مفاهیم پایه‌ی پروژه:

استفاده از یادگیری ماشین در تحلیل مسیرهای پیام‌رسانی سلولی

مریم رضائی

استاد مشاور: دکتر فاطمه منصوری

نیم‌سال دوم ۱۴۰۲-۱۴۰۱

پیش زمینه و هدف

- برای استفاده از یادگیری ماشین برای تحلیل مسیرهای پیام‌رسانی سلولی نیاز است مفاهیم پایه‌ای اعم از چیستی پیام‌رسانی سلولی، مسیرهای پیام‌رسانی و چگونگی تحلیل آن‌ها را درک کنیم.
- در ادامه به تعریف این مفاهیم از پایین با بالا می‌پردازیم تا در نهایت روند کلی مورد نظر برای انجام این تحلیل را شرح داده و چگونگی تشخیص روش‌های مناسب تحلیل را دریابیم.

فهرست مفاهیم پایه

(۱)
بیان ژن و اهمیت آن

چیستی ژن‌ها، ارتباط آن‌ها با پروتئین‌ها و اثر آن‌ها در کارکرد سلول‌ها

(۲)
ژن‌های دارای بیان متمایز

تفاوت بیان ژن‌ها میان نمونه‌ی سالم و بیمار برای تعیین ژن‌های تاثیرگرفته

(۳)
مسیرهای پیام‌رسانی سلولی

اثر رشته‌ای از پروتئین‌ها در انتقال اطلاعات از بیرون به درون یا میان بخش‌های سلول

(۴)
تحلیل مسیرهای پیام‌رسانی

مقایسه‌ی لیست ژن‌های دارای بیان متمایز با لیست مسیرهای پیام‌رسانی برای تحلیل

(۱.۴)
تحلیل غیرمبتنی بر توپولوژی

روش‌های تحلیل مسیر که تنها با لیست ژن‌های مسیر کار می‌کنند (نسل ۱ و ۲)

(۲.۴)
تحلیل مبتنی بر توپولوژی

روش‌های تحلیل مسیر که با شبکه‌ی ژن‌های مسیر کار می‌کنند (نسل ۳)

بیان ژن و اهمیت آن

(۱)

- بیان ژن فرآیندی است که در آن با استفاده از اطلاعات درون ژن، یک محصول کاربردی تولید می‌شود؛ این محصولات اکثر پروتئین‌ها هستند.
- تکنولوژی‌های گوناگون برای اندازه‌گیری میزان بیان ژن در سلول موجودند، مانند RNA-seq.
- پروتئین‌های تولیدی از بیان ژن کار سلول را تعیین می‌کنند و مقدار یک پروتئین تولیدی در یک سلول در هر لحظه بیانگر تعادل مسیرهای بیوشیمیایی آن است.
- بنابراین با مقایسه‌ی میزان پروتئین‌های تولیدی (یعنی بیان ژن) با حالت سالم می‌توان درستی کارکرد سلول را تشخیص داد.

- Nature Education. (2014). Gene Expression. Scitable. https://www.nature.com/scitable/topicpage/gene-expression-14121669

ژن‌های دارای بیان متمایز

(۲)

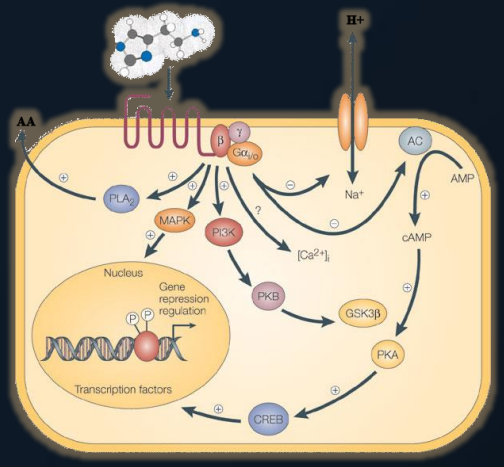
- ژن‌های دارای بیان متمایز با استفاده از روش‌های تحلیل بیان متمایز ژن (DGE) تعیین می‌شوند.
- در این تحلیل، مقادیر بیان ژن میان نمونه‌های مورد نظر (سالم و بیمار) با عادی سازی داده مقایسه می‌شوند.
- برنامه‌های دارای روش‌های متفاوتی برای انجام تحلیل DGE موجودند، مانند TMM، DESeq و edgeR.
- از میان دو برنامه‌ی با بیشترین استفاده (DESeq و edgeR) DESeq دارای میزان کشف اشتباه محدودی است.
- لیست ژن‌های دارای بیان متمایز اطلاعات کاربردی درمورد وضعیت نمی‌دهند؛ چالش، استخراج معنی از این لیست است.

- Westwood, J. (2018c). Using Transcriptomics to Study Behavior. Molecular-Genetic and Statistical Techniques for Behavioral and Neural Research, 267–288. https://doi.org/10.1016/b978-0-12-804078-2.00012-x

مسیرهای پیام‌رسانی سلولی

(۳)

- در فرآیند پیام‌رسانی سلولی اطلاعاتی خاص از سطح سلول به مایع درون سلولی و در نهایت هسته‌ی آن منتقل می‌شود.
- مسیرهای پیام‌رسانی سلولی مسیرهایی از پروتئین‌ها با انواع کارایی‌ها هستند که یکدیگر را فعال کرده و عمل انتقال اطلاعات را انجام می‌دهند.



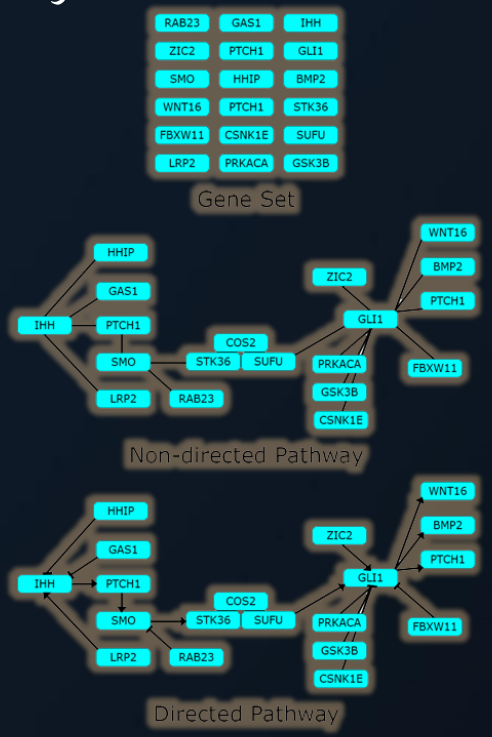
- مسیر شیء گسسته‌ای‌ست که کارکرد آن از برهم کنش اجزا شکل می گیرد.
- برای مثال اجزا پروتئین‌ها و روابط فعال‌سازی یا بازداری هستند.

- Damodaran, T. (2009). Molecular and Transcriptional Responses to Sarin Exposure. Elsevier EBooks, 665–682. https://doi.org/10.1016/b978-012374484-5.00044-4
- Nature Education. (2014b). Signaling cascades within a cell. Scitable. https://www.nature.com/scitable/content/signaling-cascades-within-a-cell-can-interact-14673527
- Garcia-Campos, M. A., Espinal-Enríquez, J., & Hernández-Lemus, E. (2015). Pathway Analysis: State of the Art. Frontiers in Physiology, 6. https://doi.org/10.3389/fphys.2015.00383

مسیرهای پیام‌رسانی سلولی

(۳)

- هر مسیر به صورت داده‌ای از اجزای مربوط به هم می‌باشد.
- شیوه‌ی ثبت داده‌ی مسیر به ۲ صورت است: مجموعه‌ی ژن‌ها یا توپولوژی مسیر (که علاوه بر خود ژن‌ها، ارتباطات را نیز دارد)؛ صورت دوم خود از ۲ نوع جهت‌دار و بی‌جهت است.
- به طور کلی این ۳ نوع خصوصیات زیر را دارند:
  - ✓ مجموعه ژن‌ها: لیست اجزاء زیستی مربوط
  - ✓ توپولوژی مسیر بی‌جهت: شبکه‌ای از اجزاء زیستی به عنوان رأس و ارتباطات به عنوان یال میان هر دو رأس دارای برهم کنشی.
  - ✓ توپولوژی مسیر جهت‌دار: شبکه‌ای با یال جهت‌دار برای اثر یک رأس بر دیگری.



- Garcia-Campos, M. A., Espinal-Enríquez, J., & Hernández-Lemus, E. (2015). Pathway Analysis: State of the Art. Frontiers in Physiology, 6. https://doi.org/10.3389/fphys.2015.00383

تحلیل مسیرهای پیام‌رسانی

(۴)

- یک روش برای استخراج معنی از لیست ژن‌های دارای بیان متمایز، گروه کردن مجموعه‌ای از ژن‌های مربوط به هم است؛ یعنی تعیین مسیرهای پیام‌رسانی سلولی تاثیر پذیرفته.
- در این صورت از پیچیدگی با کاهش تعداد اشیاء کم می‌شود، و کارهایی که ژن‌ها با هم در آن نقش دارند شناسایی می‌شود.
- برای این کار مسیرهای پیام‌رسانی لیست شده در پایگاه اطلاعاتی را با لیست ژن‌ها دارای بیان متمایز مقایسه می‌کنیم.
- تحلیل می‌تواند بی‌توجه به جایگاه و نقش پروتئین (غیرمبتنی بر توپولوژی) یا با توجه به آن (مبتنی بر توپولوژی) باشد.

- Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Computational Biology, 8(2), e1002375. https://doi.org/10.1371/journal.pcbi.1002375
- Nguyen, T. M., Shafi, A., Nguyen, T., & Draghici, S. (2019). Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biology, 20(1). https://doi.org/10.1186/s13059-019-1790-4
- Garcia-Campos, M. A., Espinal-Enríquez, J., & Hernández-Lemus, E. (2015). Pathway Analysis: State of the Art. Frontiers in Physiology, 6. https://doi.org/10.3389/fphys.2015.00383

تحلیل غیرمبتنی بر توپولوژی

(۱.۴)

- روش‌هایی که تنها بر اساس ژن‌های دارای بیان متمایز و مسیرهای پیام‌رسانی نمایش داده شده به صورت مجموعه ژن‌ها به تحلیل مسیرهای تاثیر پذیرفته می‌پردازند.
- این روش‌ها بر اساس ترتیب زمانی ارائه شدنشان به دو نسل زیر تقسیم می‌شوند:

✓ تحلیل بیش‌نمایندگی: تعیین که در کدام مسیرها ژن‌های با بیان متمایز انتخابی، بیش (یا کمتر) از اندازه نمایانند.

✓ امتیازدهی کلاس کاربردی: ژن‌های با بیان متمایز انتخاب نشده و تمام ژن‌ها با مقدار بیانشان در نظرند.

- Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Computational Biology, 8(2), e1002375. https://doi.org/10.1371/journal.pcbi.1002375
- Nguyen, T. M., Shafi, A., Nguyen, T., & Draghici, S. (2019). Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biology, 20(1). https://doi.org/10.1186/s13059-019-1790-4
- Garcia-Campos, M. A., Espinal-Enríquez, J., & Hernández-Lemus, E. (2015). Pathway Analysis: State of the Art. Frontiers in Physiology, 6. https://doi.org/10.3389/fphys.2015.00383

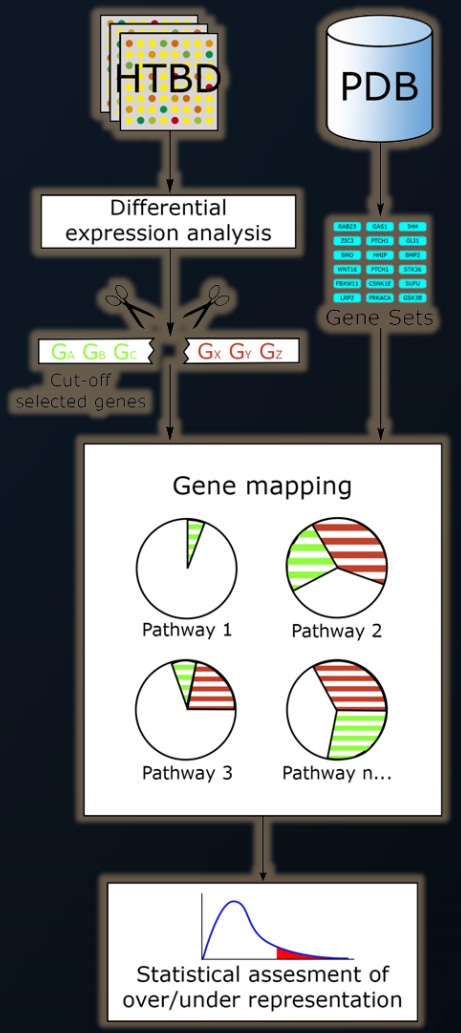
(۱.۴) نسل اول: روش‌های تحلیل بیش‌نمایندگی (ORA)

- مراحل این روش‌ها به شکل زیرند:

✓ از لیست ژن‌ها که میزان تمایز بیانشان محاسبه شده است، ژن‌های دارای تمایز بیان بیش از حدی برای ورودی جدا می‌شوند.

✓ برای هر مسیر، ژن‌های ورودی که عضو مسیر هستند شمرده می‌شوند.

✓ مسیرهایی که ژن‌ها در آن‌ها بیش (یا کمتر) از اندازه نمایانند به ترتیب احتمال تاثیر پذیریشان به دست می‌آیند.



- معایب این روش‌ها شامل: حذف بعضی ژن‌ها از ورودی، دادن وزن یکسان به ژن‌ها (بی‌توجه به سطح بیان یا توپولوژی) و مستقل دانستن مسیرها از یکدیگر است.

(۱.۴) نسل دوم: روش‌های امتیازدهی کلاس کاربردی (FCS)

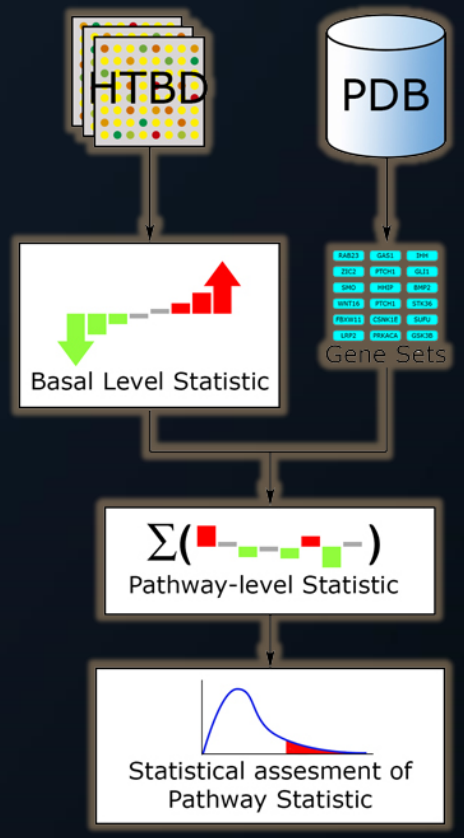
- مراحل این روش‌ها به شکل زیرند:

✓ با این فرض که حتی تمایز بیان کوچک نیز برای ژن‌های مربوط (درون یک مسیر) مهم است، میزان تمایز بیان تمامی ژن‌ها ورودی می‌شود.

✓ آماره‌ی ژن‌های مسیر برای محاسبه‌ی آماره‌ی مسیر با هم ترکیب می‌شوند.

✓ با مقایسه‌ی ژن‌های در هر مسیر (خوددار) یا دسته ژن درون مسیر با بیرون مسیر (رقابتی)، مسیرهای دارای آمار چشمگیر تشخیص داده می‌شوند.

- معایب این روش‌ها شامل: دادن وزن یکسان به ژن‌ها (بی‌توجه به توپولوژی) و مستقل دانستن مسیرها از یکدیگر است.



تحلیل مبتنی بر توپولوژی

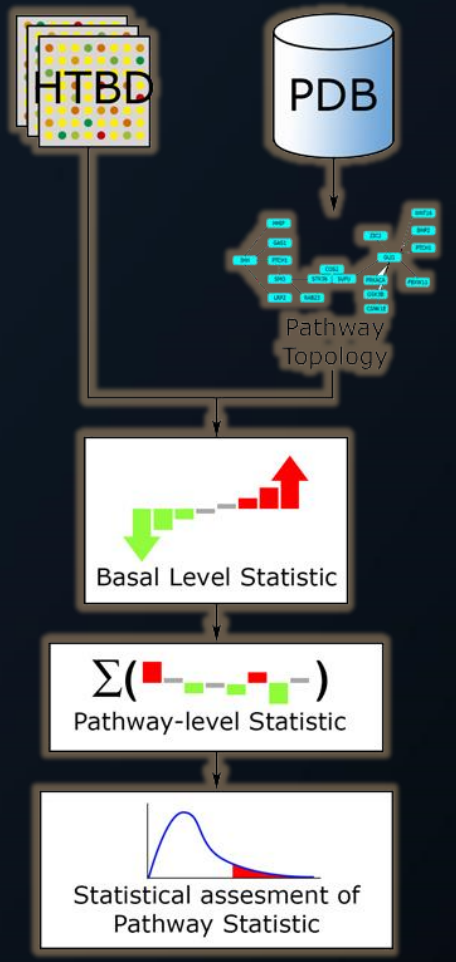
(۲.۴)

- روش‌هایی که از علاوه بر ژن‌ها و میزان تمایز بیان آن‌ها، از نمایش شبکه‌ای مسیرهای پیام‌رسانی (مبتنی بر توپولوژی) در ورودی برای تحلیل استفاده می‌کنند.
- در این صورت ضعف اصلی دو نسل قبل رفع شده و علاوه بر تمایز بیان خود ژن‌ها، وابستگی‌ها، روابط و برهم کنش‌های میان ژن‌ها نیز در نظر گرفته می‌شود.
- این گروه از روش‌ها نسل سوم تحلیل مسیرها را شکل میدهند.

- Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Computational Biology, 8(2), e1002375. https://doi.org/10.1371/journal.pcbi.1002375
- Nguyen, T. M., Shafi, A., Nguyen, T., & Draghici, S. (2019). Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biology, 20(1). https://doi.org/10.1186/s13059-019-1790-4
- Garcia-Campos, M. A., Espinal-Enríquez, J., & Hernández-Lemus, E. (2015). Pathway Analysis: State of the Art. Frontiers in Physiology, 6. https://doi.org/10.3389/fphys.2015.00383

(۲.۴) نسل سوم: روش‌های مبتنی بر توپولوژی مسیر (PTB)

- مراحل این روش‌ها به طور کلی مانند دو نسل قبل بوده و فقط روابط نیز برای تعیین آماره‌ی ژن‌ها و مسیرها اضافه می‌شوند.
- روش‌های این نسل گوناگون هستند؛ برای مثال:
  - ✓ روش ScorePAGE آماره‌ی ژن را از محاسبه‌ی شباهت هر دو ژن میابد.
  - ✓ روش IF آماره‌ی ژن را از جمع تمایز بیان و تابعی خطی بر اساس روابط با دیگر ژن‌های شبکه و آماره‌ی مسیرها از جمع مقدار ژن‌ها میابد.
- معایب این روش‌ها شامل: دشواری یافتن داده‌ی درست تمام انواع توپولوژی سلول‌ها، دشواری در نظر گرفتن پویای مدل مسیرها و مستقل دانستن مسیرها از یکدیگر است.



جمع‌بندی و روش‌کار

با توجه به مفاهیم ذکر شده، برای روند کلی تحقیق مراحل زیر را داریم
که در هر مرحله نیاز است روش مناسب تحلیل انتخاب شود.

۱	تهیه داده‌های عادی شده و یافتن ژن‌های دارای بیان متمایز از داده‌های سالم و بیمار
۲	تعیین مسیرهای پیام‌رسانی سلولی تاثیرگرفته صحیح برای بیماری مورد نظر
۳	آموزش دادن ماشین با تعدادی از نمونه‌ها (ژن‌ها + مسیرها + پاسخ)
۴	استفاده از ماشین مورد آموزش برای پیشبینی دیگر نمونه‌ها

با تشکر از توجه شما

به امید موفقیت این پروژه‌ی عظیم ☺