دانشگاه تربیت مدرس
دانشکده فنی و مهندسی
گروه مهندسی فناوری اطلاعات

گزارش سمینار

کاربرد داده‌کاوی پروتئوم در دسته‌بندی سرطان و کشف زیست‌نشانگرها

دانشجو

رسول نوروزی

استاد راهنما

دکتر امیر البدوی

تابستان ۱۳۹۶


بسم الله الرحمن الرحيم

چکیده

سرطان بیماری بسیار پیچیده است که در سطح مولکولی در بدن پدیدار می‌شود. سرطان‌ها در شکل سلول‌ها، الگو و کمیت زیست‌نشانگرها به‌خصوص پروتئین‌های موجود در بافت و سلول تأثیر می‌گذارند. در این میان علم پروتئومیکس که به بررسی الگوها و میزان بیان پروتئین‌ها در بافت و سلول می‌پردازد می‌تواند یکی از کلیدهای حل مسئله سرطان باشد. پروتئومیکس شامل رویکردها و فن‌آوری‌های متنوعی است که دراین‌بین رویکرد پروتئومیکس شناختی و مقایسه‌ای و فن‌آوری طیف‌سنج جرمی در پیوستگی استفاده از روش‌ها تحلیلی سطح بالا همچون داده‌کاوی در دسته‌بندی سرطان و کشف زیست‌نشانگرهای مربوطه می‌تواند بسیار راهگشا باشد. در این گزارش به بررسی رویکردهای علم پروتئومیکس و فن‌آوری‌های بکار رفته در آن خواهیم پرداخت و سپس تمرکز مطالعات را بر روی پروتئومیکس شناختی و مقایسه‌ای با داده‌های حاصل از فن‌آوری طیف‌سنج خواهیم گذاشت. پس از آن به علم داده‌کاوی و نحوه استفاده از داده‌کاوی با داده‌های حاصل از طیف‌سنج جرمی برای دسته‌بندی و کشف زیست‌نشانگرها خواهیم پرداخت و در آخر هم مروری بر تحقیقات گذشته خواهیم داشت.

کلیدواژه‌ها: طیف‌سنج جرمی، سرطان، داده‌کاوی، زیست‌نشانگر

فهرست مطالب

عنوان **شماره صفحه**

فصل دوم پروتئومیکس و کاربردهای علوم کامپیوتری در داده های حاصل از طیف سنج جرمی

فصل سوم داده‌کاوی و کاربرد آن در دسته‌بندی و کشف زیست‌نشانگرهای سرطانی

فصل چهارم جمع‌بندی و نتیجه‌گیری

فهرست شکل‌ها و نمودارها

موضوع **شماره صفحه**

فهرست جدول‌ها

موضوع **شماره صفحه**

فصل اول

مفاهیم و کلیات

۱,۱ مقدمه

سرطان بیماری پیچیده‌ای است که در سطح مولکولی در بدن انسان ظاهر می‌شود بنابراین برای شناخت بهتر سرطان نیاز است در ابتدای این فصل نگاهی به انسان در سطح مولکولی داشته باشیم در ادامه سرطان را از دیدگاه مولکولی و ارتباط زیست‌نشانگرها با سرطان در بدن انسان را تشریح خواهیم کرد در قدم بعدی در راستای مطالب عنوان‌شده بیان مسئله خود و دلیل استفاده از پروتئین‌ها به‌عنوان زیست‌نشانگرها منتخب خواهیم پرداخت. در قسمت بعدی از ضرورت پرداختن به بیماری سرطان خواهیم گفت سپس برخی از اصطلاحات که برای وارد شدن به دنیای علم پروتئومیکس نیاز است را تعریف خواهیم کرد، همچنین علائم و اختصارات استفاده شده را فهرست خواهیم کرد پس‌ازآن ساختار کلی تحقیق را شرح خواهیم داد و در آخر خلاصه‌ای از مطالب بیان‌شده در فصل را مرور خواهیم کرد.

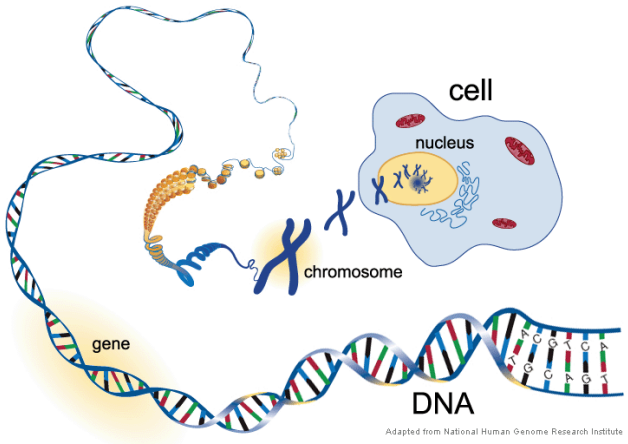
۱,۲ انسان از منظر زیست‌شناسی

از منظر زیست‌شناسی می‌توان انسان را از منظر چهار ابرمولکول[1]: کربوهیدرات‌ها، پروتئین‌ها، اسیدهای‌نوکلئوتید و لیپیدها نگریست. گرچه مولکول‌های بسیار دیگری در ساختار بدن وجود دارند اما موارد ذکرشده مهم‌ترین و تأثیرگذارترین آن‌ها می‌باشند. این ابرمولکول‌ها به‌صورت پلیمرهای بلندی می‌باشند که از ترکیب مونومورها (همچنان که مرواریدهای یک گردنبند به کمک یک‌رشته به هم متصل شده‌اند) با کمک سنتز پسابش[2] ایجادشده‌اند. هرکدام از این ابرمولکول‌های بیولوژیکی دارای زیرمجموعه‌ها و به‌تبع آن دارای خواص و وظایف منحصربه‌فردی در بدن می‌باشند. به‌طور مثال کربوهیدرات‌ها شامل سه دسته منوساکاریدها، دی‌ساکاریدها و پلی‌ساکاریدها می‌شوند که وظایفی همچون منبع ذخیره کننده انرژی و تشکیل ساختار دیواره‌های برخی از سلول‌ها را دارا می‌باشند. همچنان لیپیدها که شامل دسته‌های: چربی‌ها و روغن‌ها، واکس‌ها، فسفولیپیدها و آستروئیدها می‌باشند که وظایفی همچون ذخیره انرژی، ایجاد عایق برای سلول‌ها، ایجاد غشای سلولی و... بر عهده‌دارند؛ اما در ارتباط با اسیدهای نکلئوتید و پروتئین‌ها با توجه به این‌که اساس مطالعه ما در این تحقیق می‌باشند نگاه دقیق‌تر و موشکافانه‌تر خواهیم داشت.

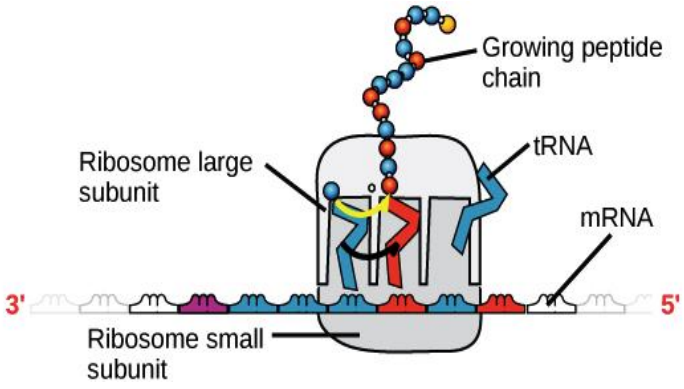
[1] macromolecule
[2] dehydration synthesis

۱,۲,۱ اسیدهای نوکلئوتید

اسیدهای نوکلئوتید از واحدهایی به نام نوکلئوتید تشکیل‌شده‌اند که در طبیعت به دو شکل اسید دئوکسی ریبونوکلئیک (DNA) واسید ریبونوکلئیک (RNA) یافت می‌شوند. هر واحد نوکلئوتید از سه بخش حلقه‌های نیتروژن دار، قند با پنج کربن و حداقل یک گروه فسفات تشکیل‌شده است. DNA معمولاً به‌صورت مارپیچ دوگانه یافت می‌شوند. چهار نوع مختلف از مولکول‌های نوکلئوتیدی در زنجیره DNA یافت می‌شوند. این چهار مولکول ازنظر ساختمانی شباهتی بسیاری به هم دارند لیکن دارای تفاوت‌های کافی جهت تشخیص‌شان توسط فرآیندهای مربوطه نیز می‌باشند. قسمت‌های ویژه از هر چهار نوکلئوتید که به‌صورت حلقوی هستند، باز نامیده می‌شود. این حلقه‌ها در اثر برقراری پیوند شیمیایی بین اتم‌های کربن و نیتروژن به وجود آمده‌اند. سایر اتم‌ها (فسفر، هیدروژن و اکسیژن) به‌طور مستقیم به این حلقه‌ها متصل شده‌اند چهار حلقه مختلف آدنین، تیمین، گوانین و سیتوزین نامیده شده و توسط حروف A,T,C,G نمایش داده می‌شوند. واحدهای نوکلئوتیدی توسط گروه فسفات به هم متصل می‌شوند. ترکیب‌های نوکلئوتیدی قواعدی دارند. تیمین همراه با آدنین و سیتوزین همراه با گوانین جفت می‌شوند.(شکل ۱-۱) این چهار حلقه همچون حروف الفبا برای یادداشت‌برداری و نگه‌داری اطلاعات ژنتیکی عمل می‌کند.(علی پور، محمد،۱۳۸۴)RNA برخلاف DNA معمولا از یک رشته بلند تشکیل‌شده است. یک نوکلئوتید در RNA می‌تواند شامل ریبوز (قند پنج کربن)، یکی از چهار حلقه نیتروژنی (A,U,G یا C) و یک گروه فسفات شود. RNA انواع مختلفی دارد که هرکدام دارای وظایفی خاصی در سلول می‌باشند مانند mRNA، tRNA و...



شکل ۱-۱از سلول تا دی‌ان‌ای (وبسایت سازمان ملی ژنتیک انسانی آمریکا)



شکل۲-۱ فرآیند ترجمه شدن و ساخت پروتئین (وبسایت سازمان ملی ژنتیک انسانی آمریکا)

اسیدهای نوکلئوتید و بخصوص DNA ابرمولکول‌های کلیدی برای ادامه حیات انسان می‌باشند. DNA اطلاعات وراثتی که از والد به فرزند منتقل می‌شود را دارا است همچنان که دربردارنده اطلاعات چگونگی، میزان و زمان ساخت پروتئین‌ها و عملکرد سلول‌ها، بافت‌ها و ارگان‌ها است. در یوکاریت‌ها (موجودات دارای هسته در سلول)، DNA معمولاً به تعدادی از قطعات بسیار بلند و باریک خطی به نام کروموزوم که دارای تعداد مشخصی می‌باشند شکسته می‌شود. برای مثال تعداد کروموزوم‌های سلول انسان ۲۳ جفت (۴۶ تا) است؛ اما شکل کروموزوم در پرویوکاریت‌ها (موجودات فاقد هسته در سلول مانند ویروس‌ها) به‌صورت مدور و حلقه‌ای شکل است. کروموزوم‌ها هرکدام می‌توانند حاوی ۱۰ تا ۱۰۰ هزار ژن باشند که هرکدام دستورالعمل ساخت فراورده موردنیاز سلول را مهیا می‌کنند مانند دستورالعمل ساخت پروتئین‌ها که توالی اسیدهای آمینه آن را مشخص می‌کنند که به این عمل اصطلاحاً

کدگذاری[3] گفته می‌شود. قبل از این‌که این اطلاعات استفاده شود برای ساخت پروتئین‌ها، RNA ها یک کپی از ژن‌ها تهیه می‌کنند سپس این کپی از ژن‌ها به ریبوزوم، ماشین سلولی که توالی RNA را می‌خواند (فرآیند ترجمه شدن[4]) و از آن برای ساخت پروتئین‌ها استفاده می‌کند. (شکل ۲-۱)

۱,۲,۲ پروتئین‌ها

پروتئین‌ها از فراوان‌ترین مولکول‌های ارگانیک در بدن موجودات زنده هستند و بیشترین تنوع در ساختار و عملکرد را در بین ابرمولکول‌های زیستی دارند. یک سلول به‌تنهایی می‌تواند شامل صدها هزار پروتئین با عملکرد مجزا و منحصر باشد. تمام پروتئین‌ها از یک یا تعدادی از زنجیره‌های آمینواسید تشکیل‌شده است که هرکدام از آن‌ها یک پلی‌پپتاید[5] نامیده می‌شود. آمینواسیدها مونومورهایی هستند که تشکیل‌دهنده پروتئین‌ها می‌باشند. درمجموع تابه‌حال ۲۰ نوع آمینواسید در پروتئین‌ها یافت شده است.

پروتئین‌ها دارای نقش‌های متنوعی در سلول‌ها و ارگان‌های بدن می‌باشند؛ که دودسته مهم آن‌ها به شرح زیر است:

آنزیم‌ها: دسته‌ای از پروتئین‌ها می‌باشند که به‌عنوان کاتالیزور برای سرعت بخشی به واکنش‌های بیوشیمیایی عمل می‌کنند. هر آنزیم یک یا چند مولکول را برای کاتالیز آن در واکنش‌های بیوشیمیایی به‌عنوان زیرمجموعه خود به رسمیت می‌شناسد. به‌طور مثال آنزیم بزاق آمیلاز[6]، برای شکستن مولکول آمیلاز (نوعی نشاسته) به قطعات کوچک قندی است.

هورمون‌ها: دسته دیگری از پروتئین‌های مهم در بدن موجودات زنده می‌باشند که فرآیندهای فیزیولوژی بخصوصی مانند رشد، توسعه، متابولیسم و تولیدمثل را به عهده‌دارند. برخی از هورمون‌ها زیرمجموعه لیپیدها هستند و به پایه-استروئید شناخته می‌شوند و مابقی زیرمجموعه پروتئین‌ها معروف به پایه-پپتید قرار می‌گیرند.(Garrett and Grisham, 2005, Watson et al., 2003)

۱,۳ سرطان

سرطان یک بیماری ژنتیکی است که ۲۷۷ نوع بیماری را شامل می‌گردد. تغییرات ژنتیکی باعث ازهم‌گسیخته شدن نظم طبیعی تقسیم و تمایز سلول‌ها می‌شود. همچنین در محیط‌زیست ما بیش از یکصدهزار نوع مواد شیمیایی وجود دارد که فقط ۳۵ هزار از آن آنالیز شده و حدود ۳۰۰ عدد از آن‌ها تولید سرطان می‌کند. هنوز ۶۵ هزار مواد شیمیایی باقیمانده در طبیعت آزمایش نشده است. در حال حاضر بیش از ۵۰ درصد بیماری‌های سرطانی را معالجه می‌نماییم مخصوصاً اگر این بیماری در مراحل اولیه تشخیص داده شوند.(پارسا, ۲۰۱۲) موارد زیر برخی از انواع سرطان می‌باشند که در سلول‌های به خصوصی در بدن شروع می‌شوند:

- کارسینوما: در سلول‌های مخاطی به وجود می‌آید که توانایی حمله به ماهیچه‌های اطراف خود را نیز دارا است. سرطان‌های سینه، پروستات، ریه و روده بزرگ از نوع کارسینوما می‌باشند.

[3] encode
[4] translation
[5] polypeptide
[6] Salivary amylose

- سارکوما: یک تومور بدخیم در استخوان یا بافت‌های نرم است. سارکوما اساس سرطان‌های استخوانی است.
- لنفاوی: سرطانی است که غدد لنفاوی را موردحمله قرار می‌دهد و قادر به انتقال به سایر بافت‌های بدن نیز است.
- سرطان خون[۷]: نوعی سرطان که بافت سلول‌های خون، گلبول‌های سفید خون و مغز استخوان را شکل می‌دهد.

از عوامل اساسی ایجاد سرطان می‌توانیم به: الکل، دخانیات، اضافه‌وزن و چاقی، مشکلات در سیستم ایمنی بدن، عفونت و ... اشاره‌کنیم.(Shukla et al., 2016)

سرطان از روش‌های جراحی، شیمی‌درمانی، اشعه درمانی، ایمنو درمانی، ژن‌درمانی و یا تلفیقی از آن‌ها درمان می‌شود.

۱,۳,۱ اساس مولکولی

اما آنچه در سرطان برای ما بیشترین اهمیت را دارا است اساس مولکولی و سازوکار ایجاد آن است. در یک تعریف ساده می‌شود سرطان را با فعالیت نابجای پروتئین‌های کنترل‌کننده چرخه سلول که موجب ازدیاد در تقسیم سلولی می‌شود نامید. یک سلول سرطانی تفاوت‌های با یک سلول سالم دارا است. سلول‌های سرطانی بدون نیاز به سیگنال‌های لازم برای تکثیر برخلاف سلول‌های سالم، شروع به تقسیم می‌کنند. همچنین سیگنال‌هایی را که برای توقف تکثیر می‌باشند را نادیده می‌گیرند. سلول‌های سالم تنها قادر به تکثیر بین ۴۰-۶۰ بار در طول عمر خود هستند اما برای سلول‌های سرطانی این چرخه به‌مراتب بیشتر است. سلول‌های سرطانی قابلیت حرکت به سمت سایر بافت‌ها و ارگان‌های بدن را دارا می‌باشند که به آن اصطلاحاً متاستاز[۸] گفته می‌شود همچنان که می‌توانند موجب رشد ماهیچه‌های خونی شوند که اصطلاحاً رگ زایی[۹] گفته می‌شود. سلول‌های سرطانی همچنان از فرمان خودکشی سلولی[۱۰] که تحت شرایط خاصی مانند معیوب شدن سلول‌ها برای آن برنامه‌ریزی می‌شوند سرپیچی می‌کنند.(Sosa et al., 2014)

سرطان‌های مختلف شامل جهش‌های ژنتیکی[۱۱] مختلفی نیز می‌باشند و هرکدام تغییرات بخصوصی در ساختار ژن‌ها ایجاد می‌کنند. به‌طورکلی جهش ژنتیکی در دو نوع از تنظیم‌کننده‌های چرخه سلولی موجب توسعه سرطان می‌شود:

۱. تنظیم‌کننده‌های مثبت (آنکو ژن‌ها[۱۲]) که ممکن است بیش‌ازاندازه فعالیت کنند. سلول‌ها تا زمانی که بیش‌فعالی نرسیده باشند پروتو-آنکوژن نامیده می‌شوند ولی ممکن است با تغییر توالی اسیدآمینه‌های پروتئین موجب تغییر در شکل پروتئین موجب همیشه فعال پروتئین شوند. از سوی دیگر ممکن است موجب تقویت شوند، به این صورت که یک سلول مقادیر زیادی کپی از یک ژن تهیه کند و درنتیجه موجب ازدیاد تولید در پروتئین‌ها شود. همچنین بسیاری از پروتئین‌ها سیگنال‌ها و دستورات تکثیر را توسط پروتو-آنکوژن‌ها زمانی که شرایط مساعد باشند منتقل می‌کنند ولی ممکن است یک جهش باعث بیش فعالی پروتئین‌ها شود و سیگنال‌های تکثیر را حتی زمانی که شرایط تکثیر مناسب نیست انتقال دهد.

---

[7] Leukemia
[8] metastasis
[9] angiogenesis
[10] apoptosis
[11] mutation
[12] oncogenes

۲. تنظیم‌کننده‌های منفی که سرکوب کننده‌ها[13] تومور نیز نامیده می‌شوند ممکن است غیرفعال شوند. زمانی که DNA آسیب می‌بیند یک پروتئین به نام P53 چرخه سلولی را متوقف می‌کند. این توقف باعث می‌شود تا این پروتئین آنزیم‌های ترمیم‌کننده DNA را فعال کند. اگر آسیب ترمیم پیدا کرد این پروتئین اجازه ازسرگیری فعالیت چرخه سلولی را می‌دهد در غیر این صورت آخرین وظیفه خود یعنی موجب سازی خودکشی سلولی را فراهم می‌کند تا جهش ژنتیکی ادامه پیدا نکند.(Otto and Sicinski, 2017, Joerger and Fersht, 2016)

۱,۴ بیان مسئله

پس از تکمیل پروژه ژنوم انسان، تمرکز محققین به سمت شناخت ساختار، عملکرد و تعامل پروتئین‌ها و نقش آن‌ها در بیماری‌ها که توسط ژن‌ها تولید می‌شود معطوف گردیده است. این تغییر جهت به دلیل: ۱. سطح بیان mRNA معمولاً بیانگر میزان دقیق پروتئین‌های فعال در سلول نیست ۲. توالی ژن‌ها تغییرات پس از ترجمه اصلاحات پروتئین‌ها را که ممکن است برای عملکرد درست پروتئین‌ها ضروری باشد بیان نکند ۳. مطالعات ژنتیک قادر به توصیف فرآیندهای پویای سلولی نیست.(Li et al., 2004) پروتئوم به سری کامل پروتئین‌های بیان‌شده در یک‌لحظه خاص در یک سلول موردنظر اشاره دارد ، بااین‌حال امروزه سطح این تعریف، از سلول به بافت، اندام و ارگانیسم نیز تعمیم داده‌شده است.(شیردل et al., 2013) به‌تبع آن پروتئومیکس یک واژه کلی بوده که به مطالعه پروتئین‌ها در ابعاد گسترده از جنبه‌های، پروتئومیکس شناختی و مقایسه‌ای، هستی‌شناسی[14] پروتئین‌ها، تعاملات پروتئین-پروتئین، مسیرهای متابولیکی و توصیف پروتئین‌ها است که پژوهش حاضر به پروتئومیکس شناختی و مقایسه‌ای پروتئین‌هایی که با طیف‌سنج جرمی حاصل‌شده‌اند می‌پردازد.(Lam et al., 2014)

تغییرات بالینی ممکن است در الگوی پروتئومیکی یک ارگان یا بافت بازتاب یابد؛ بنابراین امکان‌پذیر است که پروتئین‌های موجود در یک نمونه خاص برای تشخیص بیماران سرطانی از غیر سرطانی استفاده شود.(Li et al., 2004) با پیشرفت روش‌های نمونه‌گیری و تفکیک پروتئین‌ها میزان داده‌های قابل‌دسترسی به‌طور فزاینده‌ای افزایش‌یافته است که به‌سادگی قابل تصویر و تفسیر نیستند و با توجه این‌که در داده‌های پروتئینی معمولاً تعداد نمونه‌ها به دلیل دشواری و هزینه در نمونه‌گیری کم می‌باشند و ابعاد یا به عبارتی تعداد ویژگی‌های آن‌ها بسیار زیاد است، نیاز به ابزار تحلیلی سطح بالا همچون داده‌کاوی و الگوریتم‌های پیشرفته هوش مصنوعی و یادگیری ماشین احساس می‌شود. هدف اصلی و اساسی استخراج اطلاعاتی است که منجر به کشف الگوهایی برای دسته‌بندی سرطان و شناسایی زیست‌نشانگرهایی همچون پروتئین‌ها که به‌طور بالقوه ابزاری قدرتمند برای شناسایی، تشخیص و پیشگیری از بیماری‌ها به‌خصوص سرطان است.(Thomas et al., 2006)

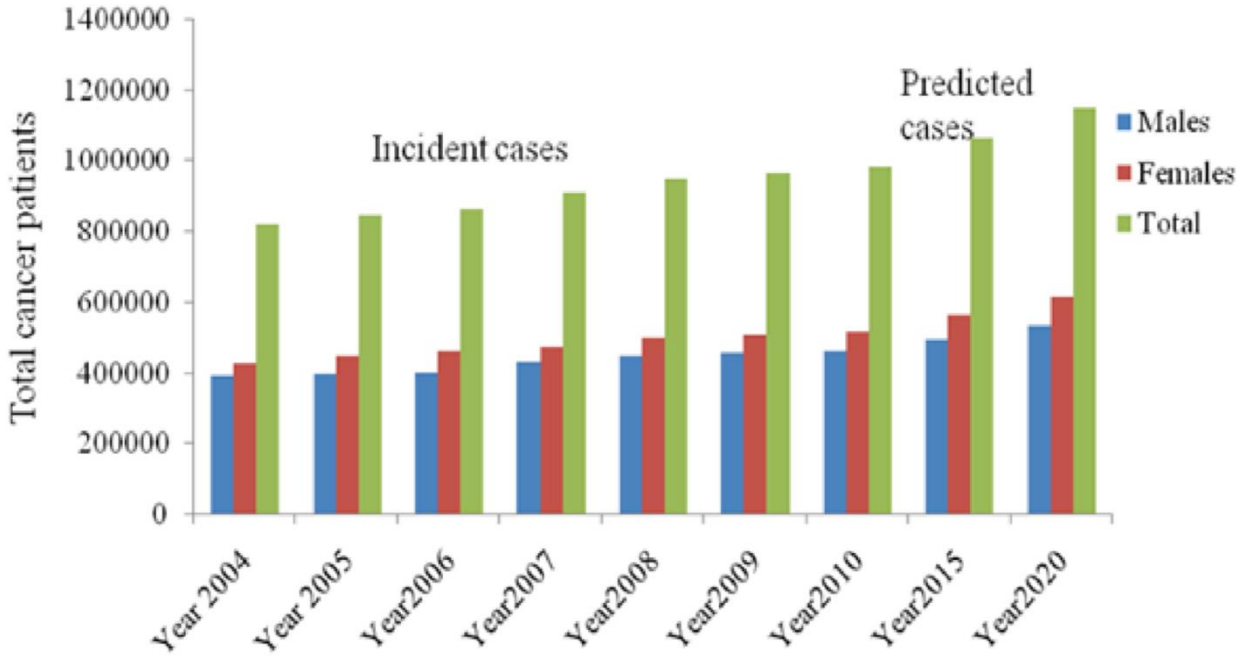
[13] Tumor suppressor
[14] ontology

۱,۵ ضرورت مسئله

بنا بر آمار سازمان جهانی بهداشت سالانه ۸,۲ میلیون نفر انسان در سراسر جهان به دلیل سرطان می‌میرند که این مقدار ۱۳٪ از کل میزان مرگ‌ومیر جهان است. پیش‌بینی می‌شود که به این میزان تا ۷۰٪ تا دو دهه آینده اضافه شود. بیشترین میزان سرطان کشف‌شده به ترتیب مربوط به سرطان ریه (۱,۸ میلیون، ۱۳٪ از کل)، سرطان سینه (۱,۷ میلیون، ۱۲٪ از کل) و کلورکتال (۱,۴ میلیون نفر، ۹,۷٪ از کل) است. تخمین زده می‌شود تا سال ۲۰۲۵ این میزان به ۱۹,۳ میلیون نفر به دلیل رشد جمعیت افزایش یابد.(شکل ۳-۱) (Shukla et al., 2016)

تعداد بیماران سرطانی سال‌به‌سال رو به افزایش است و این خود یک معضل پزشکی است نه‌فقط ازنظر بهداشت و درمان بلکه ازنظر اقتصادی می‌تواند کشورها را تا حد ورشکستی اقتصادی پیش ببرد. اگر سرطان‌ها در مرحله اول تشخیص داده شوند به‌طور کامل قابل معالجه هستند و اگر در مراحل دوم تشخیص داده شوند در حدود ۷۰ درصد شانس معالجه دارند و گر در مراحل سوم تشخیص داده شوند در حدود ۳۰ درصد شانس بهبودی دارند و اگر تشخیص در مرحله چهارم بوده باشد به‌طور حتم به بافت‌های دیگر گسترده شده است، شانس معالجه و بهبودی در حدود پنج درصد است که پنج سال ادامه حیات داشته باشد.

در سه دهه گذشته، محققین اطلاعات زیادی را درباره ژن‌ها و پروتئین‌ها و نقش آن‌ها در تولید سلول‌های طبیعی و سرطانی گزارش نموده‌اند. یکی از اکتشافات مهم آن‌ها، نقش ژن‌های جهش‌یافته در تولید سلول‌های سرطانی بوده است. عوامل محیطی باعث جهش‌های ژنتیکی می‌شوند در حال شناسایی هستند. با کمک از روش‌های مختلف مولکولی قادر هستیم که قدرت بیان ژن‌ها و پروتئین‌های معیوب را تعیین نماییم. حتی پیدا کردن زیست‌نشانگرهای جدیدی مانند پروتئین که شاخص یک نوع سرطان هستند در تشخیص زودرس و معالجه به‌موقع بیماری سرطان کمک‌های شایان توجهی را می‌نماید که این امر با به‌کارگیری علوم رایانه‌ای به‌خصوص داده‌کاوی ممکن می‌شود. پس از تعیین شکل فضایی پروتئین‌های معیوب می‌توان داروهای ضد سرطان جدیدی را ساخت که بتواند سلول‌های در حال سرطانی شدن را مورد هدف قرار بدهند تا از تولید و رشد آن‌ها به سلول‌های سرطانی جلوگیری شود. (پارسا, ۲۰۱۲)



شکل ۳-۱یش بینی نرخ رشد سرطان تا سال ۲۰۲۰ (Shukla et al., 2016)

۱,۶ تعاریف و اصطلاحات

قبل از ورود به فصل آتی جهت بررسی زوایای موضوع و درک مسئله نیاز به تعریف برخی از اصطلاحات عمده داریم که در اینجا به آن‌ها اشاره خواهیم کرد.

آنتی‌بادی[۱۵]: آنتی‌بادی یا ایمونوگلوبولینها پروتئین‌هایی می‌باشند که توسط سیستم ایمنی بدن برای آمیختن با آنتی‌ژن‌های سلول‌های (مانند باکتری و ویروس) بیگانه جهت خنثی‌سازی آن‌ها استفاده می‌شوند.

آنتی‌ژن[۱۶]: آنتی‌ژن‌ها مولکول‌های بیگانه هستند که موجب تولید آنتی‌بادی در ارگان می‌شوند.

پپتید: پپتیدها زنجیره‌های کوچکی از آمینواسیدها می‌باشند.

پلی‌پپتاید: یک زنجیره ابرمولکولی از آمینواسیدها که ممکن است یک پروتئین یا یکی اجزا پروتئین باشد.

آنزیم: پروتئین‌هایی می‌باشند که به‌عنوان کاتالیزور در فرآیندهای شیمیایی داخل بدن شرکت می‌کنند.

[15] Antibody
[16] antigen

جرم مولکولی (M)[۱۷]: جرم یک مولکول از یک ماده است. وزن مولکول عمدتاً به دالتون[۱۸] بیان می‌شود. دالتون (Da) به‌عنوان یک دوازدهم جرم کربن-۱۲ (ایزوتوپ کربن) تعریف می‌شود و به‌طور تقریبی برابر است با: $1.66*10^{-24}g$

نقطه ایزو الکتریک[۱۹]: نقطه ایزو الکتریک یک پروتئین، میزان PH یک حلال است که در آن نقطه پروتئین ازلحاظ بار الکتریکی در حالت خنثی است. PH، یا سطح هیدروژن، میزان اسیدی بودن یک حلال را اندازه‌گیری می‌کند. برای حلال‌های خنثی PH برابر هفت است.

اثرانگشت جرم-پپتاید[۲۰]: مجموع جرم مولکولی پپتیدهایی که با تجزیه آنزیماتیک پروتئین تولیدشده است. جرم پپتیدها به‌وسیله طیف جرمی مشخص می‌شود. درنتیجه اثرانگشت جرم-پپتاید برای شناسایی پروتئین‌ها با مقایسه این اثرانگشت با جرم پپت ایده‌ای حاصل از تجزیه توالی‌های پروتئین محاسبه‌شده در حالت تئوریک[۲۱] در بانک داده‌ها است.

طیف جرمی[۲۲]: طیف جرمی برداری است که شدت بار به جرم یک مورد یا تحت یک شرایط خاص را نمایش می‌دهد.

توصیف پروتئین‌های بیان‌شده[۲۳]: یک بردار است که بیانگر شدت بار به جرم یک مورد خاص در میان موارد و شرایط متفاوت است.

[17] Molecular mass
[18] dalton
[19] Isoelectric point
[20] Peptide-mass fingerprint
[21] In silico approach
[22] Mass spectrometry
[23] Protein expression profile

۱,۷ اختصارات

جدول ۱-۱ اختصارات

اختصار	واژه کامل
DNA	**Deoxyribonucleic acid**
RNA	**Ribonucleic acid**
MS	**mass spectrometry**
WHO	**world health organization**
NCBI	**National Center for Biotechnology Information**
DC	**decision tree**
SVM	**support vector machine**
ANN	**artificial neural network**
CV	**cross-validation**
RS	**random sampling**
PPI	**protein-protein interaction**
PPT	**protein post translation**
CNS	**central neural system**
MS-TOF	**mass spectrometry Time Of Flight**
SELDI-MS	**surface enhanced laser mass spectrometry**
MALDI-MS	**matrix assisted laser ionization mass spectrometry**
ESI-MS	**electrospray ionization mass spectrometry**

۱,۸ آرایش کلی گزارش

این تحقیق از چهارفصل تشکیل‌شده است که در فصل اول به بیان کلیات و مفاهیم و همچنین تشریح مسئله خواهیم پرداخت، ضمناً در پایان هر فصل یک جمع‌بندی و خلاصه‌ای از مطالب بیان‌شده آن فصل خواهیم داشت. در فصل دوم علم پروتئومیکس و کاربردهای علوم کامپیوتری در آنالیز داده‌های حاصل از طیف‌سنج جرمی را بررسی خواهیم کرد در فصل سوم به داده‌کاوی و مرور ادبیات داده‌کاوی در پروتئوم جهت دسته‌بندی و کشف زیست‌نشانگرهای سرطانی خواهیم داشت. در فصل آخر خلاصه‌ای از فصل‌های گذشته را بیان خواهیم کرد و فرصت‌ها و چالش‌های داده‌کاوی پروتئوم جهت تشخیص سرطان را بیان کرده و سپس داده‌ها در دسترس و منابع دسترس‌پذیر را معرفی خواهیم کرد.

۱,۹ خلاصه فصل

انسان تقریباً متشکل از چهار ابر مولکول کربوهیدرات‌ها، پروتئین‌ها، اسیدهای نوکلئوتید و لیپیدها شده است که در این میان پروتئین‌ها به دلیل فراوانی، رفتار پویا، تغییرات پس از ترجمه و مطابق نبودن میزان بیان آن‌ها با mRNA به دیگر ابرمولکول‌ها برای شناخت بیماری‌های پیچیده مانند سرطان برتری دارد. سرطان بیماری است که از تقسیم غیرعادی سلولی ایجاد می‌شود که به دلیل نبود پروتئین خاص، فعالیت زیاد یک پروتئین یا کم فعالیتی یک پروتئین ایجاد می‌شود. سرطان دارای انواع مختلفی از سارکوما، خون، لنفاوی و ... که سالانه جان میلیون‌ها انسان را می‌گیرد و مسئله سرطان می‌تواند موجب ضعیف شدن کشورها و حتی ورشکستی آن‌ها نیز شود. هرگونه تغییر بالینی در بدن ما باعث تغییر در میزان، نوع و فعالیت‌های پروتئین‌های موجود در بافت و یا ارگان‌های ما خواهد داشت که نتیجتاً سرطان نیز این تغییر را در الگوی پروتئین‌های بدن ما می‌گذارد. به پروتئین‌های بیان‌شده در بافت یا ارگان ما در یک‌لحظه خاص پروتئوم گفته می‌شود و علم بررسی پروتئوم را پروتئومیکس می‌گویند. داده‌های حاصل از پروتئوم بسیار پیچیده بوده و ازنظر ویژگی بسیار زیاد است بنابراین برای کشف تغییرات الگوی پروتئین‌ها که منجر به شناسایی زیست‌نشانگرهایی مانند انواع پروتئین که در بیماری‌های مثل سرطان نقش کلیدی دارند و همچنین برای دسته‌بندی بیماری‌های همچون سرطان نیاز به تحلیل‌های سطح بالایی همچون داده‌کاوی و علوم هوش مصنوعی احساس می‌شود. با کمک نتایج حاصل از این تحلیل‌ها می‌توان شناخت بهتری نسبت به سرطان پیدا کرد و درنتیجه داروها و روش‌های درمانی مناسبی را جهت مقابله با این بیماری ساخت.

فصل دوم

پروتئومیکس و کاربردهای علوم کامپیوتری در داده های حاصل از طیف سنج جرمی

۲,۱ مقدمه

در فصل قبل تعریفی از پروتئوم و علم پروتئومیکس ارائه کردیم. هدف از این فصل پاسخ به پرسش‌هایی زیر است:

- روش‌های پروتئومیکس برای تفسیر و توضیح بیماری‌ها چیست؟
- کدام نمونه مناسب است؟
- پروتئین‌های در نمونه چگونه تفکیک، کمیت شماری و توصیف می‌شوند؟
- و رویکردهای برخورد با داده‌های حاصل از طیف‌سنج جرمی چیست؟

۲,۲ انواع پروتئومیکس

تحقیقات پروتئومیکس شامل نگرش‌های متنوعی به مسائل می‌شود. طیف این نگرش از اندازه‌گیری فراوانی و مقایسه پروتئین در دو نمونه آزمایش و کنترلی تا بررسی برهم‌کنش پروتئین در یک شبکه است. در این بخش به معرفی این رویکردها و هدف آن‌ها می‌پردازیم. لازم است که مجدداً تأکید کنیم در این پژوهش تمرکز بر روی روش‌های پروتئومیکس شناختی و مقایسه‌ای با استفاده از طیف‌سنج جرمی که همان روش توصیف پروتئین‌هاست می‌پردازیم.

۲,۲,۱ پروتئومیکس شناختی و مقایسه‌ای

هدف از پروتئین کمی اندازه‌گیری مطلق و یا مقایسه‌ای فراوانی پروتئین‌ها است. مزیت رویکرد مقایسه‌ای استفاده در مطالعات کمی برای اندازه‌گیری تغییرات پروتئین‌های بیان‌شده در یک بافت یا سلول است. رویکرد متداول مقایسه دودسته بیمار و سالم، یا سلول‌های تحت معالجه با یک دارو و سلول‌های دیگر و یا موارد دیگر که به دودسته کنترل و غیر کنترل تقسیم می‌شوند. پروتئین‌های با کاهش یا افزایش فراوانی در سلول شناسایی و کمی شماری شده تا تغییرات در بیان این پروتئین به من منظور کنترل و پایش بیماری اندازه‌گیری شود. این مورد بسیار پراهمیت است که بیماری‌هایی از قبیل سرطان، دیابت و بیماری‌های قلبی دارای عوامل متعدد ژنتیکی بوده و مکانیسم‌های و کنش و واکنش‌های پیچیده‌ای را دارا می‌باشند. (Elo and Schwikowski, 2012) شکل ۲-۱ شمایی از جریان کاری فرآیند شناسایی زیست‌نشانگرها را نمایش می دهد.

۲,۲,۲ تغییرات پس از ترجمه پروتئین‌ها

پروتئین‌ها در پاسخ به پیام‌های داخل سلول و خارج از سلول دچار تغییرات پس از ترجمه می‌شوند. این تغییرات پس از ترجمه یا اختصاراً PTM تنظیم‌کننده مکانیزم‌های فرآیندهای سلولی است و شناسایی این تغییرات پیش‌نیاز درک عملکرد پروتئین‌ها در سلول است. تعدادی مختلفی از PTM شناسایی‌شده است که شامل اکستیلاسیون، متیلاسیون، فسفوریلاسیون و گلیکوزیله شدن است. در میان این‌ها گسترده‌ترین مطالعات بر فسفوریلاسیون صورت گرفته است به این دلیل که تقریباً در تمامی فرآیندهای سلولی این تغییرات پس از ترجمه اتفاق می‌افتد. بنابراین تغییرات پس از ترجمه در *پروتئین‌ها می‌توانند* به‌عنوان عاملی برای بررسی و شناسایی بیماری مورداستفاده قرار گیرد. هدف شناسایی پروتئین‌هایی که دچار تغییرات پس از ترجمه‌شده، شناسایی مکان این تغییرات در سلول، اندازه‌گیری فراوانی آن‌ها و تعیین میزان ارتباط چندین PTM با یکدیگر است. تعداد نرم‌افزارها و ابزارها برای استفاده به‌طور فزاینده‌ای در چند سال گذشته افزایش پیداکرده است. به‌طور مثال PTMScout (http://ptmscout.mit.edu) وب‌سایتی برای مشاهده، بررسی و آنالیز داده‌های پرتوان حاصل از PTM که شامل بانک اطلاعاتی از آزمایش‌ها PTM تا با یکدیگر با ابزار مقایسه‌ای مجموعه داده است. به‌عنوان مثالی دیگر ابزار scan-x است که هدف آن پیش‌بینی مکان‌های فسفوریلاسیون با استفاده از طیف وسیعی از داده‌های عمومی فسفوپروتئومیکس است. مرجع (Gallego and Virshup, 2007) نگاهی مروری به این رویکرد داشته است.

۲,۲,۳ مکان‌یابی پروتئین‌ها[1]

پروتئین‌ها باید در مکان‌های به خصوصی از سلول حضورداشته باشند تا بتواند عملکرد مناسب داشته باشند و عدم قرارگیری در جای نامناسب می‌تواند موجب عملکرد نادرست سلول شود. همچنان که پروتئین‌هایی که موجب یک بیماری می‌شوند تمایل دارند تا در زیر سلول‌های همان بخش ظاهر شوند. با این تفاسیر نقش برداری از مکان پروتئین‌ها در زیر سلول‌ها می‌تواند منجر به شناسایی زیست‌نشانگرهایی نوینی برای شناسایی بیماری‌ها و ساخت و توسعه داروهای جدید باشد. برای مکان‌یابی از ترکیب روش‌های کامپیوتری و آنالیزهای پرتوان مانند طیف‌سنج جرمی استفاده می‌شود.(Elo and Schwikowski, 2012)

۲,۲,۴ برهم‌کنش پروتئین‌ها[2]

پروتئین‌ها به‌ندرت به‌تنهایی فعالیت می‌کنند و معمولاً در شبکه از ارتباطات پیوسته در فرآیندهای سلولی شرکت می‌کنند و هرگونه تغییر و اختلال در این ارتباطات پروتئین که اختصاراً PPI[3] گفته می‌شود می‌تواند زمینه‌ساز بیماری‌ها همچون سرطان شود. با شناخت این تعاملات پویای پروتئینی تشخیص بیماری تا و ساخت داروهای درمانی تسهیل می‌شود ازاین‌رو کاربردهای شبکه‌های پروتئینی درزمینه شناخت عوامل ژنتیکی بیماری‌ها، شناخت ویژگی‌های شبکه‌های پروتئینی و ارتباط زیر شبکه‌های پروتئینی به بیماری‌ها رو به گسترده شدن است. همچنین از شبکه‌های پروتئین‌ها برای دسته‌بندی بیماری تا و به‌خصوص سرطان استفاده می‌شود. مرجع (Wang et al., 2015) مروری بر شناسایی نشانگرهای سرطانی با رویکرد شبکه در تعاملات زیست‌نشانگرها در سلول انجام داده است.

---

[1] Protein localization
[2] Protein interaction
[3] Protein-protein interaction

شکل ۲۱-۲جریان کاری مراحل شناسایی زیست نشانگرها (Elo and Schwikowski, 2012)

۲,۳ روش‌های نمونه‌گیری برای کشف نشانگرهای سرطانی

طراحی مناسب آزمایش امری حیاتی برای موفقیت یک مطالعه است. با طراحی ضعیف یک آزمایش امکان تعیین این‌که آیا یک مشاهده به‌درستی تغییرات زیستی را نمایش می‌دهد یا این تغییرات تنها مربوط به مسائل فنی است، وجود ندارد. هزینه‌های بالا آزمایش‌ها پروتئومیکی معمولاً منجر به طراحی ضعیف آزمایش‌ها می‌شود و درنتیجه منجر به نتایج ضعیف یا اشتباه در مطالعه می‌شود.(Elo and Schwikowski, 2012)

زیست‌نشانگرها که مولکول‌های زیستی (ابر مولکول‌های ذکرشده در فصل اول) بوده و اطلاعات بسیاری را در رابطه با بیمارها دارا می‌باشند و می‌توانند برای تشخیص، پیش‌بینی، ارزیابی ریسک بیماری‌ها مورداستفاده قرار بگیرند، میزان و نوع آن‌ها در روش‌های نمونه‌گیری مختلف، متفاوت است درنتیجه تشخیص و انتخاب درست این نمونه‌ها در نتایج آزمایش بسیار تأثیرگذار است و نظر به این‌که این امر موردی حیاتی و اثرگذار در مطالعات پروتئومیکس است در این بخش نگاهی هرچند گذار اما تحلیلی و انتقادی به این روش‌ها خواهیم داشت.

۲,۳,۱ مایعات زیست پذیر[۴]

سروم (مایعی است که پس از حذف پروتئین‌های عامل لختگی خون از پلاسما باقی می‌ماند) و پلاسما (بخش مایع خون که در آن سلول‌های خونی شناور است) رایج‌ترین نمونه‌های بیمارستانی برای شناخت و اثبات زیست‌نشانگرها است به دلیل دسترسی و شناخته‌شده بودن در مراکز درمانی است. مایعات زیست پذیر دیگر شامل اور CSF[۵]، بزاق، مایع آسیت و ... هم به‌سرعت در حال شناخته شدن به‌عنوان منبعی برای یافتن زیست‌نشانگرها است. بیشتر نمونه‌های پیش‌آمده بیشتر برای مطالعات ابتدایی کشف زیست‌نشانگرها مناسب است و در عمل برای استفاده نهایی در مراکز درمانی به دلیل دشواری در نمونه‌گیری و هزینه‌بر بودن قابل‌استفاده نیستند. در مقابل، اوره به دلیل سهولت در دستیابی و شیوه غیرتهاجمی بودن (عدم آسیب به بدن برای به دست آوردن نمونه) و سرعت در پردازش آن بسیار مناسب است. پروتئوم اوره منبع غنی از زیست‌نشانگرهای بیماری برای آنالیز است.

مایعات زیست پذیر مجموعه‌ای پیچیده از پروتئین‌ها است که از تومورها و بافت‌های سالم ناشی می‌شود؛ بنابراین، ممکن است موجب عدم دقت در بیان پروتئین‌های مقایسه‌ای بین دو گروه شود. مقادیر زیاد پروتئین‌ها در مایعات زیست پذیر (از قبیل آلبومین و هموگلوبین) ممکن است نقش پروتئین‌های با میزان کم اما دخیلی در سرطان را کمرنگ کند. برای غلبه به این مشکل نیاز به فن‌های متنوع آزمایشگاهی است.

۲,۳,۲ بافت[۶]

مقایسه بیان پروتئین‌های شناخته‌شده بین بافت تومور و بافت سالم مجاور می‌تواند اطلاعات ارزشمندی را برای تشخیص بسیاری از بیماری به‌خصوص سرطان فراهم کند. چندین مزیت در استفاده از بافت برای پیدا کردن زیست‌نشانگرها وجود دارد که شامل این حقیقت می‌شود که زیست‌نشانگرهای کشف‌شده به‌طور واضح منشأ تومور می‌باشند. به‌هیچ‌عنوان نشانگرهای تومور در سلول‌های تومور پنهان نمی‌شوند؛ بنابراین، تمرکز بر روی بافت محلی تومور در قیاس با مایعات زیست پذیر بسیار بیشتر است و درنتیجه شناسایی نشانگرهای منتخب ساده‌تر است. اغلب با استفاده از فن‌های LCM و تجزیه‌وتحلیل پروتئوم می‌توان منبع دقیق نشانگرهای پروتئینی داخل تومور را کشف کرد. بااین‌وجود مطالعات پروتئوم به کمک نمونه‌ها بافتی دارای ایراداتی نیز است. اول اینکه روشی تهاجمی است، دسترسی به برخی از تومورها و بافت‌ها محدود است. به‌علاوه، بافت تومور ترکیبی ناهمگون از سلول‌های بدخیم همچنان که از بافت‌های همبند، بافت چربی و سلول‌های التهابی است که نمایانگر چالش‌های فنی برای دریافت و آنالیز نمونه و کمبود دقت در شناسایی نشانگرهای بالقوه است.

۲,۳,۳ سل لاین[۷]

برخلاف بافت، سل‌لاین‌های تومور می‌تواند بیان‌کننده متقارن جمعیت سلول‌ها باشد که می‌شود به‌سادگی در آزمایشگاه کشت کرد، می‌تواند در مقدار بسیار زیاد در دسترس باشد و می‌تواند برای طیف وسیعی از مطالعات کشف نشانگرها مورداستفاده قرار بگیرد. همچنین آماده‌سازی بخش‌های زیر سلولی شامل غشای پلاسما، سیتوزولی یا بخش هسته سلول، ترشحات، اگزوموز از طریق کشت برای مطالعات پروتئومیکسی نسبتاً ساده است. سلول‌های تومور ممکن است در محیط آزمایشگاهی به‌منظور شبیه‌سازی شرایط آسیب‌دیدگی ژن‌ها، بیان زیاد ژن‌ها و پاسخ آن‌ها به روش‌های درمانی قبل از مطالعات پروتئومیکسی استفاده شود. بااین‌وجود

[4] biofluids
[5] Central nervous system
[6] Tissue
[7] Cell line

استفاده از سل‌لاین‌ها دارای چندین محدودیت است ازجمله کمبود دسترسی به سل‌لاین‌های مناسب تومورها و بافت‌های مخاطی به‌عنوان گروه کنترل در مطالعات، تغییرات در ژن یا بیان پروتئین‌های شناسایی‌شده که ممکن است در شرایط کشت دوبعدی اتفاق بیافتد و نبود تعامل بین سلول‌های تومور با سلول‌های استرومال و یا ایمنی که شرایط محیط تومور در بدن را مشخص می‌کند.(Yang et al., 2015, Panis, 2015)

۲,۴ فن‌آوری‌های آنالیز پروتئومیکس

قدم کلیدی در پروتئومیکس جداسازی و نمایش مجموعه‌ای از پروتئین‌های موجود در یک نمونه است. آنالیز پروتئومیکس شامل فن‌آوری‌های متنوعی است که هرکدام دارای مزایا و معایب می‌باشند. در این قسمت نگاهی به این روش‌ها خواهیم داشت.

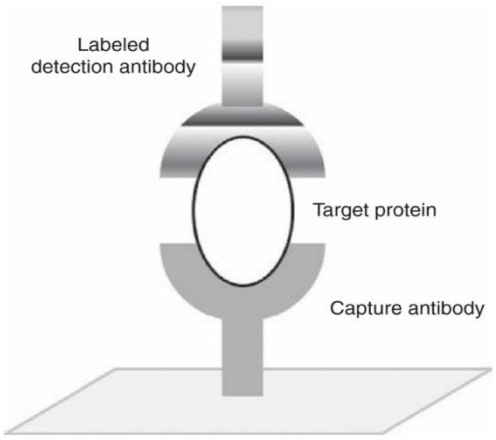
۲,۴,۱ فنّاوری ریزآرایه‌های پروتئینی

ریزآرایه‌های پروتئینی می‌تواند به دودسته عمده: ریزآرایه‌های فاز مستقیم و ریزآرایه‌های فازمعکوس تقسیم شود. در ریزآرایه‌ها فاز مستقیم مولکول‌ها با یک الگوی ماتریسی در یک سطح کوچک بی‌حرکت قرار می‌گیرند. این مولکول‌های بی‌حرکت با آنالیت‌ها (نمونه‌ای مورد ارزیابی) مورد آزمایش جهت تعیین حضور یا فراوانی پروتئین‌ها موردنظر واکنش می‌دهند. این مولکول‌ها می‌تواند شامل آنتی‌بادی‌ها، پروتئین‌ها، تکه‌هایی از پروتئین‌ها، آنزیم‌ها و یا پپتیدها باشند که با توجه به هرکدام از آن‌ها به‌عنوان مولکول گیرنده استفاده شود فن‌ها و کاربردها در آنالیز متفاوت است. به‌طور مثال ریزآرایه‌های آنتی‌بادی شامل دو فنّاوری ساندویچ ریزآرایه و فنّاوری تک-آنتی‌بادی ریزآرایه است. درروش ساندویچ ریزآرایه، دو آنتی‌بادی تعیین‌کننده پروتئین برای پروتئین هدف مورداستفاده قرار می‌گیرد. آنتی‌بادی گیرنده بر روی سطح برای گیر انداختن پروتئین هدف در روی سطح بی‌حرکت قرار می‌گیرد. سپس آنتی‌بادی دوم که با فلورسنت برچسب‌گذاری شده به پروتئین هدف که توسط آنتی‌بادی اول که پروتئین را گیر انداخته است می‌چسبد. (شکل ۲-۲)

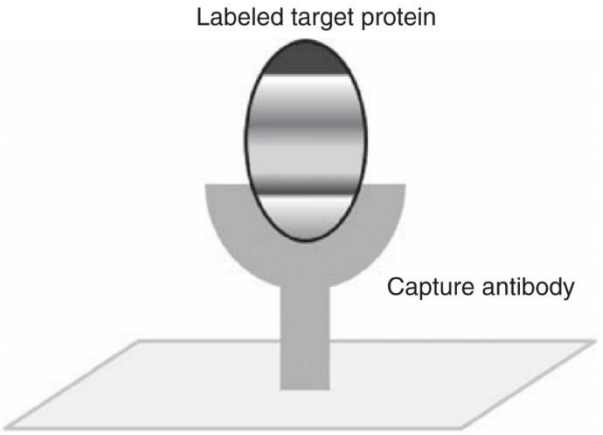
درروش آرایه‌های تک-آنتی‌بادی، نمونه مورد آزمایش جهت شناسایی مستقیم پروتئین‌ها توسط آنتی‌بادی‌های بی‌حرکت بر روی سطح برچسب‌گذاری می‌شوند.(شکل ۲-۳)

در ریزآرایه‌های فازمعکوس، آنالیت‌ها در یک سطح بی‌حرکت باقی می‌مانند. این مورد امکان ارزیابی هم‌زمان مقادیر کوچکی از بافت از نمونه‌های زیاد و متنوع را می‌دهد. بسته به طراحی ریزآرایه‌های پروتئین، چه آنالیت‌ها چه مولکول‌های گیرنده، به‌منظور شناسایی پروتئین‌ها برچسب‌گذاری می‌شوند. از روش‌های رایج برچسب‌گذاری شامل فلوئور سنت، نورتابی شیمیایی[8] و رادیواکتیویته هست که روش فلورسنت پرطرفدارترین آن‌ها است. (Dziuda, 2010)

[8] chemiluminescence



شکل ۲۲-۲روش ساندویچ ریز آرایه(Dziuda, 2010)



شکل ۳-۲روش آرایه‌های-تک‌آنتی‌بادی(Dziuda, 2010)

۲,۴,۲ ژل الکتروفوز دوبعدی پلی آکریل آمید (2D-PAGE)

ژل‌الکترفوز دوبعدی پروتئین‌ها را در بعد اول بر اساس بار آن‌ها و در بعد دوم بر اساس جرم آن‌ها جدا می‌کند. 2D-PAGE یک روش کمی پرتوان است که اجازه می‌دهد بیش از ۱۰ هزار پروتئین را با وضوح‌بالا تفکیک کرد. یکی از ایرادات این روش آن است که پروتئین‌های با مقادیر زیاد موجب نامفهوم سازی و پنهان ماندن پروتئین‌های با مقادیر کم شود. روش 2D-DIGE(two-dimensional in-gel electrophoresis) یکی از انواع روش ژل‌الکتروفوز دوبعدی پلی آکریل آمید است. در این روش پروتئین‌ها دو نمونه متفاوت با کمک رنگ فلورسنت برچسب‌گذاری می‌شوند و سپس الکتروفوز می‌شوند.(Dubitzky et al., 2007)

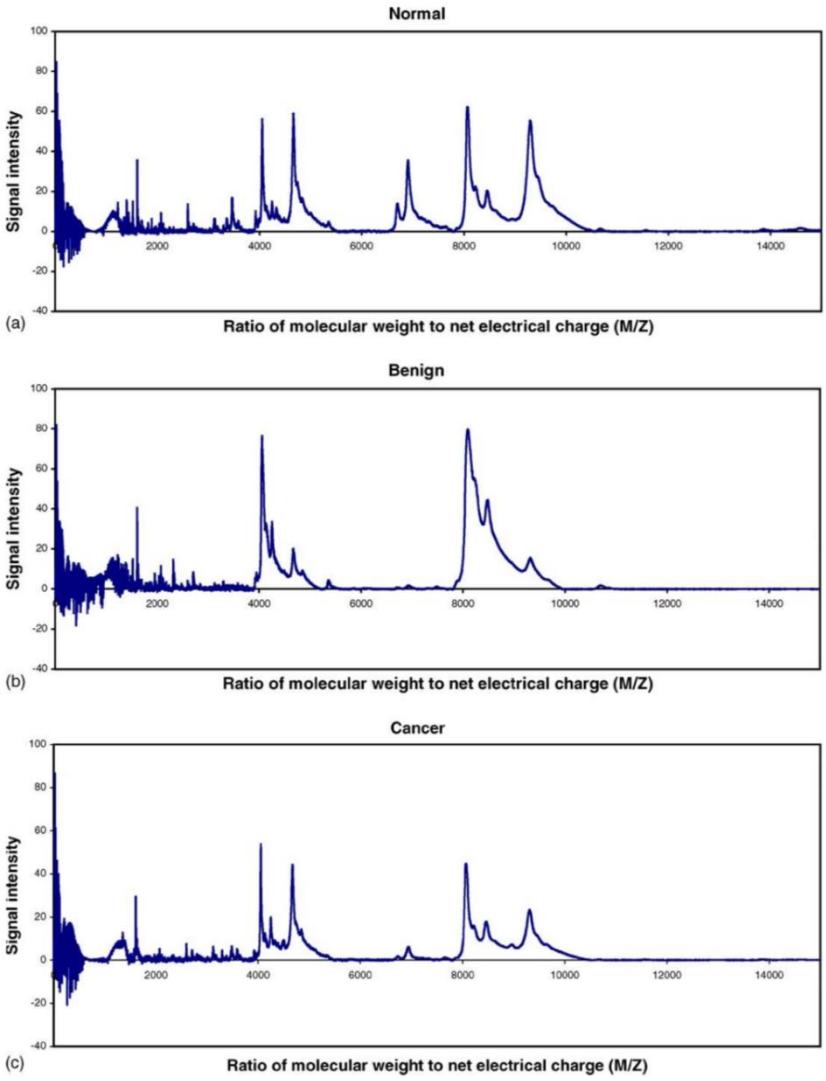
۲,۴,۳ طیف‌سنج جرمی[9]

طیف‌سنج جرمی یا به‌اختصار MS نقشی اساسی در شناسایی پروتئین‌ها و فرآیند پس از ترجمه آن‌ها بازی می‌کند. طیف‌سنج جرمی از سه بخش اساسی تشکیل‌شده است: ۱) منبع یونی که پروتئین‌ها را به گازهای یونی تبدیل می‌کند. ۲) ارزیاب جرمی که نسبت جرم به بار (m/z) یون‌ها را اندازه‌گیری می‌کند. ۳) و یک آشکارساز که تعداد یون‌های مشاهده‌شده برای یک مقدار مشخص m/z را شمارش می‌نماید. به‌طور عمده دو منبع یون‌ساز electrospray ionization(ESI) و matrix-assisted laser desorption/ionization(MALDI) موجود هست. همچنان time-of-flight(TOF) به‌عنوان ابزار آنالیز در MALDI به‌کاربرده می‌شود. به‌طور مختصر نمونه‌های پروتئینی با مولکول‌های با چینش ماتریسی مخلوط شده و سپس به لکه‌های بلوره‌ای شده تبدیل می‌شوند و بر روی صفحات فلزی قرار می‌گیرند. سپس تابش‌های پرتاب‌شده از لیزر موجب پراکنده شدن و یونیزاسیون مخلوط می‌شود. پروتئین‌های یونیزه از اتاق یونی عبور می‌کنند و به آشکارساز برخورد می‌کنند. بنا بر ولتاژ بکار رفته و شتاب یون‌ها، میزان بار به جرم (m/z) هرکدام از یون‌ها را می‌توان اندازه گرفت و سپس نمایش داد. Surface-enhanced laser desorption time-of-flight(SELDI-TOF) گونه جدیدی از MALDI-TOF است. عنصر کلیدی در SELDI-TOF MS تراشه‌های پروتئینی با سطح شیمیائی برای به دام انداختن دسته‌ای از پروتئین‌ها در شرایط خاص است. MALDI MS و SELDI-

[9] Mass Spectrometry

TOF MS فنّاوری‌های با حساسیت بالا هستند و برای شناسایی و تشخیص پروتئین‌ها در حجم بالا بسیار مناسب می‌باشند.(Dubitzky et al., 2007)

در پروتئومیکس رایج‌ترین روش برای شناسایی پروتئین‌ها با به‌کارگیری MS، رویکرد پائین به بالا است. در این رویکرد مولکول‌های مورد ارزیابی پپتیدها هستند که با فرآیند تجزیه آنزیماتیک پروتئین‌ها نمونه حاصل‌شده‌اند که مراحل آن را شرح دادیم. طیف ایجادشده از تکه‌های پپتیدها به tandem MS spectra شناخته می‌شود که نتایج حاصل از آن بصری سازی می‌شود که موجک‌های موجود در تصویر حاصل نشان‌دهنده آمینواسیدهای حاضر در پپتیدها است؛ اما در این رویکرد نتایج حاصل نمایش‌دهنده پپتیدها است و نیاز به مرحله دیگری برای شناسایی پپتیدها برای پیش‌بینی پروتئین‌ها موجود در نمونه است. این فرآیند می‌تواند با استفاده از نرم‌افزارهای جستجوی توالی در پایگاه داده‌ها مانند mascot استفاده شود که این روش در بخش‌های بعدی بررسی خواهد شد؛ اما رویکرد بعدی، از بالا به پائین است که MS مستقیماً برای ارزیابی پروتئین‌هایی سالم و تجزیه نشده استفاده می‌شود؛ اما این رویکرد نیاز به تجهیزات بیشتر و پیچیده‌تر و درنتیجه پرهزینه‌تر برای بکار گیری است که همین عوامل باعث عدم استقبال از این رویکرد است.(Swan et al., 2013) شکل ۲-۴ نمونه‌ای از طیف جرمی بصری سازی شده را نمایش می‌دهد.



شکل ۴-۲ طیف جرمی حاصل از سروم خون با فناوری SELDI

این طیف نمایش دهنده سه دسته سرطان خوشخیم، سرطان بدخیم و سالم را نشان می دهد.(Li et al., 2004)

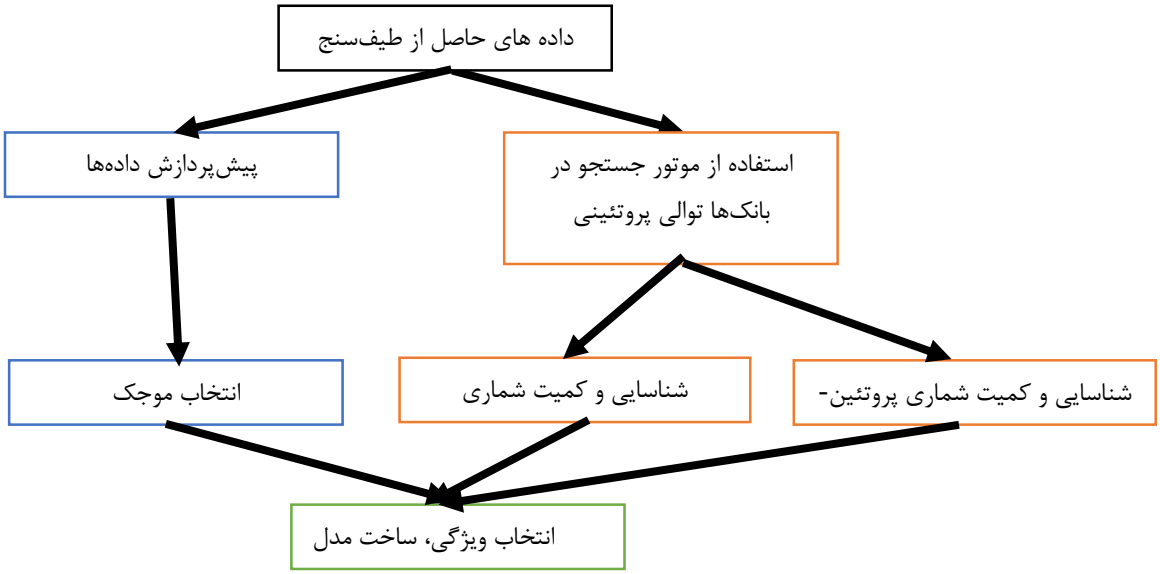
۲,۵ فرصت‌ها و چالش‌های آنالیز داده‌ها در طیف‌سنج جرمی

استفاده از طیف‌سنج جرمی فرصت و روزنه‌های پیشرفت بسیاری را گشوده است ولی بااین‌وجود دچار کاستی‌هایی نیز است. محور اساسی در شناسایی پروتئین‌ها با استفاده از طیف‌سنج جرمی، کشف زیست‌نشانگرهای بیماری‌ها است. این زیست‌نشانگرها می‌توانند برای اهداف مختلفی مانند شناسایی فرآیندهای بیولوژیکی، شناسایی و پیش‌بینی بیماری‌ها و میزان پیشرفت بیماری‌های شناخته‌شده در بدن استفاده شود. همچنان که برای بررسی تأثیرگذاری یک داروی خاص و توسعه و ساخت داروهای جدید بکار

می‌رود. فرصت دیگر حاصل از طیف‌سنج جرمی، نگرش به پروتئین‌ها از دیدگاه شبکه و مسیرهای ارتباطی آن‌ها است؛ اما با تمام این فرصت‌ها دارای محدودیت‌هایی نیز است. به‌طور مثال برای نمونه‌گیری‌های متعدد به دلیل محدودیت‌های زمانی و همچنین هزینه زیاد مناسب نیست. همچنین پیدا کردن پروتئین‌هایی با میزان کم در نمونه که در فرایند ایجاد بیماری مشارکت دارند دشوار است. مورد بعدی هم عدم قطعیت در شناسایی درست پروتئین‌ها بنا به‌توالی پپتیدی آن‌ها در بانک‌های اطلاعاتی است.

۲,۶ آنالیز داده‌های طیف‌سنج جرمی

داده‌های حاصل از طیف‌سنج جرمی مجموعه‌ای موجک‌ها برحسب بار به جرم است که برای شناسایی پپتیدها است. ضرورت استفاده از روش‌های علوم کامپیوتری برای شناسایی پروتئین‌هاا از موجک‌های شناسایی‌شده و مقایسه نمونه‌ها امری اجتناب‌ناپذیر است. بدین منظور دو رویکرد متفاوت جهت برخورد با موجک‌های حاصل از طیف‌سنج جرمی وجود دارد. رویکرد اول مستقیماً بر روی موجک‌ها پردازش و عملیات داده‌کاوی را انجام می‌دهد که به آن انتخاب موجک[10] گفته می‌شود و در رویکرد دوم استفاده از موتورهای جستجو در بانک‌های پروتئینی و توالی یابی شده جهت شناسایی و توصیف پروتئین‌ها است که این رویکرد نیز در پیوستگی استفاده از برچسب‌گذاری در حین آماده‌سازی نمونه‌ها برای آنالیز که تقریباً روشی آزمایشگاهی است استفاده می‌شود و یا از روش‌ها بدون برچسب‌گذاری و مبتنی بر روش‌های کامپیوتری جهت کمی شماری پروتئین‌ها استفاده می‌کند. شکل ۵-۲ جریان آنالیز داده‌های طیف‌سنج جرمی را به‌طور خلاصه بیان می‌کند.



شکل ۲۵-۲جریان آنالیز داده‌های حاصل از طیف سنج جرمی

[10] Peak picking

۲,۶,۱ انتخاب موجک

در انتخاب موجک داده‌های حاصل از طیف‌سنج جرمی فارغ از این‌که کدام پپتیدها و کدام پروتئین‌ها در نمونه حاضر هستند مورد پردازش قرار می‌گیرند در عوض موجک‌هایی با بیشترین سیگنال به‌عنوان زیست‌نشانگر کاندید می‌شوند. کاتاجاما و اریزیک دو ایراد این روش را برشمردند اول این‌که یک رویکرد مستقیم برای شناسایی پروتئین ارائه نمی‌کند و برای این منظور تحلیل‌های بیشتر نیاز است. دومین ایراد این است که پیش‌پردازش داده‌ها شامل نرمال‌سازی، همگام‌سازی موجک‌ها و کاهش نویز امری ضروری و اجتناب‌ناپذیر است.(Katajamaa and Orešič, 2005) درواقع بدون پیش‌پردازش داده‌ها امکان مقایسه موجک‌های کاندید بین نمونه‌ها وجود ندارد و دقت مدل‌های ارائه‌شده به‌شدت کاهش میابد. در جدول ۲-۱ نمونه‌ای از داده‌های حاصل از طیف‌سنج جرمی که برای پردازش توسط روش‌های انتخاب موجک فراهم‌شده‌اند نمایش داده‌شده است. به این جدول ماتریس شدت M/Z نیز گفته می‌شود.

۲,۶,۲ موتورهای جستجو

در طی فرآیند طیف‌سنج جرمی، جرم پپتیدهای شناسایی‌شده حاضر در نمونه مورد تحلیل قرار می‌گیرد. جرم این پپتیدها هم‌راستا با جرم تکه‌هایشان جهت این‌که کدام پپتیدها حضور دارند و به کدام پروتئین‌ها متعلق می‌باشند موردبررسی قرار می‌گیرد.

نرم‌افزارهای جستجو از قبیل Mascot برای بررسی احتمال حضور پروتئین‌ها توسعه داده‌شده است. این نرم‌افزار در ارتباط بانک توالی پروتئینی مانند UniportKB و NCBInr[11] و همچنان از بانک داده‌های خاص مانند sgn متعلق به TOMATO[12] و Tair متعلق به Arabidopsis[13] استفاده می‌کند. پپتیدهای حاضر در نمونه آنالیز شده با طیف‌سنج جرمی بررسی‌شده و با قیاس تناسب جرم به حجم پپتیدهای نمونه با پپتیدهای پروتئین‌های شناخته‌شده در بانک داده‌ای موردنظر شناسایی و پیش‌بینی می‌شود. نیلسون و همکاران معتقدند که این روش به‌طور کامل دقیق نیست به خاطر شباهت توالی برخی از پروتئین‌ها و نسبت کم پروتئین‌های توالی شده به‌کل پروتئین‌های موجود در بدن انسان که موجب عدم قطعیت و ضعف این روش می‌شود.(Neilson et al., 2011)

همان‌طور که ذکر کردیم شناسایی پروتئین با کمک موتورهای جستجو در پیوستگی استفاده از روش‌های برچسب‌گذاری و بدون برچسب‌گذاری است. روش‌های و رویکردهای برچسب‌گذاری متنوعی وجود دارد بنا به این‌که جز روش‌های آزمایشگاهی برای کمیت شماری پروتئین‌ها بوده و از موضوع بحث ما خارج است از بیان آن‌ها صرف‌نظر می‌کنیم؛ اما ضعف این روش شامل هزینه‌های بالای آزمایشگاهی، محدودیت در تعداد نمونه‌های که می‌توان آنالیز کرد و همچنین عدم سازگاری با برخی از نمونه‌های که برای تحلیل گرفته می‌شود.

[11] http://www.ncbi.nlm.nih.gov/
[12] http:// solgenomics.net
[13] http://www .arabidopsis.org

اما روش دیگر تحلیل داده‌های بدون اســتفاده از برچسب‌گذاری و روش‌های اضافی آزمایشگاهی است. دو روش بدون برچسب‌گذاری برای کمیت شماری پروتئین‌ها وجود دارد ۱) شدت سنجی سیگنال ۲) طیف شماری.

درشدت سنجی سیگنال از ناحیه زیر سطح منحنی (AUC) طیف‌ها موجک‌ها برای مقایسه مقدار پپتیدهای حاضر در نمونه استفاده می‌شـود. درروش دوم طیف مشاهده‌شده برای پپتیدهای یک پروتئین جمع زده می‌شود. مرجع (Wong and Cagney, 2010) مروری به این رویکردها داشته است. تعداد زیادی نرم‌افزار منبع باز و تجاری بدون برچسب شمارشگر موجود است. به‌عنوان مثالی از نرم‌افزارهـای رایگـان برای روش AUC می‌توان از MSInspect و MSQuant نـام برد و برای روش طیف شـــمـاری می‌توان از نرم‌افزارهای PepC و APEX نام برد که می‌تواند برای کمیت شـماری روش‌های بدون برچسـب‌گذاری استفاده شود سپس اعداد و کمیت‌های حاصل با استفاده از روش‌های کامپیوتری جهت شناسایی پروتئین‌ها مورد تحلیل قرار می‌گیرند. روش‌های کامپیوتری به دلیل عدم اسـتفاده از روش‌های پیچیده آزمایشـگاهی به‌سـادگی قابل‌اسـتفاده می‌باشند. درنتیجه انعطاف بیشتری در نمونه‌گیری و آماده‌سـازی نمونه‌گیری وجود خواهد داشـت. رویکردهای کمی مقادیر عددی با توجه به پروتئین شناسایی‌شده از طیف‌سنج جرمی تولید می‌کنند که این می‌تواند یک مزیت باشـــد برای زمانی که می‌خواهیم از روش‌های داده‌کاوی برای دســـته‌بندی بیماری‌های همچون سـرطان و یا شـناسـایی زیسـت‌نشانگرهای موثر در بیماری مثل سرطان با قیاس دودسته از نمونه‌های سالم و بیمار حاصل شـود، اسـتفاده کنیم. (Swan et al., 2013) جدولی از نرم‌افزارهای مورداستفاده در پروتئومیکس در پیوست ((الف)) آمده است. جدول ۲-۲ نمونه‌ای از پروتئین‌های شـناسـایی‌شـده را در یک ماتریس نشـان می‌دهد که به آن ماتریس بیان پروتئین[14] نیز گفته می‌شود.

جدول ۱-۲ماتریس شدت M/Z (Dziuda, 2010)

در اینجا ستون اول بیانگر جرم به بار یک پپتید خاص و ردیف های ستون اول بیانگر نمونه و سایر مقادیر بیانگر شدت سیگنال یک پپتید در یک نمونه است.

	Class 1					***	Class 2		***	Class J		
m/z	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	***	***	***	***	***	Sample $N-1$	Sample N
500.0223	25.9588	26.6526	24.4485	29.7677	26.1579						24.4203	25.9005
500.4307	28.6677	29.4781	26.9261	33.0075	29.8984						27.7164	29.0940
500.8573	33.8643	35.1272	31.9417	38.9422	34.3449						31.4630	33.0437
501.2751	38.0995	40.2232	35.2459	43.2378	36.7521						33.3736	35.1533
501.6931	39.6762	42.3087	35.7506	44.4585	38.0613						34.1921	36.1327
502.1113	39.7707	42.6244	35.3323	44.2172	38.0231						34.1357	36.2976
***												
***												
***												
15985.13	5.1260	4.9562	5.1565	4.4309	5.6946						4.5777	4.5680
15988.53	5.1059	4.9439	5.1494	4.4263	5.6690						4.5746	4.5636
15992.61	5.0613	4.9013	5.1072	4.4050	5.5761						4.5602	4.5554
15995.77	5.0525	4.8931	5.0944	4.4004	5.5560						4.5534	4.5514
15998.93	5.0297	4.8816	5.0752	4.3936	5.5237						4.5499	4.5437

[14] Expression protein matrix

جدول ۲-۲ماتریس بیان پروتئین

ستون اول بیان‌گر نام پروتئین‌ها، در سایر ستون‌ها ردیف اول نمایش‌گر نام نمونه و مقادیر جدول بیان پروتئین‌ها بر حسب M/Z

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	***	***	***	***	***	Sample $N-1$	Sample N
Variable 1	50.6083	45.8562	56.6637	66.0191	48.7441						20.3157	22.0191
Variable 2	81.3635	80.8688	86.9904	99.9598	80.8768						35.7969	48.9452
Variable 3	30.7451	21.9715	28.7226	25.6417	25.2696						16.3164	15.8974
Variable 4	25.2859	24.6415	32.1971	30.4539	23.9356						32.3733	65.2203
Variable 5	75.9739	99.6320	95.6099	73.0236	50.5500						86.1237	94.3348
Variable 6	31.3223	61.3896	40.8125	30.7833	25.0216						43.7728	47.9578
***												
***												
***												
Variable p-3	26.2785	36.7642	30.6385	30.8155	24.0845						40.8044	43.6642
Variable p-2	17.9427	1 9.6682	20.0322	20.0845	16.7964						34.6466	40.4123
Variable p-1	19.5918	20.5732	22.3328	22.2665	18.7692						19.7486	25.0679
Variable p	75.2334	90.1458	89.2247	75.6991	75.5829						87.5451	74.6747

۲,۷ خلاصه فصل

یکی از روش‌های پروتئومیکس شـامل شناسایی، توصیف و کمیت شماری پروتئین‌ها است که گاهی نیز به پروتئومیکس قیاسی نیز شـناخته می‌شود. روش دیگر بررسی تغییرات پس از ترجمه پروتئین‌ها در سلول است. روش دیگر مکان‌یابی پروتئین‌ها است که به بررسـی مکان پروتئین‌ها در سـلول می‌پردازد و در آخر برهم‌کنش پروتئین‌ها که به تعاملات پروتئین‌ها با یکدیگر در یک شبکه نگاه می‌کند. برای جمع‌آوری داده می‌توان از مایعات زیسـت پذیر، بافت‌ها و سـل لاین‌ها اسـتفاده کرد که مایعات زیست پذیر با توجه به گردش آن‌ها در بدن و به‌تبع آن دارا بودن اطلاعات بسـیار از بدن و راحتی جمع‌آوری نمونه روش مناسـب و پیشـنهادی بسیاری از تحقیقات اسـت. برای آنالیز این نمونه‌ها از فنّاوری‌های تراشه‌های پروتئینی، ژل الکتروفوز دوبعدی پلی آکریل آمید (2D-PAGE) و طیف‌سنج جرمی می‌توان استفاده کرد که هرکدام از این فنّاوری‌ها دارای انواع مختلف و زیرمجموعه‌های خاص خود هستند. در این میان طیف‌سـنج جرمی به دلیل توان بالایان در آنالیز انواع نمونه‌ها و توان بالا در آنالیز حجم بالایی از پروتئین‌ها و غیر جانب‌دارانه بودن یعنی در ابتدای آزمایش از پروتئین‌های بیان‌شـده خبری نداریم به‌عنوان فنّاوری مرسوم توسط کلینیک‌ها و محققین استفاده می‌شـود. در برخورد با داده‌های حاصـل از طیف‌سنج جرمی دو رویکرد وجود دارد. این‌که مستقیماً روی موجک‌های حاصل پردازش صـورت گیرد و یا این‌که ابتدا به شـناسایی پروتئین‌ها اقدام کرده و سپس به پردازش آن‌ها به پردازیم که این ایراد رویکرد اول برای کشف زیست‌نشانگرها نیاز به تحلیل‌های اضافی است و همچنین به‌کارگیری پیش‌پردازش داده‌ها امری اجتناب‌ناپذیر است.

فصل سوم

داده‌کاوی و کاربرد آن در دسته‌بندی و کشف زیست‌نشانگرهای سرطانی

۳,۱ مقدمه

فرآیند داده‌کاوی پروتئوم شـــامل پیش‌پردازش داده‌ها (بیشـــتر برای زمانی که مســـتقیم بر روی موجک ها پردازش و مدل‌ســـازی می‌کنیم)، انتخاب ویژگی و کاهش بعد (به دو منظور افزایش دقت دســـته‌بند و کشـــف زیســـت‌نشـــانگرها که موجب تمایز دودســـته می‌شوند)، ساخت مدل دسته‌بند، نمونه‌برداری و در آخر ارزیابی عملکرد مدل است. در این فصل شش مدل دسته‌بند درخت تصمیم (DC)[۱]، شـــبکه‌های عصـــبی مصـــنوعی (ANN)[۲]، جنگل تصـــادفی (RF)[۳]، مدل‌های مبتنی بر قاعده[۴]، بردارهای ماشـــین پشـــتیبان (SVM)[۵] و بیز ساده (NB)[۶] را موردبررسی قرار داده و در آخر مروری بر تحقیقات پیشین خواهیم داشت.

۳,۲ داده‌کاوی و کاربرد آن در پروتئومیکس

داده‌کاوی رویکردی اســت که در علم و زمینه‌های تجاری برای استخراج مفاهیم معنادار و قابل‌استفاده از مجموعه داده‌های بزرگ و پیچیده استفاده می‌شود. فرآیند داده‌کاوی معمولاً تکرارپذیر است و داده‌های ناشناخته و دارای اطلاعات بالقوه به کمک ابزار تحلیلی پیچیده مورد کاوش و کشف قرار می‌گیرد. فرآیند کشف شامل پیدا کردن روابط و الگوها در داده‌های خام است که بشود از آن برای تصـمیم‌گیری و آنالیزهای بیشـتر استفاده کرد. تکامل شیوه‌های داده‌کاوی از پیشرفت در هوش مصنوعی، آمار و مدریت انبار داده‌ها حاصـل‌شـده است.(Thomas et al., 2006) داده‌کاوی می‌تواند شامل گونه‌های نظارتی، یا همان مدل‌های دسته‌بندی[۷] باشد که در این‌گونه داده‌ها هرکدام دارای برچسـب هسـتند که دسـته آن داده را مشـخص می‌کند، داده‌ها به دو دسـته داده‌های آموزشی و تسـت تقسیم می‌شوند، با کمک داده‌های آموزشی مدل خود را می‌سازیم و سپس با داده‌های تست مدل خود را ارزیابی می‌کنیم در مقابل گونه غیر نظارتی که همان مدل‌های خوشه‌بندی[۸] است، داده‌ها فاقد برچسب هستند و بر اساس شباهت‌ها با کمک مدل‌های

---

[1] Decision tree
[2] Artificial neural network
[3] Random forest
[4] Rule-based classifiers
[5] Support vector machine
[6] Nave bayes
[7] classification
[8] clustering

خوشـه بند در یک گروه قرار می‌گیرند که به این گروه‌ها خوشـه گفته می‌شـود.(Swan et al., 2013) در این بخش ما بر روی گونه‌های نظارتی و به‌تبع مدل‌های دسـته‌بند تمرکز خواهیم کرد. داده‌کاوی با نظارت در پروتئومیکس توصیفی و مقایسه‌ای معمولاً با دو هدف دنبال می‌شـود ۱) دسته‌بندی سرطان تا (بدخیم و خوش‌خیم، سرطانی و سالم، سرطان نوع A و سرطان نوع B و ...) ۲) کشف زیست‌نشانگرهای سرطانی که از مقایسه دسته تا مثلاً دسته سالم و سرطان، سرطان خوش‌خیم و بدخیم و ... به دست می‌آید که لازمه حصـول به هدف دوم انتخاب ویژگی و کاهش بعد است. اگر ما در فرآیند داده‌کاوی مستقیماً به سراغ موجک‌ها برویم، پس از انتخاب ویژگی و سـاخت مدل نیاز اسـت تا بر روی موجک‌های منتخب تحلیل اضـافی جهت روشـن‌سـازی اینکه هرکدام از این موجک‌های منتخب مربوط به کدام پروتئین است، صورت گیرد. داده‌کاوی داده‌های پروتئینی به دلیل کم بودن نمونه تا(سطرها) و زیاد بود ویژگی تا(سـتون تا) بسیار چالش‌برانگیز است و نیاز به ابزارهای تحلیلی سطح بالا داده‌کاوی است. شکل شمایی از جریان مراحل داده‌کاوی بر داده‌های پروتئومی به‌منظور دسـته‌بندی سـرطان را نمایش می‌دهد. شکل ۳-۱ جریان کاری داده کاوی بر روی داده‌های حاصل از طیف سنج جرمی را زمانی که داده‌ها به صورت ماتریس بیان پروتئین باشند نمایش می‌دهد.

۳,۲,۱ پیش‌پردازش داده‌ها

داده‌های حاصـل از MS دارای نویز بوده و هرگونه ساخت مدل دسته‌بند قبل از پیش‌پردازش ممکن است نتایج گمراه‌کننده‌ای را به بار آورد. بیشـتر تحقیقات منتشرشده از نرم‌افزارهای که برای پیش‌پردازش داده‌ها توسعه داده شده‌اند استفاده کرده‌اند. به‌طور مثال نرم‌افزار ciphergen که برای پیش‌پردازش داده‌های حاصـلSELDI-MS اسـتفاده می‌شود. این نرم‌افزار مکان و شدت هرکدام از پروتئین‌های در نمونه را یافته و فعالیت‌های پیش‌پردازش شامل کاهش خط مبنا[1]، همگام‌سازی موجک تا نرمال‌سازی شدت موجک تا و شـناسـایی موجک تا را انجام می‌دهد. کاهش خط مبنا نویزهای شیمیایی و الکترونیکی را پاک می‌کند. معمولاً کاهش خط مبنا در دو مرحله انجام می‌شود. اول تخمین خط مبنا با روش‌های پارامتریک و غیر پارامتریک می‌شود و سپس داده‌های اصلی از پارامتر تخمین زده‌شـده کسـر می‌شـود. زمانی که کاهش خط مبنا تکمیل شـد نوبت بعدی پیش‌پردازش نرمال‌سـازی است. به دلیل اینکه موجک‌های یک طیف مقدار فراوانی نسـبی یک پروتئین را توصـیف می‌کنند برای مقایسـه معنی‌دار بین طیف‌های مختلف (منظور نمونه تا) نرمال‌سازی صورت می‌گیرد.

۳,۲,۲ نفرین بعد

مهم‌ترین چالش در داده‌کاوی پروتئوم مسـئله ابعاد داده‌هاست بدین معنی که تعداد نمونه‌ها کم و تعداد ویژگی‌ها (موجک‌ها) بسیار زیاد است و این درست برعکس داده‌های تجاری مانند خرده‌فروشی‌ها که تعداد ویژگی‌ها (ستون تا) کم ولی تعداد نمونه‌ها (سطرها) بسـیار زیاد است. یک رویکرد مستقیم می‌تواند این باشد که هرکدام از موجک تا را به‌عنوان یک ویژگی در نظر بگیریم اما متأسفانه تعداد ویژگی در داده‌های پروتئوم بسـیار زیاد بوده و می‌تواند بین ۳۰۰۰-۱۵۰۰۰ ویژگی باشـد. این مشـکل در داده‌کاوی به‌عنوان نفرین بعد یا مشـکل ابعاد مطرح اسـت. ابعاد بالا معمولاً بدین معناست که تعداد زیادی از ویژگی‌های بی‌استفاده و بی‌اثر باعث پنهان ماندن ویژگی‌های کلیدی یک مجموعه داده می‌شـوند. ابعاد بالا علاوه بر اشـغال حافظه و کاهش سـرعت پردازش می‌تواند دقت و درستی مدل دسته‌بند را هم تحت تأثیر قرار دهد.(Wang et al., 2017)

غربال طیف داده‌ها به‌منظور شـناسایی موجک‌ها، به‌عنوان استخراج ویژگی شناخته می‌شود. بر اساس این فرآیند هر گروهی از نقاط M/Z که داخل یک گروه قرار می‌گیرد به‌وسیله میانگین یا ماکزیمم داده‌های آن گروه توصیف می‌شود. درنتیجه این مقادیر نماینده

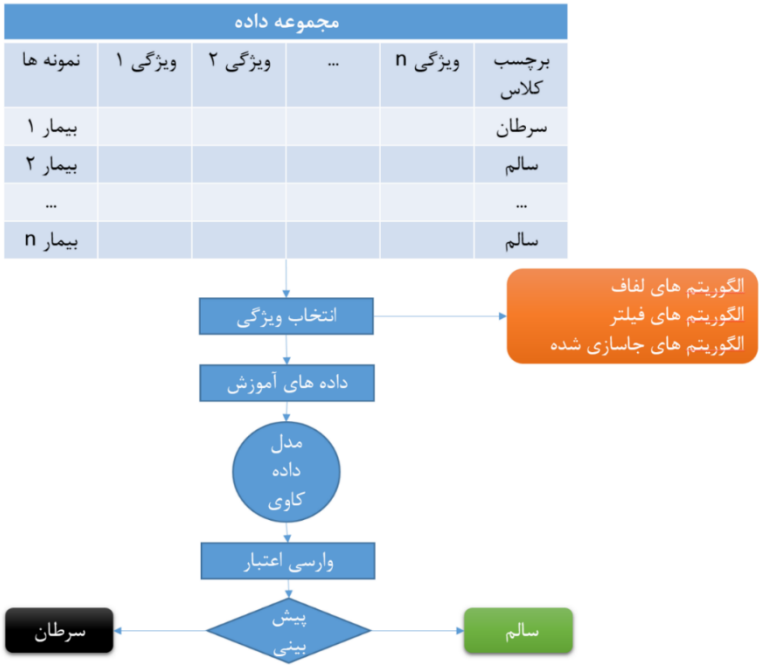
[1] Baseline reduction

به‌عنوان ویژگی (موجک) در نظر گرفته می‌شود. این گروه‌ها می‌توانند مستقل یا غیرمستقل، هم‌اندازه یا متغیر، پوشا یا غیر پوشا باشند. با تغییر اندازه گروه‌ها پژوهشگر می‌تواند فرآیند استخراج داده‌ها را بهینه کند.

اما در مقابل انتخاب و استفاده از ویژگی‌های به‌خصوص برای ساخت مدل موجب دقت و کارآمدی بیشتر دسته‌بند می‌شود. فرآیند داده‌کاوی با کاهش ویژگی سرعت و سهولت بیشتر پیدا می‌کنند و نتایج تفسیرپذیر تر می‌شوند. همچنین برای کشف زیست‌نشانگرها کاندید که موجب تمایز دو دسته از هم می‌شوند استفاده از انتخاب ویژگی را اجتناب‌ناپذیر می‌کند. پس پاک کردن ویژگی‌های غیر ضرور و بی‌استفاده امری ضروری است. البته باید به این نکته توجه داشت که کاهش ابعاد همیشه تضمین‌کننده نتایج موفقیت‌آمیز در انتخاب ویژگی نیست بنابراین لازم است تا متغیرهای انتخاب‌شده اعتبار سنجی شوند.

۳,۲,۳ روش‌های انتخاب ویژگی

به‌طور کلی روش‌های انتخاب ویژگی به سه دسته روش فیلتر[۲]، روش لفاف[۳] و روش جاسازی‌شده[۴] طبقه‌بندی می‌شود. روش‌های فیلتر مستقل از الگوریتم دسته‌بند بوده و به‌طور جداگانه داده‌ها رو موردبررسی قرار می‌دهد. روش‌های لفاف مدل دسته‌بند یادگیرنده بر مبنای زیرمجموعه‌ای از ویژگی‌های انتخاب‌شده مورد ارزیابی قرار می‌دهند. اگرچه این روش وابستگی بین متغیرها را بررسی می‌کند اما ریسک بیش برازش را افزایش داده و همچنین از لحاظ محاسباتی بسیار سنگین است. روش‌های جاسازی شده زیر مجموعه‌ای بهینه از ویژگی‌های انتخاب شده را با ساخت انتخاب ویژگی داخل الگوریتم ارزیابی می‌کند. اگرچه این روش در قیاس با روش‌های فیلتر کمی محاسبات بیشتری را نیاز دارد اما در عوض در کنش و واکنش مستقیم با مدل دسته‌بند می‌باشد.(Jagga and Gupta, 2015)



شکل ۱-۳جریان کاری داده کاوی بر روی داده های طیف جرمی (Jagga and Gupta, 2015)

[2] Filter method
[3] Wrapper method
[4] Embedded method

۳,۲,۴ مدل‌های دسته‌بندی

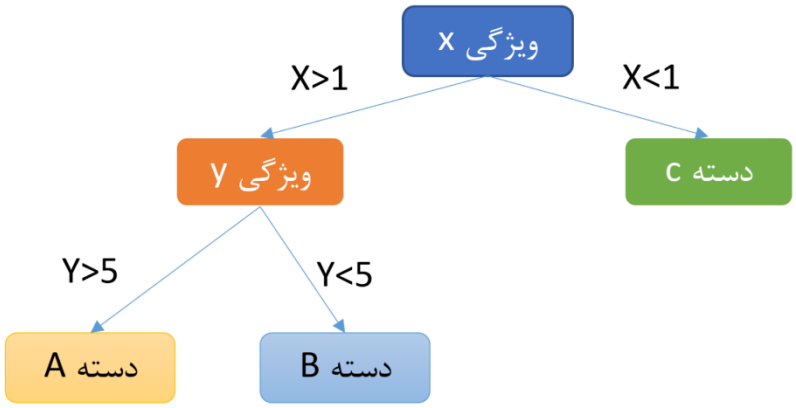
بیز ساده: بیز روشی برای دسته‌بندی پدیده‌ها، بر پایه احتمال وقوع یا عدم وقوع یک پدیده است و در نظریه احتمالات با اهمیت و پرکاربرد است. اگر برای فضای نمونه‌ای مفروضی بتوانیم چنان افرازی انتخاب کنیم که با دانستن این‌که کدام‌یک از پیشامدهای افراز شده رخ‌داده است، بخش مهمی از عدم اطمینان تقلیل میابد.

این قضیه ازآن‌جهت مفید است که می‌توان از طریق آن احتمال یک پیشامد را با مشروط کردن نسبت به وقوع و یا عدم وقوع یک پیشامد دیگر محاسبه کرد. در بسیاری از حالت‌ها، محاسبه احتمال یک پیشامد به‌صورت مستقیم کاری دشوار است. با استفاده از این قضیه و مشروط کردن پیشامد موردنظر نسبت به پیشامد دیگر، می‌توان احتمال موردنظر را محاسبه کرد.

دسته‌بندی مبتنی بر قاعده: هم‌زمان با پیدایش علم داده‌کاوی در دهه ۹۰ الگوریتم‌های استخراج قوانین وابستگی از پایگاه داده‌ها نیز پا به عرصه گذاشت. نویسندگان زیادی در زمینه استخراج قوانین وابستگی در پایگاه داده‌ها بحث کرده‌اند رابرت. اس (۲۰۰۳) در مقاله خود اقدام به مقایسه الگوریتم‌های مهم استخراج قوانین وابستگی پرداخت است. در این مطالعه مزیت‌های و معایب سه الگوریتم مهم مورداستفاده در استخراج قوانین وابستگی یعنی sampling، apriori و partitioning پرداخته است.

اساساً ارتباط میان مجموعه اشیا (چیزها) وابستگی‌های جالب توجهی هستند که منجر به امکان آشکارسازی الگوهای مفید و قوانین وابستگی برای پشتیبانی تصمیم، پیش‌بینی‌های مالی، سیاست‌های بازاریابی، وقایع پزشکی و خیلی کاربردهای دیگر می‌شود. در حقیقت توجهات زیادی را در تحقیقات اخیر به خود جلب کرده است. تحلیل وابستگی‌ها یک حالت غیر نظارتی داده‌کاوی است که به جستجو برای یافتن ارتباط در مجموعه‌ها می‌پردازد. (**تیمورپور، بابک؛ نجفی حیدر، ۱۳۹۴**)

درخت تصمیم: درخت تصمیم یک مدل یادگیرنده است که در یک ساختار درخت مانند نمونه تا را تفکیک می‌کند. شکل یک درخت ساده تصمیم را نمایش می‌دهد که یک مجموعه داده را بر اساس مقادیر دو ویژگی به سه کلاس تقسیم کرده است. درخت‌های تصمیم ساده برای درک نحوه دسته‌بندی بسیار قابل‌فهم هستند. نمونه‌های جدید بر اساس میزان پیروی از هر یک از سه‌شاخه موجود بر اساس ویژگی‌هایشان دسته‌بندی می‌شوند. روش‌های از قبیل C4.5 با یک درخت خالی شروع کرده و مرتباً داده‌ها را تقسیم می‌کنند، شاخه‌های درخت را می‌سازند تا زمانی که تمام نمونه‌های یک شاخه به یک دسته خاص تعلق گیرد، برگ‌های درختان بر اساس معیارهای خاصی ساخته می‌شود. میزان خطا در شاخه‌های درختان به‌اندازه کافی کم است.(Han et al., 2011) (شکل ۳-۲)



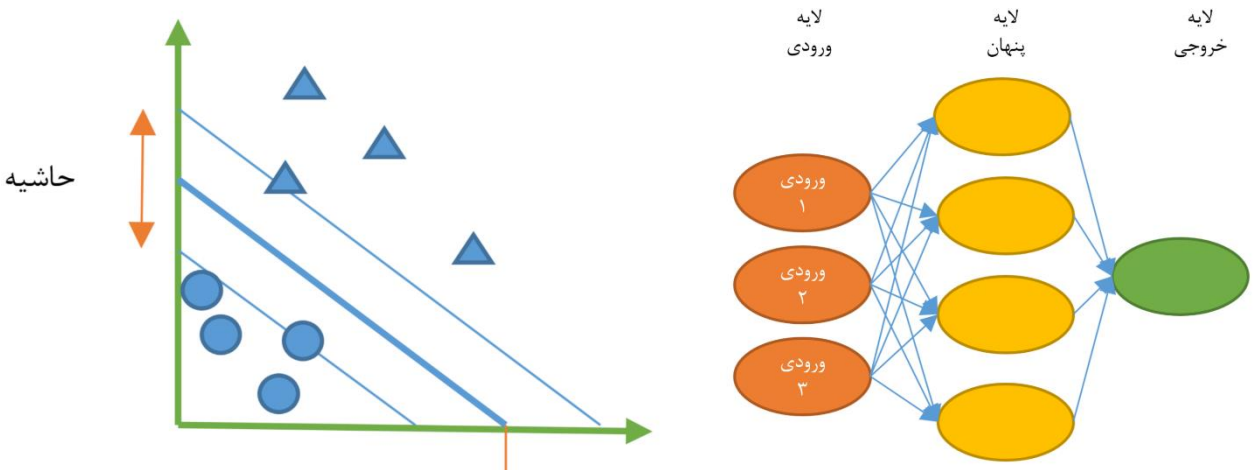
شکل۲-۳درخت تصمیم (Swan et al., 2013)

جنگل تصادفی: جنگل تصادفی بر مبنای مجموعه‌ای از درخت‌ها که بر اساس داده‌های آموزشی مدل شده‌اند ساخته می‌شود. هرکدام از درختان تصمیم به یک زیرمجموعه از ویژگی‌های نمونه تا دسترسی دارند و در آخر هنگام پیش‌بینی داده‌های تست، هرکدام از درختان یک دسته را برای داده موردنظر پیش‌بینی می‌کنند. هرکدام از دسته‌ها که بیشترین رأی را بیاورد به‌عنوان دسته آن داده موردنظر انتخاب می‌شود.(Meyfroidt et al., 2009)

بردارهای ماشین پشتیبان: این مدل به‌وسیله vapnik توسعه داده شد، شکل ... یک الگوریتم SVM را نمایش می‌دهد. این الگوریتم بر اساس مفهوم جدا پذیر بودن خطی داده‌ها را دسته‌بندی می‌کند. ویژگی به خصوصی که SVM را تعریف می‌کند عبارت‌اند از: ۱) تعیین معیاری که بهترین دسته‌بند خطی را بر اساس بیشینه کردن حاشیه تا تعریف می‌کند. ۲) شناسایی بردارهای پشتیبان که کم‌ترین تعداد نمونه داده‌های آموزشی نیاز است تا بهترین دسته‌بند خطی تعریف شود، به این دلیل که آن‌ها در مرز حاشیه تا قرار دارند. ۳) استفاده از کرنل‌ها برای انتقال ویژگی‌های اصلی به یک فضای بزرگ‌تر غیرخطی به‌منظور ایجاد جدا پذیری خطی است. الگوریتم SMO[5] یکی از پرطرفدارترین الگوریتم‌های SVM است.(شکل ۳-۴)

شبکه‌های عصبی مصنوعی: شبکه‌های عصبی مصنوعی با الهام از ساختار و عملکرد مغز انسان توسعه داده شده‌اند. این شبکه از مجموعه‌ای از عناصر محاسبه‌گر (نرون) که بر اساس یک الگوی درون ارتباطی گسترده به هم متصل شده‌اند. در کل هر نرون یک متغیر از یک دسته‌بند خطی است، اما با آمیختگی این نرون‌ها یک مدل دسته‌بندی پیچیده غیرخطی ساخته می‌شود که می‌تواند برای حل مسائل پیچیده استفاده شود. شکل ۳-۳ یک مدل شبکه عصبی مصنوعی را نمایش می‌دهد.(Swan et al., 2013).

[5] Sequential minimal optimization



۳شکل۴-۳شبکه های عصبی (Jagga and Gupta, 2015)

۳شکل۳-۳ بردارهای ماشین پشتیبان (Swan et al., 2013)

جدول۱-۳مقایسه مدل‌های دسته‌بند(Swan et al., 2013)

روش	مزایا و معایب	سرعت یادگیری	سهولت تفسیر پذیری
بیز ساده	مزایا:قابلیت ساده و سریع در بکارگیری، مناسب برای مجموعه داده های با داده های گم شده، معایب: فرض مستقل بودن ویژگی ها از هم	۱	۴
درخت تصمیم	خروجی این الگوریتم به سادگی قابل تفسیر است اما بستگی به نوع الگوریتم مورد استفاده و پیچیدگی درخت ساخته شده دارد، همچنان مناسب برای مجموعه داده ها با داده گم شده	۲	۱
جنگل تصادفی	روشی کارآمد برای مجموعه داده های بزرگ اگرچه در مقابل outlierها حساس نیست	۴	۳
دسته بندهای مبتنی بر قاعده	قواعد تولید شده به سادگی قابل خواندن است، مناسب برای کشف زیست نشانگرهای پنهان، اما امکان بیش برازش نیز دارد	۳	۱
بردارهای ماشین پشتیبان	استفاده از کرنل برای فراگیری توابع پیچیده، با این وجود بسیار کند بوده و چندین پارامتر نیز توسط کاربر باید تعریف شود	۵	۵
شبکه های عصبی مصنوعی	نتایج خروجی قابلیت خواندن ندارد و آموزش مدل ممکن است بسیار آهسته صورت گیرد.	۵	۵

۳,۲,۵ وارسی اعتبار[6]

از چالش‌های بزرگ داده‌کاوی و ساخت مدل بر روی داده‌های پزشکی اعتبار سنجی مدل بنا شده بر اساس داده‌های آموزشی است. بدین منظور داده تا به دو بخش داده‌های آموزش و داده‌های تست تقسیم می‌شوند و مدل بر اساس داده‌های آموزشی توسعه پیدا

[6] Cross-validation

می‌کند سپس به کمک داده‌های تست مدل اعتبار سنجی می‌شود. یکی از رویکردهای ساده استفاده از نمونه‌برداری تصادفی RS[۷] برای تقسیم داده‌ها به دو دسته داده‌های آموزشی و تست است؛ اما به دلیل این که فرآیندهای نمونه‌گیری پزشکی دشوار، پیچیده و هزینه‌بر است معمولاً تعداد نمونه بسیار کم است و از طرفی تعداد داده‌های آموزشی بر روی عملکرد و دقت مدل بسیار تأثیرگذار است. برای غلبه بر این مشکل از استفاده از روش‌های پیچیده‌تر نمونه‌گیری از مجموعه داده‌هاست که یکی از این روش‌ها وارسی اعتبار است. در این روش مجموعه داده به دو بخش داده‌های آموزش و تست تقسیم می‌شود و سپس مدل بر اساس داده‌های آموزش توسعه پیدا کرد و با داده‌های تست ارزیابی می‌شود. این رویه چند بار تکرار می‌شود و در آخر میزان خطا میانگین خطاهای به‌دست‌آمده از تکرار مراحل قبل است. یکی از رویکردهای رایج استفاده از الگوریتم وارسی اعتبار K-Fold است. در این روش مجموعه داده به k زیرمجموعه تقسیم شده و هر بار از یکی از این زیرمجموعه‌ها به‌عنوان داده تست و از k-1 داده دیگر به‌عنوان داده آموزش استفاده می‌شود این رویه انقدر ادامه میابد تا تمام زیرمجموعه‌ها یک بار به‌عنوان داده تست انتخاب شوند. روش دیگر استفاده از الگوریتم وارسی اعتبار یکی-بیرون[۸] است. این روش دقیقاً مشابه رویکرد k-Fold است تنها با این تفاوت که تعداد k برابر برابر با تعداد نمونه‌هاست اما این روش حجم محاسبات را بالا می‌برد.(Thomas et al., 2006)

۳,۲,۶ بررسی عملکرد مدل

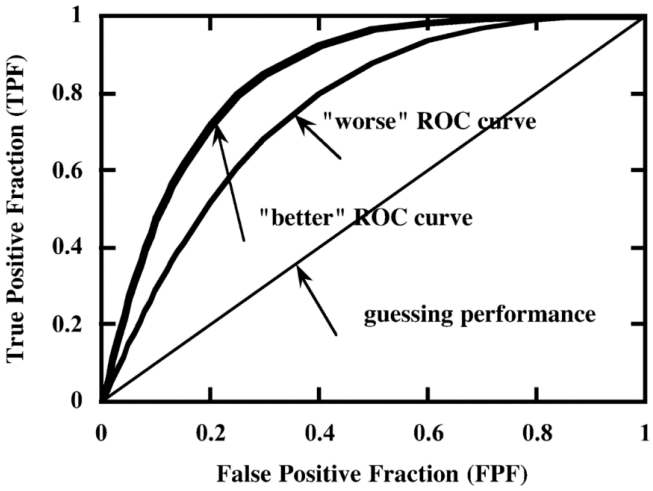
آخرین مرحله از فرآیند داده‌کاوی، مرحله بررسی عملکرد است. در این قسمت به چند روش برای ارزیابی عملکرد مدل‌های دسته‌بند در مواجه با داده‌های تست می‌پردازیم.

دقت[۹]: دقت یک مدل بر اساس نسبت میزان داده‌هایی که به‌درستی دسته‌بندی شده‌اند به مجموع داده‌هاست. معیار دقت با توجه به این اگر یک دسته به‌خصوص به‌طور معناداری بیشتر از سایر دسته‌ها باشد می‌تواند گمراه‌کننده باشد.

حساسیت و ویژگی[۱۰]: برای دودسته حاصل از مدل چهار خروجی: درست-مثبت، درست-منفی، غلط-مثبت و غلط-منفی امکان‌پذیر است. حساسیت نسبت تعداد نمونه‌های مثبت که به‌درستی دسته‌بندی شده‌اند به‌کل نمونه‌های مثبت است؛ اما ویژگی در داده‌های پزشکی و تشخیصی احتمال یک شخص سالم به اشتباه به‌عنوان بیمار دسته‌بندی شود اس

منحنی ROC. این منحنی میزان حساسیت و ویژگی یک مدل تصمیم را بر اساس تمام ترشهلدهای[۱۱] ممکن تصمیم بیان می‌کند. این منحنی یک تصویر کلی از عملکرد مدل به ما می‌دهد همچنان که می‌تواند برای انتخاب بهینه سطح بهینه تصمیمات که منجر به افزایش دقت مدل دسته‌بند می‌شود استفاده کرد. همچنان می‌توان از این منحنی برای مقایسه عملکرد چندین مدل دسته‌بند استفاده کرد. شکل ۵-۳ یک نمونه از این منحنی را که برای مقایسه عملکرد دو مدل استفاده‌شده است را نمایش می‌دهد. درمجموع هر چه سطح زیر نمودار یک مدل بیشتر باشد آن مدل بهتر است.

[7] Random sampling
[8] Leave-one-out cross validation
[9] Accuracy
[10] Sensitivity and specificity
[11] threshold



شکل ۵-۳نمونه‌ای از منحنی ROC (Thomas et al., 2006)

۳,۳ مروری بر ادبیات کاربرد داده‌کاوی در دسته‌بندی و تشخیص زیست‌نشانگرهای سرطان

برای بررسی تحقیقات گذشته ما چند معیار برای انتخاب مقالات داشتیم. اول این که این مقالات برای ارائه مدل‌های دسته‌بند و یا کشف زیست‌نشانگرها چه پردازش مستقیم بر روی موجک و چه پردازش بر روی پروتئین‌های شناسایی شده به کمک روش‌های های بدون برچسب گذاری و برچسب گذاری باشد. دوم این که فرآیند ساخت مدل بر روی داده‌های حاصل از طیف‌سنج جرمی به دلیل توان بالا این روش در کشف زیست‌نشانگرها و میزان بالا پروتئین‌های استخراجی که به‌طور معمول بین ۳۰۰۰ تا ۱۵۰۰۰ و بیش از آن است. بدین منظور از فهرست ارائه‌شده توسط دو مرجع (Swan et al., 2013) و (Jagga and Gupta, 2015) استفاده کردیم. اما یکی از مشکلات اساسی این مراجع، بیشتر مقالات فهرست شده مربوط به سال‌های قبل از ۲۰۱۱ بود. برای رفع این مشکل مجدداً در موتور جستجوی گوگل اسکولار(https://scholar.google.com/)، پایگاه داده ساینس دایرکت (http://www.sciencedirect.com/) و همچنین بانک مقالات NCBI[۱۲](https://www.ncbi.nlm.nih.gov/) کلید واژه‌های "cancer","proteomic mass spectrometry","machine learning","data mining" را جستجو کرده و پس از یافتن مقالات مناسب با مقالات قبلی ادغام کرده و بخش مهم این مقالات که *مواد و روش‌های تحقیق* و بخش *نتیجه‌گیری و استنتاج بود* مورد بررسی و نقد قرار دادیم.

۳,۳,۱ بررسی مقالات و تحقیقات صورت گرفته

در داده‌کاوی پروتئوم همان‌طور که اشاره شده دو هدف دسته‌بندی سرطان‌ها و کشف زیست‌نشانگرهای سرطان‌ها دنبال می‌شود. بیشتر تحقیقات صورت گرفته در این حوزه ارائه پردازش بروی موجک و به عبارت جدول داده‌های ماتریس شدت صورت گرفته است. برخی از این تحقیقات فقط به ارائه مدل‌های دسته‌بند پرداخته‌اند مانند (Htike and Win, 2015) که از مدل ترکیبی

[12] National Center for Biotechnology Information

درخت لجستیک برای دسته‌بندی ۸۰ نمونه بیمار سرطان لوزالمعده و ۱۰۱ نمونه فرد سالم انجام داده‌اند. بیشترین مدل‌های استفاده‌شده در تحقیقات اخیر استفاده مدل الگوریتم‌های بردار ماشین پشتیبان و درخت تصمیم است و این به دلیل توانایی بالای این دو الگوریتم در برخورد با داده‌های گمشده است. نکته‌ای که قابل‌توجه است استفاده از انتخاب ویژگی برای کاهش بعد است که منجر به افزایش دقت معنادار دقت دسته‌بند فارغ از هدف تحقیق که خواه برای کشف زیست‌نشانگر یا خواه صرفاً برای دسته‌بندی باشد. تحقیق (Guan et al., 2009) یک مثال عالی برای تائید این موضوع است. این تحقیق برای دسته‌بندی ۳۵ نمونه سرطان تخمدان و ۳۵ نمونه کنترلی است و برای دسته‌بندی از الگوریتم بردار ماشین پشتیبان استفاده کرده است. زمانی که از این الگوریتم استفاده کرده است به دقت ۸۳,۳٪ با وارسی اعتبار یکی-بیرون رسیده است. در نوبت دیگر از یک الگوریتم انتخاب ویژگی مبتنی بر ماشین پشتیبان همراه مدل استفاده کرده است و این بار به دقت ۹۷,۲٪ رسیده است که این مقدار تفاوت معناداری را نمایش می‌دهد. همچنان این تحقیق برای نمونه‌گیری برای انتخاب داده‌های تست مقایسه‌ای بین دو روش تقسیم تصادفی داده‌ها به دو دسته تست و آموزشی و روش وارسی اعتبار که از الگوریتم یکی-بیرون استفاده شد پرداخته است که به‌وضوح برتری و دقت روش وارسی اعتبار اثبات شده است. البته باید به این نکته توجه داشت که نتایج این تحقیق بر روی پردازش یک مجموعه داده به دست آمده و نتیجه‌ای قطعی و اثبات شده نیست. همچنان لازم به ذکر است که در مرحله نمونه‌برداری برای بررسی عملکرد تقریباً نزدیک به تمام تحقیقات از روش‌های وارسی اعتبار استفاده کرده‌اند. در روش‌های جمع‌آوری نمونه مطلوب‌ترین نمونه‌ها برای کشف زیست‌نشانگرها، مایعات زیست پذیر نمونه‌های مربوط به خون است به دلیل این که خون در بدن انسان توسط قلب پمپاژ شده و در سرتاسر بدن به گردش درمی‌آید و به همین دلیل می‌تواند حاوی اطلاعات ارزشمندی جهت شناخت و توصیف نشانگرهای سرطانی باشد. در پیش‌پردازش داده‌ها بیشتر پژوهش‌ها از نرم‌افزار آماده استفاده کرده‌اند که فهرستی از مهم‌ترین نرم‌افزارهای مورداستفاده در پیوست ((الف)) آمده است. البته برخی تحقیقات مانند (Ushijima et al., 2007) از روش‌های خاصی برای پیش‌پردازش داده‌ها استفاده کرده‌اند به‌طور مثال برای نرمال‌سازی داده‌ها از فرمول زیر استفاده کرده‌اند به‌طوری که V_{max} نشان دهنده شدیدترین سیگنال مشاهده شده و V_{min} نشان‌دهنده ضعیف‌ترین سیگنال مشاهده‌شده است. همچنین V_i موجک موردنظر ما است که قصد نرمال‌سازی آن را داریم. اعداد به‌دست‌آمده در بازه [1-0] است.

$$NV_i = \frac{V_i - V_{\min}}{V_{\max} - V_{\min}},$$

برای انتخاب موجک استفاده از الگوریتم نزدیک‌ترین همسایه با K=10 برای عرض پنجره‌های شناسایی موجک‌ها استفاده‌شده است.

برای تشخیص موجک‌های مشترک میان سوژه‌ها موردنظر از فرمول زیر که با میانگین هسته گوسی با مراکز موجک‌ها ساخته‌شده است استفاده کرده‌اند.

$$A(x) = \frac{1}{N_G} \sum_{i=1}^{N_G} \sum_{j} \exp\left[-\frac{(x - p_{i,j})^2}{(\sigma p_{i,j})^2} \right]$$

در این فرمول N_G بیانگر اندازه هر گروه، $p_{i,j}$ بیانگر مقدار m/z از i امین سوژه و j امین موجک است و σ بیانگر پارامتر محاسبه‌شده برای پهنا موجک‌ها است.

همچنین در این تحقیق مقایسه‌ای بین روش دسته‌بند الگوریتم *AdaBosst* و مدل دسته‌بند بردار ماشین پشتیبان با الگوریتم *SVM-RFE* صورت گرفته است که نتایج آن‌ها بر روی مجموعه داده‌هایشان برتری اندک *AdaBoost* را نشان می‌دهد اگرچه این اختلاف جزئی است اما به دلیل اینکه الگوریتم *SVM-RFE* محاسبات پیچیده‌تری و زمان بری را انجام می‌دهد الگوریتم *AdaBoost* را برتر می‌دانند.

در جدول ۳-۲ تحقیقات صورت گرفته از سال ۲۰۰۲ تا سال ۲۰۰۱۵ را به‌طور خلاصه آورده‌ایم. لازم به ذکر است که ارزیابی عملکرد مدل‌های دسته‌بند بین تحقیقات یکسان نبوده است. به‌گونه‌ای که برخی از معیارهای دقت و برخی از حساسیت و ویژگی استفاده کرده‌اند و بعضاً در برخی تحقیقات از معیارهای *ROC* و اندازه‌گیری سطح زیر منحنی استفاده‌شده است.

جدول ۲-۳مرور تحقیقات گذشته دسته‌بندی سران و کشف زیست‌نشانگر

مرجع	دقت عملکرد دسته‌بند و زیست‌نشانگرهای شناسایی‌شده	روش‌های شناسایی، داده‌کاوی و ارزیابی	هدف پژوهش و مجموعه داده تا
(Sinues et al., 2015)	دقت دسته بند بالا ۹۰٪ و کشف هشت زیست نشانگر	استفاده از مدلSVM برای دسته بندی و انتخاب ویژگی	دسته بندی و کشف زیست نشانگرهای سرطان سینه با داده های بزاق دهان، ۱۴۱ نمونه بیمار و ۱۱ نمونه سالم، آنالیز داده تا SESI-MS
(Htike and Win, 2015)	دقت عملکرد دسته بند: ۷۴,۰۳۳۱٪	روش ترکیبی لوجستیک درخت برای مدل سازی، الگوریتم RELIEF برای انتخاب ویژگی، استفاده از روش TOP-HAT برای کاهش خط مبنا، استفاده از رویکرد یکی بیرون برای اعتبار سنجی	هدف دسته بندی سرطان لوزالمعده ۸۰ نمونه سرطان لوزالمعده در مقابل ۱۰۱ نمونه سالم،
(Guan et al., 2009)	بهترین دقت ۸۳٪ با استفاده از SVM و LOO-CV، رسیدن به دقت ۹۷,۲٪ با ترکیب مدل SVM و انتخاب ویژگی بر مبنای SVM، شناسایی ۳۸ زیست نشانگر با چهار رویکرد متفاوت	شناسایی با نرم افزار mzMine (v0.60)، SVMs با روش های مرتبط انتخاب ویژگی، LOO-CV,12-FOLD CV,52-20-split validation	دسته بندی و کشف زیست نشانگرهای سرطان تخمدان، ۳۷ بیمار بیماران مبتلا به سرطان تخمدان پاپیلری سروز و ۳۵ نمونه کنترلی
(Vlahou et al., 2003)	۸۵٪ دقت دسته بندی	دسته بندی بر اساس درخت-رگرسیون (CART)، اعتبار سنجی 10-FOLD، برای دسته بندی موجک استفاده از نرم افزار Ciphergen Systems، نرم افزار الگوهای زیست نشانگر (BPS)	دسته بندی سرطان بدخیم، خوشخیم و سالم در سرطان تخمدان، ۴۴ سرطان بدخیم، ۶۱ سرطان خوشخیم و ۳۴ نمونه سالم
(Le et al., 2005)	۸۵٪ دقت و پیدا کردن چند زیست نشانگر	استفاده از نرم افزار mascot و استفاده از C.SVM با استفاده از اعتبار سنجی یکی-بیرون	دسته بندی نمونه های پروستات از ۱۹ بیمار با متاستاز استخوان و ۱۹ بیمار فاقد آن
(Adam et al., 2002)	۹۰٪ از داده های تست به درستی دسته بندی شدند.	داده تست شامل: ۱۵ کنترل، ۱۵ خوشخیم و ۳۰ پروستات. شناسایی	دسته بندی سرطان پروستات و کنترل، ۹۷ نمونه کنترلی، ۹۲

		موجک با نرم افزار Ciphergen SELDI software و همگام سازی موجک تا با الگوریتم peakminer، استفاده از درخت تصمیم.	نمونه سرطان خوشخیم، ۱۹۷ سرطان پروستات
(Oh et al., 2009)	دقت ۸۷,۹٪، ۲۶ موجک به عنوان زیست نشانگرهای محتمل شناسایی شد.	پیش پردازش: گروه بندی، تصحیح خط مبنا و نرمال سازی با استفاده از نرم افزارTOFWorks	شناسایی زیست نشانگر برای سرطان پروستات، ۱۷۹ سرطان ادرنوکازسینوما و ۷۴ سرطان خوشخیم
(Ralhan et al., 2008)	سه تا از زیست نشانگرها شناخته شد.	استفاده از نرم افزارهای ProteinPilot و iTRQ. استفاده از مدل بیز ساده و اعتبار سنجی 3-FOLD	شناسایی زیست نشانگرهای سرطان سر و گردن، پنج مجموعه از چهار نمونه به همراه نمونه کنترل برای هر مجموعه
(Zhang and Chen, 2009)	۸۵٪ دقت در داده های تست، دوتا از بهترین پنج پنل پروتئین شناخته شد که شامل هفت پروتئین شد.	شناسایی پروتئین تا به صورت بدون برچسب گذاری با کمک نرم افزارEli Lilly ، استفاده از مدل شبکه های عصبی مصنوعی	توسعه پنل[۱۳] برای شناسایی زیست نشانگرهای سرطان سینه، ۴۰ نمونه پلاسما خون از سرطان سینه و ۴۰ نمونه از سالم
(Ushijima et al., 2007)	۶ زیست نشانگر شناخته شد	استفاده از نرم افزار SpecAlign برای کاهش خط مبنا، همگام سازی موجک، استفاده از مدل AdBoost برای دسته بندی و 5-FOLD برای CV	کشف زیست نشانگرها، ۶۵ نمونه از بیماران مبتلا به سرطان سینه، سپس نمونه گیری مجدد از آنها بعد از مصرف چهار هفته ای docetaxel 75 mg/m2
(Zhang et al., 2007)	نه موجک به عنوان زیست نشانگر بالقوه انتخاب در نظر گرفته شد، چهار موجک برای پیش بینی پاسخ بیماران به درمان های استاندارد، حساسیت ۹۴٪، ویژگی ۹۴٪ در در ۸۵ نمونه از مجموعه تست، ۹۴٪ حساسیت و ۹۲٪ ویژگی با ۶۶ نمونه تست.	استفاده از درخت تصمیم برای مدل سازی و استفاده از نرم افزار Ciphergen برای پیش پردازش داده تا	شناسایی زیست نشانگرها، ۱۳۲ بیمار با سرطان لنفوم B-Cell و ۷۵ نمونه کنترلی، داده های آنالیز شده SELDI-TOF-MS از سروم خون
(Cohen et al., 2011)	۹ نشانگر یافت شده با ۸۹٪ دقت در عملکرد	استفاده از مدل دسته بند مبتنی بر SVM و برای وارسی اعتبار از الگوریتم 10-FOLD استفاده شده است.	کشف زیست نشانگرهای سرطان معده، ۷۹ نمونه از بیماران مبتلا به سرطان معده و ۳۳ نمونه از افراد فاقد سرطان که ۱۰ نفر از آنها دارای التهاب معده بودند.
(Washam et al., 2013)	شناسایی ۱۳ زیست نشانگر، قدرت دسته بند با حساسیت ۹۱٪ و ویژگی ۹۳٪	استفاده از جنگل تصادفی برای مدل دسته بند، ساخت درخت ۱۰۰۰ مرتبه تکرار شده و هر بار در هر گره سه ویژگی مورد آزمایش قرار می گرفت	شناسایی زیست نشانگرها که موجب متاستاز استخوانی در سرطان سینه می شود، نمونه از ۱۱۱ زن مبتلا به سرطان سینه به دو گروه تقسیم شده که گروه اول شامل ۴۱ نفر متاستاز استخوان و ۳۶ نفر دیگر فاقد آن، گروه دیگر

[13] Panel (منظور تعدادی از پروتئین ها که با هم کار می کنند و ظاهر می شوند)

			شامل ۱۷ نفر متاستاز استخون و ۱۷ نفر دیگر فاقد آن

۳,۴ خلاصه فصل

در این فصل به بررسی مراحل داده‌کاوی بر روی پروتئوم پرداختیم. پیش‌پردازش داده‌ها معمولاً برای مواردی است که مستقیماً پردازش را بر روی موجک‌ها انجام می‌دهیم. فعالیت‌های که در پیش‌پردازش استفاده می‌شود شامل کاهش خ مبنا، کاهش نویز و نرمال‌سازی و انتخاب موجک‌هاست. شش مدل درخت تصمیم، بردارهای ماشین پشتیبان، جنگل تصادفی، مدل‌های مبتنی بر قاعده، بیز ساده و شبکه‌های عصبی مصنوعی را معرفی کردیم و ضعف و قدرت آن‌ها را بیان کردیم. برای وارسی اعتبار روش‌های یکی-بیرون و K-FOLD را معرفی کردیم. سپس معیارهای ارزیابی مدل مانند ویژگی، حساسیت، دقت و همچنین ROC را شرح دادیم. در مرحله آخر به مرور ادبیات کاربرد داده‌کاوی در دسته‌بندی و کشف زیست‌نشانگرهای سرطانی پرداختیم. مشاهده کردیم که بیشترین مدل‌های استفاده‌شده بردارهای ماشین پشتیبان و درخت تصمیم بودند. همچنین استفاده از روش‌های انتخاب ویژگی علاوه بر این‌که برای کشف زیست‌نشانگرها ضروری است موجب دقت و سرعت عملکرد مدل دسته‌بند می‌شود.

فصل چهارم

جمع‌بندی و نتیجه‌گیری

۴,۱ مقدمه

در این فصل ابتدا به مرور فصل‌های گذشته خواهیم پرداخت و سپس به بیان چالش‌ها و فرصت‌های مطالعات سرطان با بکارگیری داده‌کاوی پروتئوم خواهیم پرداخت و پس از آن مجموعه داده‌های خود را شرح خواهیم و تعدادی بانک‌داده مناسب برای تحقیق معرفی خواهیم کرد.

۴,۲ مروری بر فصل‌های گذشته

در فصل اول به بیان مفاهیم و کلیات پرداختیم. دانستیم که انسان تقریباً متشکل از چهار ابرمولکول کربوهیدرات‌ها، پروتئین‌ها، اسیدهای نوکلئوتید و لیپیدها است و دراین‌بین پروتئین‌ها به دلیل رفتار پویا در بدن انسان، تغییرات پس از ترجمه و همچنین معادل نبودن میزان تولیدشان مطابق با mRNA گزینه‌ای مناسب‌تر برای مطالعات بیماری‌ها در سطح مولکولی است. سرطان را بیماری ژنتیکی در سطح مولکولی تعریف کردیم که از تکثیر غیرعادی سلول‌ها به دلیل کم‌کاری، پرکاری و یا عدم حضور پروتئینی خاص در یک بافت یا ارگان از بدن بروز می‌کند؛ بنابراین میزان پروتئین‌های بیان‌شده، الگو و رفتار آن‌ها می‌تواند معیاری مناسب برای تشخیص بیماری‌ها و کشف زیست‌نشانگرها –ابرمولکول‌ها- مرتبط با بیماری‌ها به‌خصوص در سرطان باشد. به مجموع پروتئین‌های بیان‌شده و الگوی آن‌ها در یک‌لحظه خاص در سلول یا یک بافت پروتئوم گفته می‌شود و علم بررسی پروتئوم را پروتئومیکس می‌گویند. در علم پروتئومیکس کشف این تغییرات و کشف این زیست‌نشانگرها می‌تواند منجر به ساخت داروهای جدید و همچنین ابداع روش‌های درمانی جدید نیز گردد. بیان کردیم، به دلیل این‌که نمونه‌های بیمارستانی معمولاً تعداد نمونه‌هایشان کم اما ویژگی و متغیرهای بسیاری دارند برای تحلیل و دسته‌بندی آن‌ها به‌خصوص جهت کشف زیست‌نشانگرها نیاز به بهره‌گیری از علومی مانند داده‌کاوی، یادگیری ماشین و هوش مصنوعی اجتناب‌ناپذیر است. سپس در پایان فصل اول برخی از اصطلاحات موردنیاز برای ورود به دنیای پروتئومیکس را تعریف کردیم و پس‌ازآن آرایش کلی گزارش خود را ارائه کردیم.

در فصل دوم ابتدا به سراغ روش‌ها و رویکردهای پروتئومیکس به حل مسائل پرداختیم. دانستیم که پروتئومیکس از روش‌های شناسایی، توصیف و کمیت شماری پروتئین‌ها برای مقایسه و دسته‌بندی دو دسته‌بیمار و دسته‌کنترل جهت دسته‌بندی سرطان‌ها و کشف زیست‌نشانگرها استفاده می‌کند. در رویکرد تغییرات پس از ترجمه به اتفاقات و فعالیت‌های پروتئین‌ها پس از ترجمه در سلول و بافت را مشاهده و اندازه‌گیری می‌کند. در مکان‌یابی پروتئین‌ها به بررسی مکانی پروتئین‌ها بنا به شرایط بالینی مختلف

انسان می‌پردازد و در برهم‌کنش پروتئین‌ها، تعاملات پروتئین‌ها را در یک شبکه ارتباطی با یکدیگر می‌سنجد؛ که هدف ما در این تحقیق پروتئیومیکس کمی و مقایسه‌ای بود. برای دریافت نمونه از بیماران سه نوع نمونه‌گیری مایعات زیست پذیر، بافت‌ها و سل لاین‌ها وجود دارد که باوجوداینکه بافت‌ها اطلاعات ارزشمندی را در خود دارند اما به دلیل مشکلات نمونه‌برداری و تهاجمی بودن آن‌ها توصیه نمی‌شود ولی در مقابل مایعات زیست پذیر هم به دلیل گرددش در بدن اطلاعات متنوعی در خود دارند و به‌خصوص نمونه‌های ادرار به دلیل غیرتهاجمی بودن نمونه‌های مناسب برای پروتئومیکس است. فناوری‌های بکار برده شده در پروتئومیکس را به سه دسته عمده فنّاوری ریزآرایه‌های پروتئینی، ژل الکتروفوز دوبعدی پلی آکریل آمید (2D-PAGE) و طیف‌سنج جرمی تقسیم‌بندی کردیم و عنوان نمودیم که طیف‌سنج جرمی به دلیل سهولت و توان بالا در آنالیز حجم عظیمی از پروتئین‌ها نسبت به روش‌ها دیگر بیشتر موردتوجه محققین قرارگرفته است و نکته مهم دیگر اینکه غیر جانب‌دارانه است یعنی برخلاف فناوری‌های دیگر که از قبل فرض بر وجود تعداد مشخصی پروتئین در نمونه گرفته می‌شود و برای آنتی‌بادی جهت به دام انداختن پروتئین‌ها ساخته می‌شود، است و تعداد و نوع پروتئین‌ها از قبل مشخص نیست. طیف‌سنج خود به انواع مختلف ازلحاظ فناوری لیزر تقسیم می‌شود اما دو رویکرد عمده در رابطه با آن وجود دارد رویکرد اول از پائین به بالا هست که ابتدا پروتئین‌ها را تجزیه به پپتیدها و شاخه‌های آمینواسیدهای آن‌ها می‌کنند و سپس آنالیز می‌شود و بعد از آنالیز با فن‌آوری‌های خاصی مجدداً نوع پروتئین‌ها حاصل از آنالیز شناسایی می‌کنند در رویکرد دوم یا از بالابه‌پایئین پروتئین‌ها بدون تجزیه شدن به قطعات کوچک‌تر خود مستقیماً تفکیک، کمیت‌شماری و شناسایی می‌شوند که رویکرد دوم به دلیل مشکلات آزمایشگاهی پیچیده و هزینه‌های بیشتر مرود توجه محققین نبوده است. ما هم تمرکز خود را به داده‌های حاصل از رویکرد اول گذاشتیم و دانستیم نتایج حاصل از رویکرد اول به دو صورت عمده مورد پردازش قرار می‌گیرد. اول این‌که مستقیماً به سراغ موجک‌ها حاصل برویم که این امر نیاز به پیش‌پردازش داده‌ها و همچنین آنالیزهای اضافه‌تر برای شناسایی پروتئین‌ها دارد و در روش دوم استفاده از موتورهای جستجو بانک توالی اطلاعات پروتئینی است که در پیوستگی استفاده از روش‌ها برچسب‌گذاری که روشی آزمایشگاهی است و یا غیر برچسب‌گذاری که روشی غیر آزمایشگاهی و مبتنی بر علوم کامپیوتری است.

در فصل سوم به سراغ مراحل داده‌کاوی بر روی پروتئوم جهت دسته‌بندی و کشف زیست‌نشانگرها رفتیم و بیان کردیم زمانی که مستقیماً به سراغ موجک‌ها برویم نیاز به پیش‌پردازش داده‌ها از قبیل کاهش خط مبنا، نرمال‌سازی داده‌ها و حذف نویزها داریم که برای این مسئله نرم‌افزارهای کامپیوتری مناسب و همچنین روش‌های مختلف وجود دارد. از اهمیت کاهش بعد در داده‌های پروتئینی گفتیم و بیان کردیم برای کشف زیست‌نشانگرها انتخاب ویژگی یک امر ضروری است و برای دسته‌بندی سرطان‌ها هم موجب افزایش سرعت و دقت دسته‌بند ما می‌شود سپس به سراغ مدل‌های پراستفاده دسته‌بند در تحقیقات چند سال اخیر رفتیم و شش مدل: درخت تصمیم، بیز ساده، مدل مبتنی بر قاعده، بردارهای ماشین پشتیبان، جنگل تصادفی و شبکه‌های عصبی را تشریح کردیم و دریک جدول به‌طور خلاصه آن‌ها را با یکدیگر مقایسه نمودیم. سپس به سراغ روش‌های نمونه‌گیر از داده‌ها رفتیم و بیان کردیم به دلیل دشواری در نمونه‌گیر داده‌های پزشکی تعداد سطرهای داده‌ها بسیار کم است و نیاز به روش‌های پیچیده نمونه‌برداری جهت داده تست و داده آزمایش برای ساخت و ارزیابی مدل داریم. روش وارسی اعتبار را معرفی کردیم و دو نمونه K-FOLD و یکی بیرون را که از روش‌های وارسی اعتبار است شرح دادیم. در قسمت بعدی به بیان معیارهای ارزیابی مدل دسته‌بند خود پرداختیم و معیارهای دقت، ویژگی، حساسیت و منحنی ROC را معرفی نمودیم. در بخش آخر مروری انتقادی به تحقیقات و پژوهش‌های صورت گرفته داشتیم و آن‌ها را در یک جدول خلاصه بندی کردیم. دانستیم که پرکاربردترین مدل دسته‌بند الگوریتم بردار ماشین‌های پشتیبان و درخت تصمیم است به دلیل قدرت و دقت بالای دسته‌بندی که البته نشان دادیم در همه تحقیقات این نکته لزوماً صدق نمی‌کند. همچنین لزوم استفاده از انتخاب ویژگی را با نمایش تأثیرگذاری آن در یکی از تحقیقات نشان دادیم و

برای پیش‌پردازش داده‌ها علاوه بر اســتفاده از نرم‌افزارهای کامپیوتری اســتفاده از روش‌ها ابتکاری و جدید نیز صــورت می‌گیرد که اســتفاده آن در یکی از تحقیقات را بیان کردیم. نشان دادیم برخی تحقیقات علاوه بر دسته‌بندی، در کشف زیست‌نشانگرها به دنبال یافتن پنل‌ها هســتند. پنل‌ها به مجموعه‌ای از پروتئین‌ها گفته می‌شود که با یکدیگر در سلول فعالیت می‌کنند و با یکدیگر نیز ظاهر می‌شوند.

۴,۳ برسی چالش‌ها و پیشنهاد فرصت‌ها

تعـداد زیـادی چالش همچنان در طراحی آزمایش‌ها، تحلیل داده‌ها، اســـتانداردســـازی روش‌ها و ارائه و توزیع داده‌های حاصـل از پروتئومیکس در بانک داده‌ها موجود است. اثرات پروتئومیکس در تحقیقات پزشکی روزبه‌روز در حال بیشتر شدن است و ابزارهای پروتئومیکس به‌طور گسترده‌ای برای شناسایی و درمان بیماری‌ها بکار برده می‌شـود. نقطه کلیدی در این بین، شـناسـایی زیسـت‌نشانگرها است که عدم وجود یا تغییر در فراوانی آن‌ها در حالت زیسـتی سلول‌ها تأثیرگذار است. با توجه تحقیقات قبلی در حوزه‌های کشف زیست‌نشانگرها هنوز هم نیاز به مطالعات بیشــتر جهت شــناســایی هرچه بیشــتر زیســت‌نشــانگرها وجود دارد. به‌طور مثال می‌توان از روش دســته‌بندی مبتنی بر قاعده که در بررســی و تحلیل ریزآرایه‌ها در بیوانفورماتیک اســتفاده می‌شــود بهره برد. بدین‌صـورت که شـناسـایی شبکه‌ای از زیست‌نشانگرها که در یک قاعده خاص باهم ظاهر می‌شوند. همچنین از روش‌های علوم ژئومیکس مانند آنالیز مجموعه‌ای از ژن‌ها مشخص در علم پروتئومیکس استفاده کرد. بدین‌صورت که دسـته‌ای از پروتئین‌های کاندید را در دسـته‌های کنترل و بیمار شـناسـایی و کمیت‌شماری و مقایسه کنیم. همچنان بحث پیش‌پردازش و کاهش ابعاد در پروتئومیکس جدی اســت. با وجود تلاش‌های صــورت گرفته اما همچنان در بسیاری از تحقیقات نتایج مناسب نیست. به دلیل این‌که تعداد نمونه‌ها در داده‌های پروتئومیکس کم است دقت دسته‌بندهای گزارش‌شده می‌تواند گمراه‌کننده باشد و زمانی که این داده‌ها در دنیای واقعی با مقادیر نمونه‌های بسـیار زیاد اسـتفاده شـود نرخ خطاها بسـیار بیشتر می‌شود به همین دلیل توصیه می‌شود تا به بحث پیش‌پردازش داده‌هـا و روش‌هـای جدید کاهش ابعاد در علم پروتئومیکس پرداخته شــود. همچنین جای خالی ارتباط سـایر بیماری‌ها به‌طور مثال چاقی و یا دیابت با یک سـرطان خاص یا مجموعه‌ای از سـرطان‌ها به‌شـدت احساس می‌شود.

۴,۴ داده‌های تحقیق

در فصل دوم به طور مفصل از جنس داده‌های پروتئومیکس و نمایش آنها صحبت کردیم. داده‌های در دسترس ما مجموعه داده‌های با همکاری بخش آزمایشگاه پروتئومیکس دانشگاه کالیفرنیا جنوبی است که شامل ۲۰ نمونه مبتلا به سرطان لوزالمعده، ۲۰ نمونه عفونت مزمن لوزالمعده و ۲۰ نمونه از اشخاص سالم است. نمونه‌گیری از پلاسما خون است و این داده‌ها با فن‌آوری ESI-MS آنالیز شده است. همچنین برای توصیف و شناسایی و کمیت‌شماری پروتئین‌ها از روش بدون‌برچسب و با استفاده از نرم‌افزار sequest حاصل شده‌است. شکل ۴-۱ شمایی از داده‌ها را که در یک فایل Excel است را نمایش می‌دهد.

همچنین بانک‌های داده‌ای مناسب بسیار زیادی برای داده‌ها پروتئینی موجود است مانند:

- اطلس پروتئینی انسان (http://www.proteinatlas.org/) که داده‌های پروتئین آن با فن‌آوری ریز آرایه‌ها جمع‌آوری شده‌است. بانک داده‌ای شامل ۴۴بافت نرمال و ۲۰ بافت سرطانی، ۵۶ سل لاین و سطح بیان ترنسکریپت‌ها است.
- پایگاه داده مرجع پروتئین انسانی(http://www.hprd.org/) : که اطلاعات مناسبی در رابطه با تعاملات‌پروتئینی در یک شبکه دارد.
- مرکز ملی اطلاعات زیست فن‌آوری (https://www.ncbi.nlm.nih.gov/) : اطلاعات، مقالات و کتب بسیاری در رابطه با ژنومیکس و پروتئومیس دارا است.
- بانک‌داده پروتئین (http://www.rcsb.org/pdb/home/home.do)
- بانک‌داده شناسایی پروتئین‌ها (https://www.ebi.ac.uk)
- Proteopedia (http://proteopedia.org/)
- UniProt (http://www.uniprot.org/)

	A	B	C	D	E	F	G	H	I
1	Gene	Protein	Protein description	Peptide	charge	A1	A2	A9	B2
2	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	GWMNDPNGLWYDEK	2	63976296	24769764	40046584	5498
3	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	GWM[+16]NDPNGLW	2	6197049	2523306	12923109	104
4	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	VFWYEPSQK	2	8723820	8625410	7146495	496
5	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	WIMTAAK	2	1086354	1498491	6836681	602
6	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	WIM[+16]TAAK	2	177862	12706056	851060	777
7	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	IEIYSSDDLK	2	190592160	118888592	136130736	1684
8	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	IEIYSSDDLKSWK	3	742601	1987192	2445724	477
9	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	SSMSLVRK	2	210993	484660	3281242	54
10	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	KFSLNTEYQANPETELIN	3	8995346	22224170	23039502	563
11	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	FSLNTEYQANPETELINL	3	41705792	41687748	46593780	4379
12	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	AEPILNISNAGPWSR	2	5289871	6530384	16779556	239
13	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	GLEDPEEYLR	2	522095232	784777280	649137216	56294
14	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	MGFEVSASSFFLDR	2	4875871	5925656	7153578	575
15	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	MGFEVSASSFFLDR	3	75067	1966584	2582508	198
16	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	M[+16]GFEVSASSFFLD	2	1367760	1	1	
17	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	M[+16]GFEVSASSFFLD	3	3904164	9913573	19360182	628
18	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	MGFEVSASSFFLDRGN	3	1511042	5894734	4381281	235
19	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	ENPYFTNR	2	247100000	292735456	193823904	7914
20	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	MSVNNQPFK	2	4375857	81305216	2918126	2939
21	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	M[+16]SVNNQPFK	2	4616198	5216906	1125524	1895
22	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	MSVNNQPFKSENDLSY	3	3358167	3965633	3892572	429
23	SUC2	P00724	Invertase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SUC2 PE=1 SV=1	SENDLSYYK	2	153155616	173682416	130623936	6286
24	**SUC2 Total**					1274133118	1607309228	1311073297	92276
25	AKAP13	Q12802	A-kinase anchor protein 13 OS=Homo sapiens GN=AKAP13 PE=1 SV=2	STPSLPC[+57]M[+16]\	3	383171360	396120864	449970880	10815
26	AKAP13	Q12802	A-kinase anchor protein 13 OS=Homo sapiens GN=AKAP13 PE=1 SV=2	SVSIQNITGVGNDENM	3	25716	45718	14628	2
27	**AKAP13 Total**					383197076	396166582	449985508	10818
28	ECD	O95905	Protein ecdysoneless homolog OS=Homo sapiens GN=ECD PE=1 SV=1	EEKEQNYDLTEVSESM[	4	29705026	8415698	4520073	747
29	**ECD Total**					29705026	8415698	4520073	747

شکل۱-۴نمونه‌ای از داده‌های آزمایشگاه پروتئومیکس دانشگاه کالیفرنیا جنوبی

۴,۵ خلاصه فصل

در این فصـل ابتدا مروری بر مطالعات فصـل‌های گذشـته داشتیم و خلاصه‌ای از نتجیه‌گیری‌های خود را ارائه کردیم. سپس به بیان چالش‌ها و فرصـت‌ها پرداختیم و از لزوم پرداختن به مسئله کاهش بعد و نگاه کردن به سرطان از رویکردهای دیگر صحبت به میان آوردیم و چند موضوع برای مطالعات آتی پیشنهاد دادیم. بعد از آن به معرفی داده‌های در دسترس خود و چند بانک اطلاعاتی جامع و مناسب که برای تحقیقات آتی مناسب است را معرفی کردیم.

مراجع

مراجع فارسی

۱. تیمورپور، بابک؛ نجفی حیدر، ۱۳۹۴، "داده کاوی با R به همراه متن کاوی و تحلیل شبکه های اجتماعی". ویرایش ۱، تهران: مرکز تحقیقات و توسعه سازمان اتکا.

۲. علی پور، محمد. ۱۳۸۴،" ارائه روشی برای تشخیص بیماری سرطان به کمک داده کاوی پروتئوم انسان"، پایان‌نامه کارشناسی ارشد مهندسی برق-الکترونیک، دانشکده فنی مهندسی، دانشگاه تربت معلم سبزوار

۳. پارسا, ن. ۲۰۱۲. اساس سلولی و مولکولی سرطان در انسان، مقاله مروری. مجله سلول و بافت *(Cell & Tissue Journal),* 2**,** 365-376.

۴. شیردل, س. ا., عالمی, م. & خلیفه, خ. ۲۰۱۳. مباحث نوین در زیست‌شناسی: پروتئومیکس و نانوزیست‌فناوری. زیست فناوری دانشگاه تربیت مدرس, ۴, ۹-۲۲

مراجع انگلیسی

5. ADAM, B.-L., QU, Y., DAVIS, J. W., WARD, M. D., CLEMENTS, M. A., CAZARES, L. H., SEMMES, O. J., SCHELLHAMMER, P. F., YASUI, Y. & FENG, Z. 2002. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer research,* 62**,** 3609-3614.
6. COHEN, M., YOSSEF, R., EREZ, T., KUGEL, A., WELT, M., KARPASAS, M. M., BONES, J., RUDD, P. M., TAIEB, J. & BOISSIN, H. 2011. Serum apolipoproteins CI and C-III are reduced in stomach cancer patients: results from MALDI-based peptidome and immuno-based clinical assays. *PloS one,* 6**,** e14540.
7. DUBITZKY, W., GRANZOW, M. & BERRAR, D. P. 2007. *Fundamentals of data mining in genomics and proteomics*, Springer Science & Business Media.
8. DZIUDA, D .M. 2010. *Data mining for genomics and proteomics: analysis of gene and protein expression data*, John Wiley & Sons.
9. ELO, L. L. & SCHWIKOWSKI, B. 2012. Mining proteomic data for biomedical research. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 2**,** 1-13.

10. GALLEGO, M. & VIRSHUP, D. M. 2007. Post-translational modifications regulate the ticking of the circadian clock. *Nature Reviews Molecular Cell Biology,* 8**,** 139-148.
11. GARRETT, R. & GRISHAM, C. 2005. Biochemistry: Belmont, CA. *CA: Thomson Brooks/Cole,* 32005.
12. GUAN, W., ZHOU, M., HAMPTON, C. Y., BENIGNO, B. B., WALKER, L. D., GRAY, A., MCDONALD, J. F. & FERNÁNDEZ, F. M. 2009. Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines .*BMC bioinformatics,* 10**,** 259.
13. HAN, J., PEI, J. & KAMBER, M. 2011. *Data mining: concepts and techniques*, Elsevier.
14. HTIKE, Z. Z. & WIN, S. L. 2015. Premalignant Pancreatic Cancer Diagnosis Using Proteomic Pattern Analysis. *Journal of Medical and Bioengineering Vol,* 4.
15. JAGGA, Z. & GUPTA, D. 2015. Machine learning for biomarker identification in cancer research–developments toward its clinical application. *Personalized Medicine,* 12**,** 371-387.
16. JOERGER, A. C. & FERSHT, A. R. 2016. The p53 pathway: Origins, inactivation in cancer, and emerging therapeutic approaches. *Annual review of biochemistry,* 85**,** 375-404.
17. KATAJAMAA, M. & OREŠIČ, M. 2005. Processing methods for differential analysis of LC/MS profile data. *BMC bioinformatics,* 6**,** 179.
18. LAM, S., JIMENEZ, C. & BOVEN, E. 2014. Breast cancer classification by proteomic technologies: current state of knowledge. *Cancer treatment reviews,* 40**,** 129-138.
19. LE, L., CHI, K., TYLDESLEY, S., FLIBOTTE, S., DIAMOND, D. L., KUZYK, M. A. & SADAR, M. D. 2005. Identification of serum amyloid A as a biomarker to distinguish prostate cancer patients with bone lesions. *Clinical chemistry,* 51**,** 695-707.
20. LI, L., TANG, H., WU, Z., GONG, J., GRUIDL, M., ZOU, J., TOCKMAN, M. & CLARK, R. A. 2004. Data mining techniques for cancer detection using serum proteomic profiling. *Artificial intelligence in medicine,* 32**,** 71-83.
21. MEYFROIDT, G., GÜIZA, F., RAMON, J. & BRUYNOOGHE, M. 2009. Machine learning techniques to examine large patient databases. *Best Practice & Research Clinical Anaesthesiology,* 23-۱۲۷ , .۱۴۳
22. NEILSON, K. A., ALI, N. A., MURALIDHARAN, S., MIRZAEI, M., MARIANI, M., ASSADOURIAN, G., LEE, A., VAN SLUYTER, S. C. & HAYNES, P. A. 2011. Less label, more free: approaches in label-free quantitative mass spectrometry. *Proteomics,* 11**,** 535-553.
23. OH, J. H., LOTAN, Y., GURNANI, P., ROSENBLATT, K. P. & GAO, J. 2009. Prostate cancer biomarker discovery using high performance mass spectral serum profiling. *Computer methods and programs in biomedicine,* 96**,** 33-41.
24. OTTO, T. & SICINSKI, P. 2017. Cell cycle proteins as promising targets in cancer therapy. *Nature Reviews Cancer,* 17**,** 93-115.
25. PANIS, C. 2015. Proteomic Tools for Cancer Research: Updating the Oncoproteomics. *Journal of Proteomics & Bioinformatics***,** 1.
26. RALHAN, R., DESOUZA, L. V., MATTA, A., TRIPATHI, S. C., GHANNY, S., GUPTA, S. D., BAHADUR, S. & SIU, K. M. 2008. Discovery and verification of head-and-neck cancer biomarkers by differential protein expression analysis using iTRAQ labeling, multidimensional liquid chromatography, and tandem mass spectrometry. *Molecular & Cellular Proteomics,* 7**,** 1162-1173.
27. SHUKLA, S., GUPTA, D. L. & PRASAD, B. R. 2016. Comparative Study of Recent Trends on Cancer Disease Prediction using Data Mining Techniques. *International Journal of Database Theory and Application*.۱۱۸-۱۰۷ ,۹ ,
28. SINUES, P. M.-L., LANDONI, E., MICELI, R., DIBARI, V. F., DUGO, M., AGRESTI, R., TAGLIABUE, E., CRISTONI, S. & ORLANDI, R. 2015. Secondary electrospray ionization-mass spectrometry and a novel statistical bioinformatic approach identifies a cancer-related profile in exhaled breath of breast cancer patients: a pilot study. *Journal of breath research,* 9**,** 031001.

29. SOSA, M. S., BRAGADO, P. & AGUIRRE-GHISO, J. A. 2014. Mechanisms of disseminated cancer cell dormancy: an awakening field. *Nature Reviews Cancer,* 14**,** 611-622.
30. SWAN, A. L., MOBASHERI, A., ALLAWAY, D., LIDDELL, S. & BACARDIT, J. 2013. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *Omics: a journal of integrative biology,* 17**,** 595-610.
31. THOMAS, A., TOURASSI, G. D., ELMAGHRABY, A. S., VALDES, R. & JORTANI, S. A. 2006. Data mining in proteomic mass spectrometry. *Clinical Proteomics,* 2**,** 13.
32. USHIJIMA, M., MIYATA, S., EGUCHI, S., KAWAKITA, M., YOSHIMOTO, M., IWASE, T., AKIYAMA, F., SAKAMOTO, G., NAGASAKI, K. & MIKI, Y. 2007. Common peak approach using mass spectrometry data sets for predicting the effects of anticancer drugs on breast cancer. *Cancer informatics,* 3**,** 285.
33. VLAHOU, A., SCHORGE, J. O., GREGORY, B. W. & COLEMAN ,R. L. 2003. Diagnosis of ovarian cancer using decision tree classification of mass spectral data. *BioMed Research International,* 2003**,** 308-314.
34. WANG, J., YUE, S., YU, X. & WANG, Y. 2017. An efficient data reduction method and its application to cluster analysis. *Neurocomputing,* 238**,** 234-244.
35. WANG, J., ZUO, Y., MAN, Y.-G., AVITAL, I., STOJADINOVIC, A., LIU, M., YANG, X., VARGHESE, R. S., TADESSE, M. G. & RESSOM, H. W. 2015. Pathway and network approaches for identification of cancer signature markers from omics data. *Journal of Cancer,* 6**,** 54.
36. WASHAM, C. L., BYRUM, S. D., LEITZEL, K., ALI, S. M., TACKETT, A. J., GADDY, D., SUNDERMANN, S. E., LIPTON, A. & SUVA, L. J. 2013. Identification of PTHrP (12-48) as a plasma biomarker associated with breast cancer bone metastasis. *Cancer Epidemiology and Prevention Biomarkers,* 22**,** 972-983.
37. WATSON, J. D., BAKER, T., BELL, S., GANN, A., LEVINE, M. & LOSICK, R. 2003. *Molecular biology of the gene*, Pearson/Benjamin Cummings.
38. WONG, J. W. & CAGNEY, G. 2010. An overview of label-free quantitation methods in proteomics by mass spectrometry. *Proteome bioinformatics***,** 273-283.
39. YANG, J., ROY, R., JEDINAK, A. & MOSES, M. A. 2015. Mining the human proteome: biomarker discovery for human cancer and metastases. *The Cancer Journal*.۳۳۶-۳۲۷ ,۲۱ ,
40. ZHANG, F. & CHEN, J. Y. A neural network approach to multi-biomarker panel development based on LC/MS/MS proteomics profiles: A case study in breast cancer. Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on, 2009. IEEE, 1-6.
41. ZHANG, X., WANG, B., ZHANG, X.-S., LI, Z.-M., GUAN, Z.-Z. & JIANG, W.-Q. 2007. Serum diagnosis of diffuse large B-cell lymphomas and further identification of response to therapy using SELDI-TOF-MS and tree analysis patterning. *BMC cancer,* 7**,** 235.

پیوست‌ها

پیوست ((الف))

جدول نرم‌افزارهای مورد استفاده در داده‌کاوی داده‌های طیف سنج جرمی

نرم‌افزار	کاربرد	رایگان بودن	آدرس وبسایت
mzMine	شناسایی موجک، برچسب‌گذاری، ایزوتوپ زدایی	بله	http://mzmine.github.io/
موتور جستجوی توالی در بانک‌داده‌ها			
Mascot	تعیین پروتئین‌های موجود در نمونه همچنین شامل emPAI برای شناسایی پروتئین‌ها	خیر	http://www.matrixscience.com/
Sequest	تعیین پروتئین‌های حاضر در نمونه	خیر	http://fields.scripps.edu/sequest/
X!Tandem	تعیین پروتئین‌های حاضر در نمونه	بله	http://www.thegpm.org/tandem/
شناسایی پروتئین‌ها بدون برچسب‌گذاری			
emPAI	بدون برچسب که در Mascot هست	بله	http://www.matrixscience.com/
PepC	بدون برچسب	بله	http://sashimi.svn.sourceforge.net/viewvc/sashimi/trunk/trans_proteomic_pipeline/src/Quantitation/Pepc/
APEX	بدون برچسب گذاری، کمیت شماری قطعی	بله	http://pfgrc.jcvi.org/index.php/bioinformatics/apex.html
انتخاب ویژگی/ یادگیری ماشین			
WEKA	متدهای متنوع برای انتخاب ویژگی و دسته-بندی	بله	http://www.cs.waikato.ac.nz/ml/weka
کاربری‌های مختلف			
پیکج‌های مختلف برای R:	xcms MassSpecWavelet Bioconductor packages	بله	http://www.r-project.org/ http://bioconductor.org/packages/release/bioc/html/xcms.html http://bioconductor.org/packages/release/bioc/vignettes/MassSpecWavelet/inst/doc/MassSpecWavelet.pdf

پیوست ((ب))

فهرست واژگان انگلیسی به فارسی و بلعکس

English	فارسی
macromolecule	ابرمولکول
Peptide-mass fingerprint	اثر-انگشت پپتیدی
Peak picking	انتخاب موجک
Salivary amylose	آمیلاز بزاق
Antibody	آنتی بادی
antigen	آنتی ژن
Support vector machine	بردارهای ماشین پشتیبان
Nave bayes	بیز ساده
dehydration synthesis	پسابش
polypeptide	پلی‌پپتاید
Panel	پنل
translation	ترجمه
Protein-protein interaction	تعاملات بین پروتئین ها
Post translational modification(PTM)	تغییرات پس از ترجمه پروتئین ها
oncogenes	تنظیم کننده های مثبت
Protein expression profile	توصیف پروتئین های بیان شده
Molecular mass	جرم مولکولی
Random forest	جنگل تصادفی
mutation	جهش ژنتیکی
apoptosis	خودکشی سلولی
clustering	خوشه بندی
dalton	دالتون
Decision tree	درخت تصمیم
Rule-based classifiers	دسته بند مبتنی بر قاعده
classification	دسته بندی
angiogenesis	رگ زایی
In silico approach	رویکرد کامپیوتری در شناخت پروتئین
Leukemia	سرطان خون
Tumor suppressor	سرکوب کننده های تومور
Cell line	سل لاین

Artificial neural network	شبکه های عصبی مصنوعی
Mass spectrometry	طیف سنج جرمی
baseline reduction	کاهش خط مبنا
encode	کدگذاری
Expression protein matrix	ماتریس بیان پروتئین
Tissue	ماهیچه
biofluid	مایعات زیست پذیر
metastasis	متاستاز
Protein localization	مکان یابی پروتئین ها
Isoelectric point	نقطه ایزوالکتریک
Random sampling	نمونه برداری تصادفی
chemiluminescence	نورتابی شیمیایی
ontology	هستی شناسی پروتئین
Cross-validation	وارسی اعتبار

Abstract

Cancer is a very complex disease that occurs at the molecular level in the body. Cancers affected the form of cells, the patterns and the number of biomarkers, especially proteins that are in the tissue and the cell. In this context, proteomics, which examines the patterns and expression of proteins in the tissue and the cell, can be one of the keys to solving the cancer problem. Proteomics includes a variety of approaches and technologies in which the Expression Profiling proteomics approach and mass spectrometry technology can be used in conjunction with the use of high-level analytical methods such as data mining in the classification of cancer and the discovery of relevant biomarkers. In this report, we will discuss the approaches of proteomics and the technologies that used in it, and then focus the studies on expression profiling with the data obtained from the mass spectrometry technology. Subsequently, we will look at data mining and how to use data mining on data obtained from a mass spectrometer for the categorization and discovery of biomarkers, and, finally, we will have a review of past research.

Keys: data mining, mass spectrometry, cancer, biomarkers

Tarbiat Modares University

Faculty of Engineering

Information Technology Engineering Department

Seminar report

Application of proteome data mining in Cancer classification and biomarker discovery

Student

Rasoul norouzi

Supervisor

Dr.amir albadvi

July 2017