

ابزاری از MATLAB برای غنی سازی مسیر با استفاده از امتیاز نظم مسیر مبتنی

بر توپولوژی

چکیده

پیشینه: رسیدگی به مقدار وسیع داده‌های تظاهر ژنی تولیدشده توسط روش‌های نمای نسخه‌برداری ژنوم، یک کار چالش‌برانگیز است که نیازمند ترکیبی آگاه از روش‌های پیش‌پردازش، فیلتراسیون و تحلیل دارد اگر که قرار باشد نتایج زیست‌شناختی معناداری حاصل شود. برای مثال، طیفی از آمار سنتی و رویکردهای تحلیل مسیر محاسبه‌ای برای شناسایی فرایندهای بسیار ارائه‌شده در داده‌های ریزآرایه به‌دست‌آمده از حالت‌های مختلف بیماری استفاده‌شده است. اگرچه، اکثر این رویکردها تمایلی در بهره‌برداری از کل طیف داده تظاهر ژنی یا روابط مختلف و وابستگی‌ها ندارند. قبلاً، ما ابزار تحلیل غنی‌سازی مسیر را در MATLAB که یک امتیاز نظم مسیر (PRS) را با در نظر گرفتن علامت‌دهی توپولوژی مسیر و بیش نمایندگی و بزرگی ژن‌هایی که به‌طور متفاوتی ظاهرشده‌اند به دست می‌دهد. در اینجا، این رویکرد را برای شامل شدن مسیر متابولیک گسترش دادیم و استفاده از رابط کاربر گرافیکی (GUI) را توصیف کردیم.

نتایج: با استفاده از تعدادی از جایگاه‌های ریزآرایه و گونه‌ها، مصرف‌کنندگان قادرند تا امتیازات PRS را به همراه امتیاز Z مطابق برای مقایسه محاسبه کنند. ارزیابی بیشتر اهمیت مسیر ممکن است برای افزایش اعتماد به مسیرهای به‌دست‌آمده انجام شود و مصرف‌کنندگان می‌توانند دایره المعارف Kyoto نمودارهای مسیر ژن و ژنوم را که برای تأکید بر ژن‌های نهفته علامت‌گذاری شده است، ببینند.

نتیجه‌گیری: ابزار PRS، فیلتری برای منزوی‌سازی بینش‌های زیستی از داده‌های نسخه‌برداری شده پیچیده ارائه می‌کند.

پیشینه

به‌طور فزاینده‌ای، روش‌های نمای نسخه‌برداری با بازده بالا (ریزآرایه‌ها یا به‌طور فزاینده‌ای، دنباله‌های RNA) تحقیقات علوم حیاتی مدرن را شکل می‌دهد. چنین روش‌هایی، یک دوربین مولکولی فراهم می‌کند که تصاویری از سراسر ژنوم فعالیت ژنتیکی می‌گیرد. اگرچه، داده‌های تحلیل مؤثر ریزآرایه‌ها، چالش‌هایی را مخصوصاً در رسیدگی به تعداد زیادی از ژن‌هایی که به‌طور هم‌زمان مطالعه می‌شوند ارائه می‌دهد.

تحلیل تظاهر ژن درزمینهٔ دانش برگزیده یا "تحلیل مسیر ناشی از پایه دانش" از آنجایی که این مسئله باعث کاهش در فضای تحقیقاتی از هزارها ژن تا زیرمجموعه‌ای از فرایندهای زیست‌شناختی که برای تفسیر انسانی بیشتر مهار شدنی‌تر است ضروری است [1]. طبق Khatrı و همکارانش [2]، رویکردهای غنی‌سازی مسیر می‌تواند به سه نسل تقسیم‌بندی شود:

1- تحلیل بیش نمایندگی (ORA): این تحلیل یک مسیر را با در نظر گرفتن نسبت ژن‌هایی که به‌صورت متفاوت ظاهر شده (DEG) و در هر مسیر مربوط به نسبت تمام ریزآرایه‌های DEG مشاهده می‌شود، ثبت می‌کند. این تحلیل برای ابزارهای تحلیل مسیر متعددی من جمله [3] GenMAPP, [4] GoMiner, [5] Onto-Express and [6] FatiGo به کار می‌رود.

2- امتیازدهی طبقه عملکردی (FCS): FCS یک امتیاز را به هر ژن بر اساس تظاهر آن، در مسیر می‌دهد که با آن امتیاز مسیر بر اساس امتیازات تمام ژن‌ها در مسیر محاسبه می‌شود. تعدادی از روش‌های FCS از طریق ابزارهای مستقل مانند [7] GSEA, [8] SigPathway و [9] SAFE یا ابزارهای وب مانند [10] T-profiler, Gazer [11] و [12] GeneTrail اجرا می‌شود.

3- رویکردهای مبتنی بر توپولوژی مسیر (PT): این رویکردها از توپولوژی مسیرها با دادن اوزان برای اتصالات از پیش تعیین شده بین ژن‌ها که امتیازدهی مسیر را تشکیل می‌دهند، بهره می‌گیرند. برخی رویکردهای مبتنی بر توپولوژی در دهه‌های گذشته در ادبیات توصیف شده است. طبق [13] Mitrea et al، رویکردهای مبتنی بر PT، در روشی که اطلاعات توپولوژی مسیر را به امتیاز مسیر ترجمه می‌کند متفاوت‌اند. برخی روش‌ها تنها از داده‌های توپولوژی ژن‌هایی که به صورت متفاوت ظاهر شده‌اند (DEG) در امتیاز غنی‌سازی استفاده می‌کنند (مثلاً، [14] MetaCore، [15] EnrichNet)، درحالی که دیگر روش‌ها (من جمله، [16] SPIA و [17] GANPA) از داده‌های تظاهر DEG به همراه داده‌های توپولوژی استفاده می‌کنند. متناوباً، برخی روش‌ها از داده‌های تظاهر ناشی از تمام ژن‌های ریزآرایه استفاده می‌کنند چه بین شرایط تغییر بکنند یا نکنند، برای مثال، [18] PathOlogist، [19] DEGraph و [20] ACST. نکته مهم این‌که برخی ابزار مبتنی بر PT تنها از توصیف‌های علامت‌دهی مسیر مانند Pathway- [21] Express، [22] NetGSA، [23] ScorePAGE، [24] TAPPA، [25] MetPA و [26] Clipper استفاده می‌کنند.

پیش‌ازاین، یک روش غنی‌سازی مسیر جدید را ارائه کردیم که در آن هم توپولوژی مسر و هم بزرگی تظاهر ژن، ایجاد امتیاز نظم مسیر (PRS) را تغییر می‌دهد [27]. مخصوصاً، با ترکیب داده‌های تغییر کل برای آن نسخه‌هایی که از آستانه اهمیت تجاوز می‌کنند و با در نظر گرفتن پتانسیل تظاهر ژن تغییر یافته در اثرگذاری بر نسخه‌برداری‌های پایین‌دست، مسیرهایی را شناسایی کردیم که به فرایند پاتوفیزیولوژیک تحت بررسی مرتبط است. رویکرد ما تعدادی از مسائل را که به‌طور بالقوه روش‌های غنی‌سازی را تضعیف می‌کند در نظر می‌گیرد. ما گام‌هایی برای کاهش تأثیر خطاها در نگاشت شناسه و کاهش خطای ایجاد شده توسط مسیرهای اضافی (مانند نمونه‌های چندگانه یک ژن) برداشتیم. روش‌های توپولوژی نیز باید به‌صورت مؤثر به حلقه‌ها رسیدگی کنند، لذا ما از الگوریتم جستجویی که از نظریه گراف ناشی شده است برای حل این مشکل استفاده کردیم. هم‌چنین احساس کردیم که تقسیم‌بندی دلخواهانه فرایندها به نظم بالا یا پایین، از آنجایی که تغییرات در تظاهر ژن احتمالاً در سراسر مسیرها توزیع می‌شود، ساختگی است و از این‌رو ارزیابی ما، یک ارزیابی کلی اثر بود.

در اینجا، اجرای رویکرد PRS خود را به عنوان یک ابزار مستقل که به مصرف کننده نهایی گزینه وارد کردن داده را از جایگاه ها و انواع مختلف ریزآرایه می دهد، توصیف کردیم. این ابزار هم امتیازات Z و هم PRS را به دست می دهد، تحلیل آماری ارائه می کند و اجازه مرور راه هایی که دارای ژن های نهفته ای هستند که بارنگ های مختلف علامت گذاری شده اند می دهد. گزارش خود را با این مطلب که مصرف کنندگان قادرند تا هم مسیرهای متابولیک و هم علامت دهی را تقویت کنند، ارتقا می دهیم.

اجرا

رویکرد PRS در MATLAB اجرا شد. مصرف کنندگانی که به محیط MATLAB دسترسی ندارند، می توانند کامپایلر اجرای MATLAB (MRC) را دانلود کنند تا نرم افزار توصیف شده در اینجا را با یک GUI کاربرپسند به کار گیرند. رابط PRS (شکل 1) توابع متعددی را برای مصرف کنندگان ارائه می دهد:

The screenshot shows the PRS_interface software window. On the left, there are several panels for configuring a new analysis: 'New analysis' with fields for Species (Human), Data file, Num of samples (group1, group2), FoldChange Threshold, Pvalue threshold, Normalization (RMA/GCRMA), Platform (Affy HG U133A), and Pathway type (Both); 'Normalization and Evaluation' with dataset, num of permutations, and Pvalue Threshold; and 'Open Existing analysis' with dataset and an Open button. Below these are 'Others' options: 'select a pathway from the table to visualize' and 'Clear tables'. The main panel on the right is titled 'list of enriched pathways ranked by normalized score PRS' and contains a table with 27 rows of pathway data.

Pathway Id	Pathway definition	PRS	Pvalue	FDR	Pathway Id	Pathway definition	Zscore
1	path.hsa0... Pathways in cancer - Homo sapiens (human)	26.4540	1.0000...	0.0030	1	path.hsa0... Osteoclast differentiation - Homo sapiens (human)	4.6726
2	path.hsa0... Acute myeloid leukemia - Homo sapiens (human)	24.3507	0	0	2	path.hsa0... Antigen processing and presentation - Homo sapiens (human)	4.0928
3	path.hsa0... Natural killer cell mediated cytotoxicity - Homo sapiens (human)	20.0962	0	0	3	path.hsa0... Natural killer cell mediated cytotoxicity - Homo sapiens (human)	4.0617
4	path.hsa0... Cytokine-cytokine receptor interaction - Homo sapiens (human)	19.2648	0	0	4	path.hsa0... Acute myeloid leukemia - Homo sapiens (human)	3.8778
5	path.hsa0... T cell receptor signaling pathway - Homo sapiens (human)	17.4283	1.0000...	0.0030	5	path.hsa0... T cell receptor signaling pathway - Homo sapiens (human)	3.6975
6	path.hsa0... MAPK signaling pathway - Homo sapiens (human)	17.1975	0.0080	0.0186	6	path.hsa0... Malaria - Homo sapiens (human)	3.6916
7	path.hsa0... Osteoclast differentiation - Homo sapiens (human)	16.6754	0	0	7	path.hsa0... Systemic lupus erythematosus - Homo sapiens (human)	3.5432
8	path.hsa0... Endocytosis - Homo sapiens (human)	10.8245	0.0100	0.0193	8	path.hsa0... Staphylococcus aureus infection - Homo sapiens (human)	3.3589
9	path.hsa0... Neurotrophin signaling pathway - Homo sapiens (human)	10.6538	0.0200	0.0270	9	path.hsa0... Endocytosis - Homo sapiens (human)	3.2979
10	path.hsa0... Chemokine signaling pathway - Homo sapiens (human)	10.0916	0.0270	0.0317	10	path.hsa0... Bacterial invasion of epithelial cells - Homo sapiens (human)	2.8939
11	path.hsa0... Malaria - Homo sapiens (human)	9.4949	0	0	11	path.hsa0... Thyroid cancer - Homo sapiens (human)	2.7391
12	path.hsa0... Jak-STAT signaling pathway - Homo sapiens (human)	9.3189	0.0030	0.0081	12	path.hsa0... Endometrial cancer - Homo sapiens (human)	2.7229
13	path.hsa0... Apoptosis - Homo sapiens (human)	9.2222	0.0060	0.0135	13	path.hsa0... Fc gamma R-mediated phagocytosis - Homo sapiens (human)	2.6683
14	path.hsa0... Endometrial cancer - Homo sapiens (human)	8.6814	0.0080	0.0135	14	path.hsa0... Apoptosis - Homo sapiens (human)	2.5476
15	path.hsa0... Thyroid cancer - Homo sapiens (human)	8.5206	1.0000...	0.0030	15	path.hsa0... Cytokine-cytokine receptor interaction - Homo sapiens (human)	2.4911
16	path.hsa0... Fc gamma R-mediated phagocytosis - Homo sapiens (human)	8.4456	0.0250	0.0307	16	path.hsa0... Allograft rejection - Homo sapiens (human)	2.4702
17	path.hsa0... Staphylococcus aureus infection - Homo sapiens (human)	7.4460	1.0000...	0.0030	17	path.hsa0... MAPK signaling pathway - Homo sapiens (human)	2.4690
18	path.hsa0... Amebiasis - Homo sapiens (human)	7.0131	0.0190	0.0270	18	path.hsa0... Chemokine signaling pathway - Homo sapiens (human)	2.4282
19	path.hsa0... Chronic myeloid leukemia - Homo sapiens (human)	6.5650	0.0370	0.0384	19	path.hsa0... Neurotrophin signaling pathway - Homo sapiens (human)	2.3476
20	path.hsa0... Bacterial invasion of epithelial cells - Homo sapiens (human)	6.1642	0.0140	0.0236	20	path.hsa0... Nicotinate and nicotinamide metabolism - Homo sapiens (human)	2.3330
21	path.hsa0... Antigen processing and presentation - Homo sapiens (human)	6.1463	0.0140	0.0236	21	path.hsa0... Lysosome - Homo sapiens (human)	2.2374
22	path.hsa0... Colorectal cancer - Homo sapiens (human)	5.9075	0.0410	0.0410	22	path.hsa0... Pathways in cancer - Homo sapiens (human)	2.1774
23	path.hsa0... Viral myocarditis - Homo sapiens (human)	5.1525	0.0190	0.0270	23	path.hsa0... Chronic myeloid leukemia - Homo sapiens (human)	2.1453
24	path.hsa0... Systemic lupus erythematosus - Homo sapiens (human)	4.9950	0.0320	0.0356	24	path.hsa0... Jak-STAT signaling pathway - Homo sapiens (human)	2.1006
25	path.hsa0... Allograft rejection - Homo sapiens (human)	4.4056	0.0150	0.0238	25	path.hsa0... Epithelial cell signaling in Helicobacter pylori infection - Homo sapiens (human)	1.9068
26	path.hsa0... Nicotinate and nicotinamide metabolism - Homo sapiens (human)	3.6333	0.0330	0.0356	26	path.hsa0... Viral myocarditis - Homo sapiens (human)	1.8549
27	path.hsa0... Porphyrin and chlorophyll metabolism - Homo sapiens (human)	3.5733	0.0220	0.0283	27	path.hsa0... SNARE interactions in vesicular transport - Homo sapiens (human)	1.8608

شکل 1: رابط مصرف کننده PRS که نشان دهنده یک مجموعه داده نمونه است

پیش‌پردازش داده‌های ریزآرایه

ما یک فیلتر را برای نرمال‌سازی داده‌ها از جایگاه‌ها مختلف مجدداً مهندسی نکردیم، در عوض، مصرف‌کنندگان ابتدا باید داده‌های نسخه‌برداری شده را با استفاده از یکی از ابزارهای بی‌شمار موجود پیش‌پردازش کنند. داده‌ها باید در قالب یک صفحه گسترده (Excel Spreadsheet) باشد که در آن اولین ستون باید شناسه محقق باشد و در ستون‌های بعدی باید ارزش‌های تکرار شده نرمالیزه شده تظاهر را از شرایط تست و کنترل قرار داد. اطلاعات مازاد در خصوص گونه‌ها، تعداد نمونه، تغییر کل، آستانه‌های تست t ، روش نرمال‌سازی و جایگاه موردنیاز است.

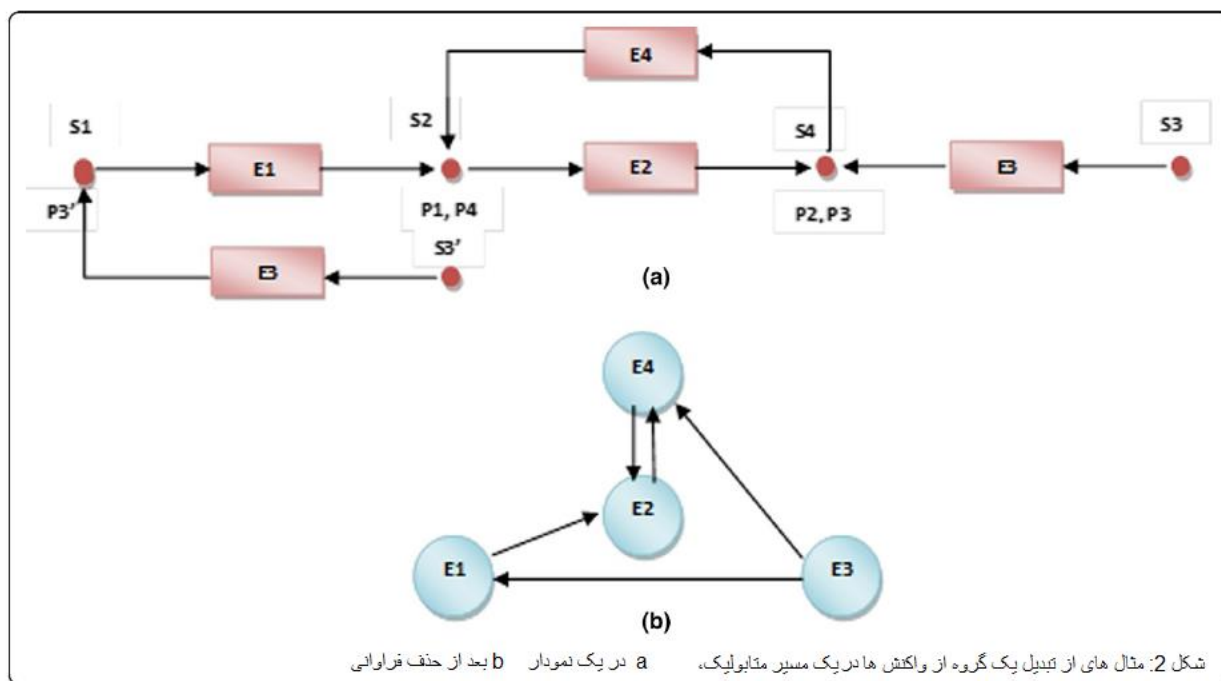
نمایش مسیر

الگوریتم اساسی ما پیش‌ازاین توصیه شد [27]. به‌طور خلاصه، دایره المعارف Kyoto تعاریف مسیر ژن‌ها و ژنوم‌ها [28] استفاده شد که در آن مسیرها در فرمت (KGML) زبان نشانه‌گذاری KEGG حفظ شده‌اند. مجموعاً ما 189 توصیف علامت‌دهی و متابولیک از KEGG وارد کردیم و این‌ها را در موارد MATLAB تجزیه و تحلیل کردیم و سپس به گراف‌های مستقیم تبدیل کردیم. پرونده‌های KGML شامل سه نوع مورد است: ورودی‌ها، روابط و واکنش‌ها. این‌ها می‌توانند به‌صورت موارد گرافیکی در نقشه مسیر مرتبط کشیده شود (پرونده اضافی 1). تنها ورودی‌ها (که متشکل از گره‌ها هستند و به‌صورت جعبه نمایش داده می‌شوند) و روابط (که به‌صورت مرز نمایش داده می‌شوند) در نمایش مسیرهای علامت‌گذاری که پروتئین‌ها (جعبه‌ها) با "روابط" پیوند داده شده‌اند، استفاده می‌شوند. تمام سه نوع در نمایش ساختار مسیرهای متابولیک برای گرفتن روابط لایه-آنزیم-محصول که در آن آنزیم‌ها (جعبه‌ها) با "روابط" پیوند دارند و ترکیبات (دایره‌ها) با "واکنش‌ها" پیوند دارند استفاده می‌شوند. برای تبدیل یک مسیر متابولیک به یک گراف به روشی منطقی، ما آنزیم‌ها را به‌عنوان گره‌ها در گراف نشان دادیم، درحالی‌که لایه‌ها و محصولات برای شناسایی جهت روابط (مرزها) بین گره‌ها استفاده شدند (شکل 2). درحالی‌که اذعان کردیم که ممکن نیست تا هرگونه تأثیر بر روی شار را با این منطق پیش‌بینی کرد، استدلال کردیم که هرگونه تغییر در تظاهر گره در یک مسیر متابولیک می‌تواند فیزیولوژیکی باشد مخصوصاً اگر گره‌ها متصل باشند.

نمایش مسیرها به صورت گراف با توجه به اینکه فراوانی را که ژن‌ها تنها یکبار در هر گراف مسیر نشان دادند کاهش می‌دهد، دارای مزیت است. یک الگوریتم جستجوی عمق اول (DFS) ناشی از نظریه گراف برای تضمین اینکه حلقه‌ها تنها یکبار شمرده می‌شوند به کار رفت.

امتیازدهی مسیر

روش ما اوزان را به تمام گره‌های مهم (همان DEGs) در یک مسیر اختصاص داد تا نیروی توپولوژیکی آن‌ها را نشان دهد (مخصوصاً تعداد گره‌های مهم پایین دست که به آن یا به صورت مستقیم یا از طریق دیگر گره‌های مهم اشاره شد همان‌طور که پیش‌ازاین توصیف شد [27]). یک PRS بر مبنای ارزش تغییر کل و وزن دهی تمام گره‌های مهم در مسیر و محاسبه شد و برای اندازه مسیر نرمال شد. هم‌چنین ما یک امتیاز Z را نیز محاسبه کردیم [29] (با بهبودی در اجراهای اولیه که در آن بعد از حذف ژن‌های فراوان از توصیفات مسیر، این کار انجام شد). نرم‌افزار دو فهرست از مسیرهای رتبه‌بندی شده طبق PRS و امتیاز Z را به دست می‌دهد که هم به برای تحلیل بیشتر هم به صورت فایل excel و هم به صورت mat. ذخیره شدند.

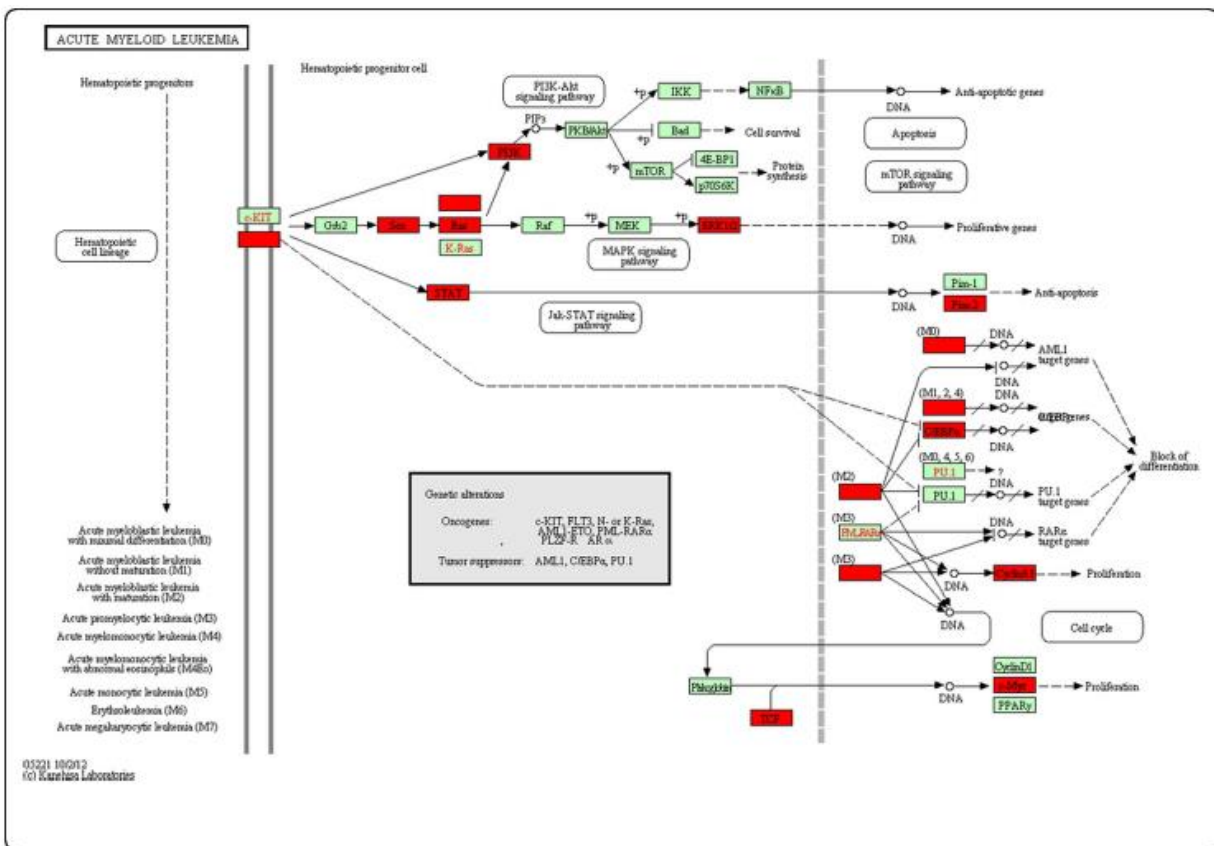


ارزیابی اهمیت مسیر

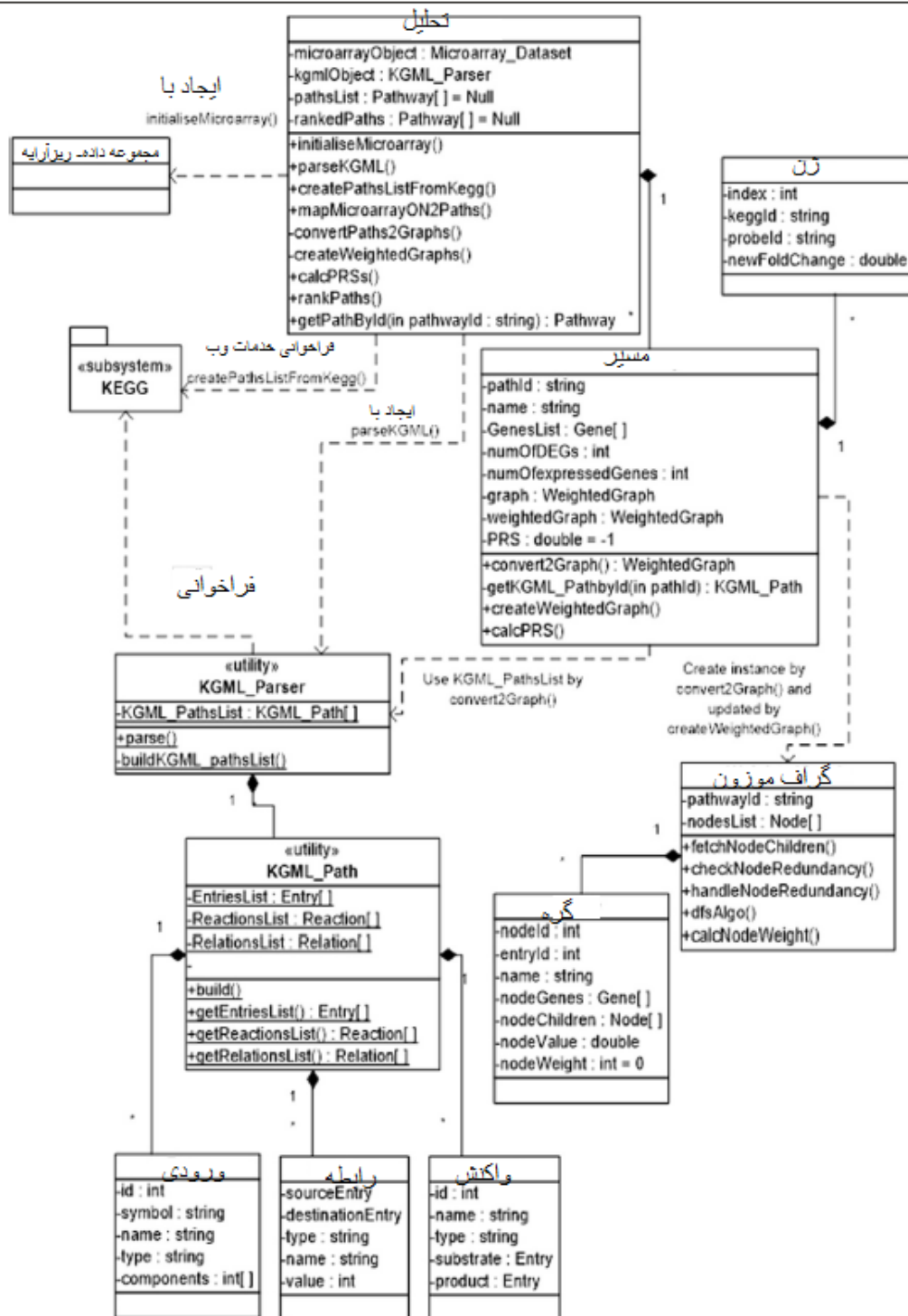
سپس ما احتمال کسب امتیازاتی حداقل به اندازه امتیاز PRS به صورت اتفاقی و با استفاده از روش جایگشت غیر پارامتریک را ایجاد کردیم. در ابتدا، مقادیر تغییر کل برای تمام ژن‌های ریزآرایه تغییر کرد. سپس این مقادیر به سمت مسیرها بازگشت و PRS دوباره محاسبه شد. این فرایند n بار تکرار شد که در آن n توسط مصرف‌کننده از طریق رابط ارائه شد (معمولاً $n=1000$). اهمیت آماری (مقدار p) هر امتیاز مسیر با مقایسه بین امتیاز مشاهده شده و امتیازات ایجاد شده رندم n برآورد شد. برای کسب ارزیابی اهمیت آماری قابل‌اعتمادتر، مقادیر p برای چندین تست اصلاحی توسط روش نرخ کشف خطا (FDR) بر اساس آستانه ارائه شده توسط مصرف‌کننده، تنظیم شدند. این مسئله به صورت دقیق‌تر در گزارش اصلی ما توصیف شده است [27].

تجسم مسیرهای غنی شده

بعد از انجام تحلیل، نتایج به صورت فایل‌هایی با فرمت mat. برای بازیابی آسان ذخیره شدند. با کلیک کردن بر روی نام مسیر از فهرست مسیرهای رتبه‌بندی شده که در جدول 1 نشان داده شده است و انتخاب گزینه تجسم یک مسیر از رابط، یک نقشه مسیر علامت‌گذاری شده نشان داده خواهد شد. درواقع، نرم‌افزار یک خدمات وب نقشه‌کشی مسیر فرا خواهد خواند که بر روی وبسایت KEGG قرار دارد و تعدادی از پارامترها را من جمله فهرستی از ژن‌های تظاهر با تغییرات کلی و رنگ‌های مشخص را عبور می‌دهد تا بین DEG ها با ژن‌های غیر نهفته تفاوت قائل شود. شکل 3 نشان‌دهنده یک نقشه مسیر معمولی است که در آن ژن‌های مهم (بالای آستانه) به رنگ قرمز مشخص شده‌اند و غیر مهم‌ها (تغییر نکرده یا ظاهر نشده) به رنگ سبز.



شکل 3: یک مسیر علامت گذاری شده معمولی، در این مورد، "KEGG" مسیر لوسمی میلوئیدی حاد " غنی شده در یک مجموعه داده AML (ورودی GEO، #GSE9476)؛ ژن های مهم قرمز رنگ و غیر مهم ها سبز هستند.



شکل 4: نمودار طبقه UML که نشان دهنده طبقه های اصلی بسته در مرحله تحلیل مسیر است

UML برای مدل سازی و توصیف نرم افزار

در اینجا، ما از زبان مدل سازی یکپارچه (UML) استفاده کردیم تا ساختار و توابع روش خود را با نمودارها توصیف، مدل سازی و تجسم کنیم. 14 نوع نمودار دسته بندی شده در سه طبقه در UML 2.0 وجود دارد [30]. اگرچه، در این مقاله ما از دوتا استفاده کردیم: نمودارهای طبقه و توالی. نمودارهای طبقه ساختارهای ثابت یا موارد اصلی در نرم افزار را نشان می دهد. شکل 4 نشان دهنده طبقه های اصلی در مرحله تحلیل مسیر است. تحلیل طبقه، طبقه اصلی است که رابطی برای اجرای تمام خدمات ارائه شده توسط ابزار ارائه می کند؛ و دارای چهار ویژگی اصلی است:

- مورد ریزآرایه: موردی از طبقه "مجموعه داده- ریزآرایه" که با فراخوانی اولیه تابع ریزآرایه () ایجاد می شود (فایل اضافی 2 را ببینید)؛ و داده های تظاهر ژنی نرمال شده را و فهرستی از تمام ژن ها با مقادیر تغییر کلی شان را دارد.
- مورد KGML: موردی از طبقه "KGML_Parser" را با فراخوانی تابع ParseKGML (فایل اضافی 3 را ببینید) ایجاد می کند؛ و ساختار پایای تمام مسیرها را به عنوان فهرستی از موارد طبقه "KGML-PATH" که با فرمت KGML تعریف شده است، داراست. یک مورد از "KGML_path" ساختار یک مسیر KEGG را نشان می دهد و متشکل از فهرست مطالب، فهرست واکنش ها و فهرست روابط است (فایل اضافی 1 را ببینید).
- فهرست مسیر: این فهرستی از موارد طبقه مسیر است که با فراخوانی تابع Createpathlistfromkegg () ایجاد می شود (فایل اضافی 4 را ببینید). این مورد در نهایت فهرستی از مسیرهای غنی سازی شده با ارجاع به یک مجموعه داده ریزآرایه را داراست.
- مسیرهای رتبه دهی شده: این مورد با فراخوانی تابع rankpaths () ایجاد می شود؛ و دارای همان فهرست مسیرهای تعریف شده توسط Pathlist می باشد، اما به صورت نزولی بر اساس مقادیر PRS رتبه بندی شده اند.

نمودارهای توالی برای نشان دادن توابع ابزار PRS طبق انواع مختلف روابط درونی بین موارد استفاده شده است. به عنوان مثال، شکل 5 نمایش دهنده توابع PRS اصلی با گام های زیر است:

1. تبدیل مسیرها به گراف توسط تابع convertpath2graph که نیاز به استفاده از مورد KGML دارد و دارای فهرستی از مطالب، روابط و واکنش های تمام مسیرهاست.

2. با استفاده از اطلاعات ذخیره شده در مورد KGML و pathlist برای هر گراف (شکل 4 را ببینید)، فهرستی از گره‌ها (که هر گره نشان‌دهنده یک یا چند ژن از مسیر اصلی است) و فهرستی از کودکان برای هر گره ایجاد می‌شود.

3. حذف ژن‌های فراوان که ممکن است بارها در همان مسیر نشان داده شده باشد. دو تابع برای رسیدگی به فراوانی گره طراحی شده است: `checknoderedundancy()` و `handlenoderedundancy()`.

4. بعد از ایجاد یک گراف برای هر مسیر، گراف‌ها با فراخوانی تابع `createweightedgraphs()` وزن دهی می‌شود که از الگوریتم DFS برای عبور دادن گره‌های هر گراف و اختصاص یک وزن برای هر گره مهم با در نظر گرفتن حلقه‌های گراف استفاده می‌کند.

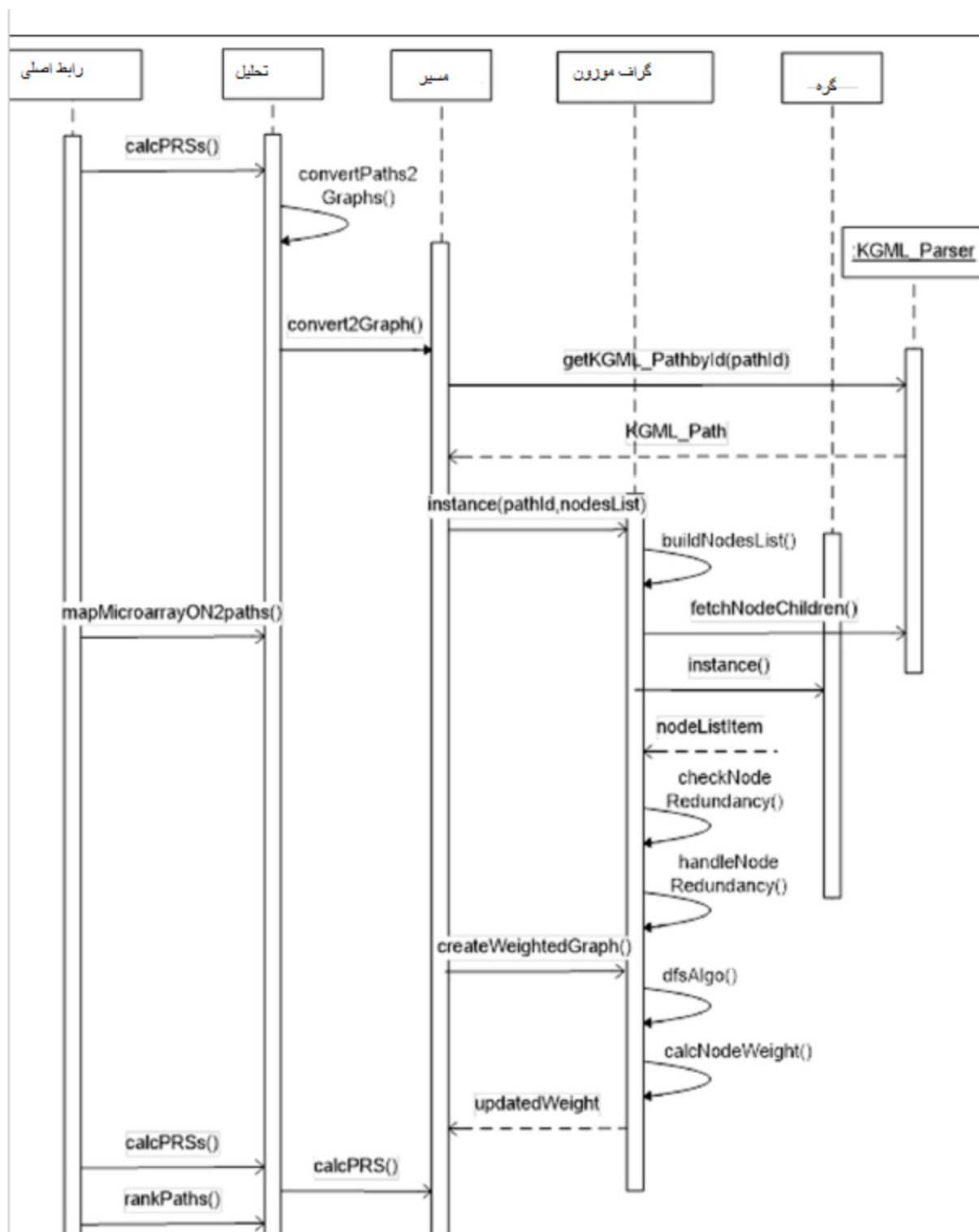
5. امتیاز نظم مسیر (PRS) به هر گراف وزن دهی شده با استفاده از اوزان گره‌های مهم در گراف و دیگر پارامترها اختصاص می‌یابد.

ما تمام این طبقه‌ها، توابع و الگوریتم DFS را با استفاده از MATLAB R2010a اجرا کردیم.

نتایج و بحث

ارزیابی موردی روش‌های تحلیل غنی‌سازی جدید بسته به توانایی آن‌ها در تشخیص فرایندهای بیولوژیکی که در حالت‌های بیماری اختلال ایجاد می‌کنند، دشوار است. ما و دیگران برای این ارزیابی با مطالعه عملکرد در طی طیفی از مجموعه داده‌های ناشی از شرایط مشخص تلاش کردیم ([27] و رفرنس‌های اینجا). با گسترش دادن الگوریتم خود برای شامل شدن مسیرهای بیوشیمی، تحلیل دیگر بر روی یک مجموعه داده انجام دادیم که اختلال متابولیک رایج را توصیف می‌کند که نوع 2 بیماری دیابت است (T2DM). داده‌ها در ابتدا توسط Taneera و همکارانش [31] که سطوح تظاهر ژن را در RNA جدا شده از جزایر پانکراس انسانی که از 9 اهداکننده جسد دیابت نوع 2 (T2D) با نمونه‌های RNA جزایر پانکراس ناشی از 54 اهداکننده جسد غیر دیابتی مقایسه کردند، ایجاد شد. این‌ها در آرایه‌های ST ژن انسانی 1.0 Affymetrix ترکیب شده بودند و مقادیر تظاهر حاصل توسط تحلیل چند آرایه‌ای قوی ((RMA قبل از آپلود شدن در کل تظاهر ژن نرمال شد (www.ncbi.nlm.nih.gov/geo;

GDS4337)#accession. ما یک فایل ورودی شامل شناسه‌های محققین Affymetrix و داده‌های تظاهر ژن نرمال شده برای هر 63 نمونه ایجاد کردیم. دیگر پارامترها نیازمند اعداد نمونه در هر گروه (9 در گروه 1، 54 در گروه 2) و مقادیر آستانه ارزش P و تغییر کل بری فیلتر کردن ژن‌های مهم بودند (در این مورد تغییر کل ≤ 1.3 و مقدار $p > 0.05$). آستانه‌های تغییر کل اتفاقی هستند و مقدار انتخاب‌شده در این مثال تعداد کافی از ژن‌های نهفته به دست داد تا نگاشت مسیر ممکن شود (در این مثال، آستانه 1.5 تنها DEG 88 به دست می‌داد). مصرف‌کننده می‌تواند انتخاب کند که برای علامت‌دهی یا مسیر متابولیک یا هر دو غنی‌سازی کند. تست آماری اضافه‌تر را می‌توان در صورت نیاز با روش جایگشت انجام داد در این مثال ما از تعداد 1000 جایگشت انجام دادیم و آستانه مقدار $p = 0.05$ است. جدول 1 و 2 به ترتیب نشان‌دهنده ده مسیر برتر رتبه‌بندی شده طبق PRS و امتیاز Z است که در آن تنها مسیرهای مهم ($FDR < 0.05$) انتخاب شدند. تعدادی از فرایندهای مربوط به T2DM با هر دو روش انجام شدند، مخصوصاً مسیرها متابولیک مانند "متابولیسم اسید آراشیدونیک" [32]. "متابولیسم اسید چرب" [33,34]، هم‌چنین فرایندهای علامت‌دهی مانند "مسیر علامت‌دهی PPAR" [35,36]. هر دو روش "مسیرهای سرطان" را شناسایی می‌کنند که از آنجایی که این توصیف، تعدادی از فرایندهای موجود در دیابت من‌جمله آپوپتز و چرخه سلولی به همراه علامت‌دهی TGF- β را دربرمی‌گیرد، جای تعجبی ندارد [37]. "تعاریف و آبشارهای انعقادی" با هر دو روش امتیاز بالایی داشتند که می‌توانست یک مثبت خطا باشد یا تغییراتی را در عروق جزایر دیابتی نشان دهد. جدای از این انتظار، گفته می‌شود که تمام دیگر مسیرهای امتیاز بالای PRS در حالت‌های بیماری نهفته می‌شوند. در مقابل، توضیح تعدادی از مسیرهای شناسایی‌شده با امتیازدهی Z دشوارتر است و لذا ممکن است مثبت‌های خطایی باشند ("شبکه ایمنی روده"، "مولکول‌های چسبندگی سلول"، "رد آلوگرافت"، "عفونت استافیلوکوکوس اورئوس"). درنهایت، روش PRS امتیاز بهتری به دو مسیر ضروری برای T2DM، "علامت‌گذاری MAPK" [38] و "دیابت نوع 2" [39] در مقایسه با امتیازدهی Z ارائه کردند. درواقع، توصیف دوم به‌وضوح تأثیر بر علامت‌دهی انسولین و adipocytokine را نشان می‌دهد که برای پاتوفیزیولوژی دیابتی‌ها در مرکزیت است.



شکل 5: نمودار توالی UML که نشان دهنده رتبه بندی مسیر و محاسبه PRS است

جدول 1: ده مسیر برتر رتبه بندی شده توسط PRS مجموعه داده های جزایر پانکراس و T2D

رتبه	نام مسیر	PRS	pvalue	FDR
1	متابولیسم اسید آراشیدونیک،	3.450412	0	0
2	تعامل گیرنده سینوکین سینوکین	1.443531	0	0
3	مسیر علامت دهی بتا TGF	1.345376	0	0
4	آبشارهای انعقاد و مکمل	1.180362	0	0
5	مسیر علامت دهی PPAR	1.030316	0.002	0.0065
6	مسیرها در سرطان	0.910555	0.004	0.0104
7	بیماری دیابت نوع 2	0.793327	0.002	0.0065
8	متابولیسم تریگلیسیرید	0.754089	0.001	0.004875
9	مسیر علامت دهی MAPK	0.736616	0.001	0.004875
10	متابولیسم اسید چرب	0.701842	0.004	0.0104

نتیجه گیری

توسعه سریع فناوریهای پر استفاده ژنومی و نهشت خروجی آنها در پایگاههای داده دسترسی آزاد، مقادیر زیادی از دادهها بیولوژیکی تولید کرده است. استخراج و تفسیر این دادهها، نوآوری در زمینه بیولوژی محاسبه‌ای را ترغیب کرده است که منجر به ظهور ابزار پیچیده برای تولید نتایج قابل اعتماد، معنادار و قابل بررسی شده است. این از این حیث اهمیت دارد که این نوع از آزمایشها گران هستند و ابزارهای جدید احتمالاً باید به تحلیل‌های قبل از این ارزش بیفزاید.

جدول 2: ده مسیر برتر رتبه دهی شده با امتیاز Z مجموعه داده های جزایر پانکراس و T2D

Rank	نام مسیر	Z-score
1	متابولیسم اسید آراشیدونیک،	6.103672
2	تعامل گیرنده سینوکین سینوکین	5.571651
3	مسیر علامت دهی بتا TGF	5.468563
4	آبشارهای انعقاد و مکمل	5.302763
5	مسیر علامت دهی PPAR	5.102405
6	مسیرها در سرطان	5.050608
7	بیماری دیابت نوع 2	4.748036
8	متابولیسم تریگلیسیرید	4.601507
9	مسیر علامت دهی MAPK	4.480696
10	متابولیسم اسید چرب	4.416682

در این مقاله، ما دو حوزه را مدنظر قراردادیم: اول، گسترش الگوریتم غنی‌سازی PRS [27] برای در بر گرفتن مسیرهای متابولیک و علامت‌دهی؛ و دوم، برای ارائه توصیفی دقیق از یک GUI که تحلیل آرایه را با PRS و امتیازدهی Z تسهیل می‌کند. ابزار بهبودیافته، به تعدادی از چالش‌ها مخصوصاً در نگاشت شناسه، فراوانی توصیف مسیر و ارزیابی اهمیت آماری، رسیدگی می‌کند. برخلاف امتیازدهی z، الگوریتم PRS توپولوژی یک مسیر (رابطه بین ژن‌ها) و بزرگی تغییرات تظاهر ژن را در نظر می‌گیرد تا مسیرهای نهفته را شناسایی کند. برای این دلایل، ما استدلال می‌کنیم که غنی‌سازی PRS، بینش‌های بیولوژیکی‌تری در مقایسه با روش استاندارد فوق هندسی به دست می‌دهد. از آنجایی که گام‌های پیش‌پردازش اضافی اتخاذ شده برای کاهش فراوانی در توصیفات KEGG به راحتی در دیگر روش‌ها بدون مهندسی مجدد قابل ملاحظه اجرا نمی‌شود، مقایسه عملکرد با دیگر روش‌های PT ممکن نیست. البته رفتار مسیرهای علامت‌گذاری و متابولیک، مشخص است. اگرچه، از آنجایی که رویکرد ما به جای پیش‌بینی تأثیری بر تابع مسیر، ارزیابی تغییرات نسخه‌برداری در یک مسیر بود، ما احساس کردیم که معقول است تا تأثیر بر مسیرهای علامت‌دهی و بیوشیمیایی را با استفاده از یک روش منفرد ارزیابی کنیم. بدین طریق، ما قادر بودیم تا مسیرهای بیوشیمیایی که در بیماری متابولیک اختلال ایجاد می‌کند، شناسایی کنیم. یک اصل اساسی این نوع از تحلیل این است که دانشمندان زیست پزشکی در بررسی‌های متعاقب اهداف نشان داده شده توسط مطالعات نمای نسخه‌برداری هدایت می‌شوند. متأسفانه، هیچ تست آماری غیرمبهمی وجود ندارد که به محققان اجازه دهد تا مطمئن باشند که هر مسیر مشخص‌شده‌ای ارزش مطالعه بیشتر را دارد (و هزینه گزاف). کاربرد رویکردهای مبتنی بر جایگشت اغلب برای تعیین احتمال یک امتیاز غنی‌سازی که به صورت اتفاقی به دست آمده است به کار می‌رود و برای تنظیم مقادیر p با FDR می‌توان اعتماد به نفس محققان را افزایش داد که یک نتیجه معنادار است.

به طور خلاصه، ما استدلال می‌کنیم که فراهم کردن گزینه انتخاب ابزار تحلیل برای محققان که با توجیهات مشخصی نسبت به آن آگاه شده‌اند، امکان ترکیب شدن یا متضاد شدن مدارک را برای تسهیل تصمیم‌گیری آگاهانه‌تر به وجود می‌آورد.

در دسترس بودن و الزامات

نام پروژه: نرم افزار PRS

صفحه اصلی پروژه: <http://www.buckingham.ac.uk/research/clore-laboratory-diabetes-obesity-and-metabolic-research/staff/maysson-al-haj-ibrahim/prs-tool/>.

سیستم عملیاتی: بسترهای نرم افزاری مستقل

زبان برنامه نویسی: MATLAB

دیگر الزامات: MATLAB 2010a یا بالاتر. اگر MATLAB بر روی کامپیوتر شخصی شما نصب نشده است، شما

ابتدا نیاز به نصب محیط MCR و سپس اجرای ابزار PRS دارید.

محدودیت های استفاده: ندارد.

فایل های اضافی

فایل اضافی 1: مواردی که تشکیل دهنده مسیرهای KEGG هستند و در یک فایل KGML نشان داده شده اند.

فایل اضافی 2: نمودار توالی UML که نشان دهنده اجرای تابع ریزآرایه "initialize" است.

فایل اضافی 3: نمودار توالی UML تابع KGML parse.

فایل اضافی 4: نمودار توالی UML که نشان دهنده اجرای جزئیات فرایند ایجاد فهرستی از مسیرهای KEGG و

نگاشت داده های ریزآرایه درون آن هاست.

References

1. Glazko GV, Emmert-Streib F: **Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets.** *Bioinformatics* 2009, **25**(18):2348–2354.
2. Khatri P, Sirota M, Butte AJ: **Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges.** *PLoS Comput Biol* 2012, **8**(2):e1002375.
3. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR: **GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways.** *Nat Genet* 2002, **31**(1):19–20.
4. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN: **GoMiner: a resource for biological interpretation of genomic and proteomic data.** *Genome Biol* 2003, **4**(4):R28.
5. Khatri P, Draghici S, Ostermeier GC, Krawetz SA: **Profiling Gene Expression Using Onto-Express.** *Genomics* 2002, **79**(2):266–270.
6. Al-Shahrour F, Díaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.** *Bioinformatics* 2004, **20**(4):578–580.
7. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102**(43):15545.
8. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies.** *Proc Natl Acad Sci USA* 2005, **102**(38):13544.
9. Barry WT, Nobel AB, Wright FA: **Significance analysis of functional categories in gene expression studies: a structured permutation approach.** *Bioinformatics* 2005, **21**(9):1943–1949.
10. Boorsma A, Foat BC, Vis D, Klis F, Bussemaker HJ: **T-profiler: scoring the activity of predefined groups of genes using gene expression data.** *Nucleic Acids Res* 2005, **33**:W592–W595.
11. Kim S-B, Yang S, Kim S-K, Kim SC, Woo HG, Volsky DJ, Kim S-Y, Chu I-S: **GAzer: gene set analyzer.** *Bioinformatics* 2007, **23**(13):1697–1699.
12. Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, Muller R, Meese E, Lenhof H-P: **GeneTrail-advanced gene set enrichment analysis.** *Nucleic Acids Res* 2007, **35**:W186–W192.
13. Mitrea C, Taghavi Z, Bokanizad B, Hanoudi S, Tagett R, Donato M, Voichița D, Drăghici S: **Methods and approaches in the topology-based analysis of biological pathways.** *Front Physiol* 2013, **4**:278.
14. MetaCore™: [<http://thomsonreuters.com/metacore/>]
15. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A: **EnrichNet: network-based gene set enrichment analysis.** *Bioinformatics* 2012, **28**(18):i451–i457.
16. Amin K: *Pathway-express: A Bioinformatics Tool for Pathway Level Analysis Using Gene Expression Data.* ; 2007. Master thesis. Wayne State University, Graduate School.
17. Fang Z, Tian W, Ji H: **A network-based gene-weighting approach for pathway analysis.** *Cell Res* 2012, **22**(3):565–580.
18. Greenblum SI, Efroni S, Schaefer CF, Buetow KH: **The PathOlogist: an automated tool for pathway-centric analysis.** *BMC Bioinformatics* 2011, **12**(1):133.

19. Jacob L, Neuvial P, Dudoit S: *Gains in power from structured two-sample tests of means on graphs.* ; 2010.
20. Mieczkowski J, Swiatek-Machado K, Kaminska B: **Identification of Pathway Deregulation – Gene Expression Based Analysis of Consistent Signal Transduction.** *PLoS One* 2012, **7**(7):e41541.
21. Khatri P, Voichita C, Kattan K, Ansari N, Khatri A, Georgescu C, Tarca AL, Draghici S: **Onto-Tools: new additions and improvements in 2006.** *Nucleic Acids Res* 2007, **35**:W206–W211.
22. Shojaie A, Michailidis G: **Analysis of Gene Sets Based on the Underlying Regulatory Network.** *J Comput Biol* 2009, **16**(3):407–426.
23. Rahnenführer J, Domingues FS, Maydt J, Lengauer T: **Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data.** *Stat Appl Genet Mol Biol* 2004, **31**:1544–6115.
24. Gao S, Wang X: **TAPPA: topological analysis of pathway phenotype association.** *Bioinformatics* 2007, **23**(22):3100–3102.
25. Xia J, Wishart DS: **MetPA: a web-based metabolomics tool for pathway analysis and visualization.** *Bioinformatics* 2010, **26**(18):2342–2344.
26. Martini P, Sales G, Massa MS, Chiogna M, Romualdi C: **Along signal paths: an empirical gene set approach exploiting pathway topology.** *Nucleic Acids Res* 2012, **41**(1):e19.
27. Ibrahim MA, Jassim S, Cawthorne MA, Langlands K: **A Topology-Based Score for Pathway Enrichment.** *J Comput Biol* 2012, **19**(5):563–573.
28. **Kyoto Encyclopaedia of Genes and Genomes**, data retrieved May 2012 from [http://www.genome.jp/kegg/]
29. Cheadle C, Vawter MP, Freed WJ, Becker KG: **Analysis of Microarray Data Using Z Score Transformation.** *J Mol Diagn* 2003, **5**(2):73–81.
30. Unified Modeling Language™ (UML®): [http://www.uml.org/]
31. Taneera J, Lang S, Sharma A, Fadista J, Zhou Y, Ahlqvist E, Jonsson A, Lyssenko V, Vikman P, Hansson O, Parikh H, Korsgren O, Soni A, Krus U, Zhang E, Jing X-J, Esguerra JLS, Wollheim CB, Salehi A, Rosengren A, Renström E, Groop L: **A Systems Genetics Approach Identifies Genes and Pathways for Type 2 Diabetes in Human Islets.** *Cell Metab* 2012, **16**(1):122–134.
32. Persaud SJ, Muller D, Belin VD, Kitsou-Mylona I, Asare-Anane H, Papadimitriou A, Burns CJ, Huang GC, Amiel SA, Jones PM: **The Role of Arachidonic Acid and Its Metabolites in Insulin Secretion From Human Islets of Langerhans.** *Diabetes* 2007, **56**(1):197–203.
33. Yaney GC, Corkey BE: **Fatty acid metabolism and insulin secretion in pancreatic beta cells.** *Diabetologia* 2003, **46**(10):1297–1312.
34. McGarry JD: **Banting lecture 2001 Dysregulation of fatty acid metabolism in the etiology of type 2 diabetes.** *Diabetes* 2002, **51**(1):7–18.
35. Sugden MC, Holness MJ: **Potential Role of Peroxisome Proliferator-Activated Receptor- α in the Modulation of Glucose-Stimulated Insulin Secretion.** *Diabetes* 2004, **53**(1):S71–S81.
36. Kim H-S, Hwang Y-C, Koo S-H, Park KS, Lee M-S, Kim K-W, Lee M-K: **PPAR- γ Activation Increases Insulin Secretion through the Up-regulation of the Free Fatty Acid Receptor GPR40 in Pancreatic β -Cells.** *PLoS One* 2013, **8**(1):e50128.