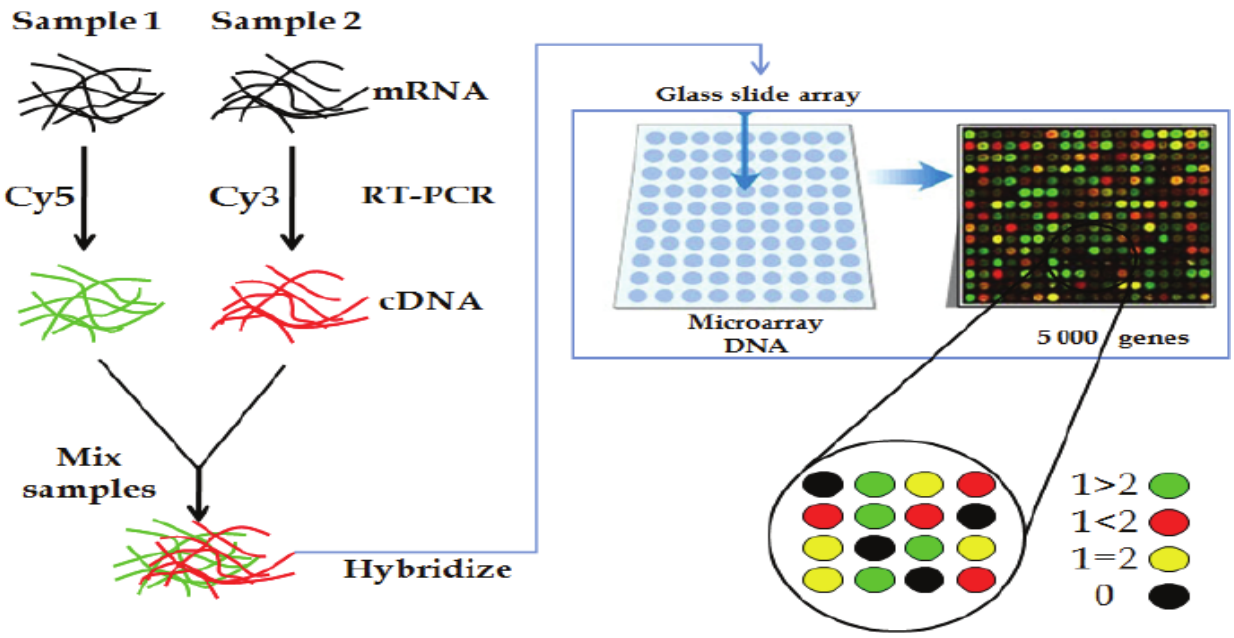
دو نوع آرایه با بیشترین کاربرد عبارتند از:

- آرایه بر پایه DNA مکمل (Complementary DNASpotted)
- آرایه بر پایه الیگونوکلئوتید (Oligo nucleotide array)

آنالیز ریز آرایه محدودیت های خاصی دارد که از آن جمله میتوان عدم توانایی در شناسایی رونوشت های جدید، دامنه ی محدود، تکرارپذیری دشوار و عدم انجام مقایسه بین آزمایش های مختلف به علت خطاهای تصادفی و معمول محققان و آزمایشگاه ها را نام برد (۱۹). برای آنالیز داده های ریز آرایه می توان از شبکه هایی موسوم به شبکه بیانی ژن GCNS استفاده کرد. این شبکه برای انواع وسیعی از مسائل بیولوژی مانند نقش اترات متقابل بین پروتئینها، کشف جایگاه اتصال فاکتور رونویسی و مدل سازی اثرات متقابل ژنتیک به کار برده می شود. چندین ابزار برای تصویرسازی و آنالیز شبکه های بیولوژی از جمله cytoscape ،VisAnt و tYNA استفاده می شود.

مراحل انجام روش microarray چیست؟

در شکل زیر مراحل تکنیک ریز آرایه برای شناسایی و بیان ژن در سلولهای سالم و بیمار ارائه شده است.



ابتدا DNA مورد نظر) که می تواند همان cDNAهای تکثیر شده از mRNA استخراج شده باشد(توسط مواد فلورسنتی نشاندار می شوند و سپس بر روی لام مخصوص ریزآرایه دو رک گیری می شود. پس از نشاندار شدن مولکولهای CDNA عملیات شستشو انجام گرفته تا اتصالات غیر اختصاصی جدا شود در مرحله بعد تشعشع فلورسنتی مربوط به اتصالات اختصاصی توسط دستگاه ثبت می گردد که میزان روشنایی که بیانگر میزان بیان است توسط نرم افزار مربوطه محاسبه و آنالیز آماری انجام میشود.

برای اینکه بفهمیم کدام ژن در مجموعه ای از شرایط بیان می شود، سلول ها در آن شرایط همراه با سلول ها در شرایط کنترل گرفته می شوند. mRNA از این سلول ها استخراج می شود. با استفاده از روش پلی T نشاندار شده به عنوان پرایمر و سنتز cDNA با استفاده از cDNA ، PCR نشاندار شده از هر دو نمونه (نمونه و شاهد) با هم مخلوط شده و در هنگام هیبریداسیون روی لام ریز آرایه شسته می شود. خروجی آزمایش به صورت بیان نوع رنگ است. هر نقطه روی اسلاید یکی از چهار رنگ است: سبز، قرمز، زرد یا سیاه. رنگ ها با بیان ژن در شرایط مختلف مطابقت دارند. لکه هایی که فقط سبز هستند در شاهد به شدت بیان می شوند در حالی که لکه هایی که قرمز هستند در نمونه به شدت بیان می شوند. لکه های زرد در هر دو نمونه و شاهد به طور مساوی بیان می شوند و لکه های سیاه ژن هایی هستند که در هیچ یک از نمونه ها بیان نمی شوند.

سوال دوم

در اول باید داده ها را به کمک کتابخانه GEOquery بخوانیم. با توجه به لینک داده شده در مستند تعریف پروژه، متوجه می شویم که نام دیتاست مورد استفاده GSE48558 و بر پایه پلتفرم GPL6244 است. این دو عبارت را به صورت یک متغیر ذخیره می کنیم تا اگر بعدها خواستیم تحلیل را روی دیتاست مشابهی انجام دهیم، این قسمت کد بدون تغییر قابل استفاده باشد.

Platforms (1)	GPL6244 [HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array [transcript (gene) version]

Series GSE48558	Query DataSets for GSE48558
Status	Public on Jul 06, 2013
Title	Expression data from normal and Malignant hematopoietic cells
Organism	Homo sapiens
Experiment type	Expression profiling by array

```
seriesName <- "GSE48558"
platformName <- "GPL6244"

gset <- getGEO(seriesName,GSEMatrix = TRUE, AnnotGPL = TRUE , destdir = "D:/Term/term7/BIO/project")

if (length(gset) > 1){
  idx <- grep(platformName, attr(gset, "names"))
} else {
  idx <- 1
}
gset <- gset[[idx]]
```

حال نوبت به انتخاب و اسم گذاری روی گروه ها می رسد. برای این کار مواردی که مربوط به بی‌ری بودند، همگی Test نامگذاری شده اند. برای مواردی که Normal بودند، نوع Source Name هم به انتها نام گروه آن ها اضافه شده است چون در ادامه کار به آن نیاز داریم. ضمن این که برای CD34 ، کلمه اضافی HSPC هم وجود داشت که از نام گروه خارج شده است.

```
gset<- gset[,which(gset$source_name_ch1 == "AML Patient" | gset$`phenotype:ch1` == "Normal")]


func <- function(x) {
  if (gset$source_name_ch1[x] == "AML Patient") {
    return("Test")
  } else {
    spll <- strsplit2(gset$source_name_ch1[x] , "\+")[1, 1]
    return(paste0("Normal_" , spll))
  }
}


gr <- sapply(1:length(gset$`phenotype:ch1`) , func)
```

حال ماتریس بیان ژن را تشکیل می دهیم و ماکزیمم و مینیمم آن را چک می کنیم:

```
expr <- exprs(gset)
print(paste0(max(expr)))
print(paste0(min(expr)))
```

خروجی بالا به ترتیب:

13.76153622

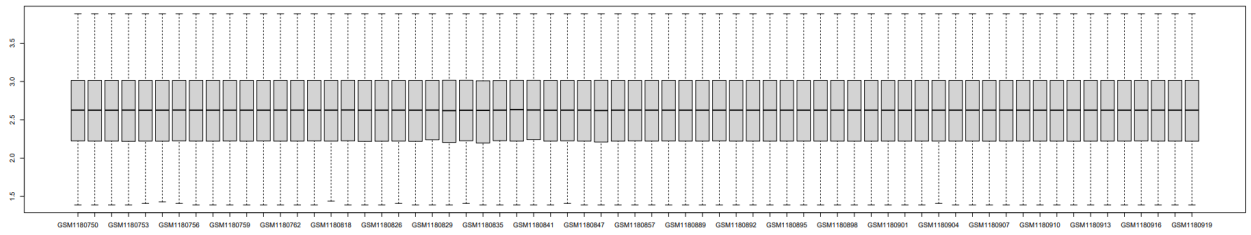
1.611473179

این نشان دهنده این است که بر پایه لگاریتم هستند داده ها

کد تغییر مقیاس لگاریتمی:

```
expr <- log2(1 + expr)
exprs(gset)<- expr
pdf("D:/Term/term7/BIO/project/plot/boxplot.pdf" , width = 32)
boxplot(expr)
dev.off()
```

همین طور داده ها را نرمالیز می کنیم ولی در صورتی که نرمالیز هم نکنیم داده ها نرمال هستند



```
expr <- normalizeQuantiles(expr)
exprs(gset) <- expr
pdf("D:/Term/term7/BIO/project/plot/boxplot_afterNormalize.pdf" , width = 32)
boxplot(expr)
dev.off()
```

برای سوالات 3 و 4 یک پروژه مجزا در Rstudio در ساخته شده است که working directory آن پوشه R-wd است و تمامی دستور ها مبتنی بر آن است.

سوال سوم

در پوشه Src یک فایل به نام microarrayAnalysis.R قرار گرفته است که تمام دستورات در آن است.

هر سه روش PCA و TSNE و MSD را در آن اجرا کرده ایم و خروجی آن را در پوشه Results قرار داده ایم که به صورت تعداد فایل pdf تفکیک شده است.

روشی بیشترین کارایی را دارد که داده ها را بیشتر از هم جدا کند یعنی نقاط مربوط به سلول های سرطانی را طوری از سلول های عادی و سالم جدا کند که به راحتی قابل تشخیص باشند و فاصله ی قابل توجهی داشته باشند. به نحوی که اگر یک نقطه جدید بی رنگ در صفحه گذاشته شد ما بتوانیم با اطمینان خوبی حدس بزنیم این نماینده یک سلول سالم است یا سلول سرطانی. با این تفاسیر، TNSE بیشترین و واضح ترین تفکیک را ایجاد می کند و روش بهتری است.

برای دیدن جزییات و نمودار ها فایل های موجود در پوشه Results را ببینید.

سوال چهارم

در تمام تحلیل های بالا، در کنار دسته بندی بر اساس Aml یا Normal، دسته بندی بر اساس source name هم انجام شده است و تمام نمودار های مرتبط به آن رسم شده است و در فایل های pdf (صفحه اول هر کدام) موجود اند، برای مشاهده به پوشه Results مراجعه کنید. در روش های MDS و PCA نمونه های B Cells نزدیک نمونه های AML هستند و تمییز آن ها دشوار به نظر میرسد، هر چند در TNSE به خوبی از هم جدا شده اند. ما باید تحلیل های بعدی را روی این AML و B Cells انجام دهیم تا بتوانیم عامل تمییز کننده ی اصلی را شناسایی کنیم. در واقع یک پارامتر تعیین کننده وجود دارد که با اندازه گیری و محاسبه ی آن میتوانیم سرطانی بودن یا نبودن یک نمونه را تشخیص دهیم، و هدف ما شناسایی آن پارامتر است. برای اینکه بهتر به آن پارامتر دست پیدا کنیم باید از دو دسته نمونه های را کنار هم قرار دهیم و مقایسه کنیم که شباهت های زیادی دارند در حالی که از دو دسته مخالف اند، در این مقایسه ما باید پارامتر های کمتری را مورد بررسی قرار دهیم تا یقین حاصل کنیم که آیا این پارامتر همان پارامتر تعیین کننده است یا نه؟ زیرا داده های شبیه به هم در بیشتر پارامتر ها مشابه اند و ما میدانیم که آن پارامتر های با مقادیر مشابه اصولا نمی توانند تفاوتی بین دسته آن ها ایجاد کنند و عامل تفاوت دسته های آن ها شوند.