مجله دانشکده پزشکی اصفهان
سال سی و پنجم/شماره‌ی ۴۲۶/هفته‌ی دوم خرداد ماه ۱۳۹۶

تاریخ دریافت: ۱۳۹۵/۱۰/۲۹
تاریخ پذیرش: ۱۳۹۵/۱۱/۱۳

طبقه‌بندی و آنالیز شباهتی پایگاه مولکولی Binding-DB: بررسی کاربرد مدل‌های طبقه‌بندی چند منظوره برای استخراج قوانین تجمعی عمومی از پایگاه‌های بزرگ مولکولی

مرضیه مختاری[۱]، احمد مانی ورنوسفادرانی[۲]

مقاله پژوهشی

چکیده

مقدمه: در این مطالعه، با استفاده از ترکیبی از روش‌های کاهش بعد داده و طبقه‌بندی، ویژگی‌های مواد دارویی مورد بررسی قرار گرفت. تعریف و آماده‌سازی مولکول‌های «غیر فعال» برای توسعه‌ی مدل‌های تفکیکی دوتایی (Two-class classifier) یکی از مشکلات عمده در مسیر استفاده از مدل‌های تفکیکی بر پایه‌ی لیگاند در روند طراحی سیستماتیک دارو می‌باشد. از این رو، با استفاده از مولکول‌های «فعال» موجود در پایگاه Binding-DB، به توسعه‌ی مدل‌های تفکیکی چند متغیره‌ی چند منظوره پرداخته شد.

روش‌ها: به این منظور، در حدود ۱۶۰۳۷۲ ریز مولکول برای ۴۵ هدف دارویی مختلف از پایگاه مولکولی Binding-DB دانلود شد و پس از بهینه‌سازی ساختار، ۱۴۹۷ ویژگی فیزیکی و شیمیایی برای هر مولکول استخراج گردید. با استفاده از الگوریتم Apriori و ترکیب آن با روش طبقه‌بندی تفکیکی خطی (Linear discriminant analysis)، ویژگی‌های مولکولی برای هر هدف دارویی به منظور تفکیک مولکول‌های فعال استخراج شد.

یافته‌ها: در نهایت، با استفاده از غربالگری مجازی در پایگاه داده‌های مولکولی Zinc و Binding-DB و محاسبه‌ی سطح زیر نمودار Receiver operating characteristic (ROC) صحت و حساسیت طبقه‌بندی مورد بررسی قرار گرفت. میزان سطح زیر نمودار ROC برای هر بهینه‌سازی پایگاه Zinc به طور میانگین برابر با ۰/۱۴۹۵ ± ۰/۸۳۴۱ و در پایگاه Binding-DB به طور میانگین برابر با ۰/۱۵۰۲ ± ۰/۸۶۱۵ بود.

نتیجه‌گیری: می‌توان با استفاده از الگوریتم ارایه شده، ویژگی‌هایی برای هر دسته از ریز مولکول‌های مرتبط با هر هدف دارویی استخراج کرد و پایگاه‌های مولکولی مختلف را برای هر هدف دارویی بهینه‌سازی کرد. سطح زیر نمودار ROC برای دو پایگاه مولکولی مورد بررسی نشان می‌دهد که روش ارایه شده، روش مفیدی برای طبقه‌بندی پایگاه‌های بزرگ مولکولی بدون استفاده از ریز مولکول‌های غیر فعال می‌باشد.

واژگان کلیدی: غربالگری مجازی، طبقه‌بندی چند منظوره، بررسی داده، طبقه‌بندی پایگاه داده، لیگاند

ارجاع: مختاری مرضیه، مانی ورنوسفادرانی احمد. **طبقه‌بندی و آنالیز شباهتی پایگاه مولکولی Binding-DB: بررسی کاربرد مدل‌های طبقه‌بندی چند منظوره برای استخراج قوانین تجمعی عمومی از پایگاه‌های بزرگ مولکولی.** مجله دانشکده پزشکی اصفهان ۱۳۹۶؛ ۳۵ (۴۲۶): ۴۰۵-۴۰۰

مقدمه

در سال‌های اخیر، منابع تولید دارو به طور کامل تغییر کرده است. از سال ۱۹۷۰ به بعد نیز نقش منابع تجربی در کشف مواد دارویی کاهش یافته است. از جمله روش‌هایی که منجر به کشف دارو می‌شود، روش‌های غربالگری مجازی و استفاده از ویژگی‌های قابل محاسبه می‌باشد. قبل از سال ۱۹۸۹، مشاهده‌ی صدها هزار ملکول همراه با فعالیت‌های زیستی و کشف مواد با قابلیت دارو شدن (Lead) از نظر عملی و کاری غیر ممکن بوده است. با ظهور High throughput screening (HTS) در طی دوره‌ی زمانی ۱۹۹۱-۱۹۸۹، مشاهده‌ی صدها مولکول در بین هزاران مولکول موجود، امکان‌پذیر شد. از آن پس، با ترکیب علم شیمی با روش‌های آنالیز و دسته‌بندی داده‌ها، کشف مواد دارویی توسعه یافت و امکان مشاهده‌ی مواد دارویی در حجم انبوهی از مولکول‌ها میسر شد.

قوانین کلی در جستجوی فضای شیمیایی و کشف مواد دارویی به صنعت داروسازی کمک شایانی می‌کند. شیمیدان مواد دارویی Lipinski و همکاران در سال ۱۹۹۷ با بررسی ۲۰۰۰ مواد دارویی و غیر دارویی که قابلیت واکنش با پروتئین‌ها را داشتند، موفق به معرفی

۱- دانشجوی کارشناسی ارشد، گروه مهندسی پزشکی و کمیته‌ی تحقیقات دانشجویی، دانشکده‌ی فن‌آوری‌های نوین علوم پزشکی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران
۲- استادیار، گروه شیمی تجزیه، دانشکده‌ی علوم پایه، دانشگاه تربیت مدرس، تهران، ایران
نویسنده‌ی مسؤول: مرضیه مختاری
Email: mokhtarymarzieh@yahoo.com

قوانین کلی در خواص مواد شدند. این قوانین به قوانین پنج‌گانه‌ی Lipinski معروف شد (۱). همیشه داروهایی که از قوانین پنج‌گانه‌ی Lipinski پیروی می‌کنند، قابل استفاده نیستند و این امر، به میزان دز مصرفی دارو بستگی دارد. برای این که یک دارو قابل مصرف باشد، باید وزن مولکولی کم، تعداد اتم‌های دهنده و گیرنده‌ی هیدروژن پایین و حلقه‌ی قابل گردش کمی داشته باشد. به این ترتیب، جستجو در کتابخانه‌های با وزن مولکولی پایین، از نظر تئوری بیشتر منجر به کشف دارو می‌شود.

در طول سال‌های اخیر، با پیشرفت در زمینه‌ی طراحی دارو و روش‌های کموژنومیک، تلاش‌های زیادی برای کشف قوانین ساده ارایه شده است. برای مثال Ghose و همکاران در سال ۱۹۹۹ با بررسی پایگاه مولکولی Comprehensive medicinal chemistry (CMC) این قوانین را گسترش دادند (۲). در بین سال‌های ۲۰۱۰-۲۰۰۰، صورت‌های جدیدتری از خواص مواد دارویی برای ترکیبات شیمیایی توسعه یافته‌اند (۴-۳).

از طرف دیگر، کامل نبودن قوانین ذکر شده در بسیاری از نمونه‌ها توسط محققین نشان داده شد و به همین منظور، برای توسعه‌ی قوانین ساده، کارآمد و جدید در این زمینه تلاش‌های بسیاری شده است و یکی از موضوعات بسیار با اهمیت در زمینه‌ی طراحی دارو می‌باشد (۵).

در سال ۲۰۰۷، محققین در زمینه‌ی آنالیز داده‌های مولکولی به بررسی ۲/۷ میلیون مولکول از ۲۳ پایگاه دارویی مختلف پرداختند و نشان دادند که تعدادی از قوانین Lipinski و همچنین قوانین بسط داده شده‌ی آن، برای ترکیبات دارویی موجود در این کتابخانه‌ها صدق نمی‌کند (۶). این موضوع، نشان می‌دهد که روندهای ناشناخته‌ای از روابط پیچیده‌ی ساختار- فعالیت در داده‌های مولکولی وجود دارد. در سال ۲۰۱۵ نیز روشی برای استخراج قوانین با استفاده از روش‌ها و مدل‌های ریاضی ارایه گردید (۷).

هدف از انجام این پژوهش، ارایه‌ی مدلی برای مولکول‌های هدف مربوط به هر پروتئین خاص در پایگاه مولکولی Binding-DB بود. با استفاده از مدل ارایه شده، می‌توان بدون نیاز به مولکول‌های غیر فعال در پایگاه‌های مولکولی همانند پایگاه داده‌ی Zinc و Binding-DB غربالگری مجازی انجام داد و برای هر هدف پروتئین خاص از بین هزاران مولکول در این پایگاه مولکولی، ریز مولکول‌هایی با قابلیت برهم‌کنش با پروتئین مورد نظر ارایه نمود. بررسی یک سری از قوانین عمومی از پایگاه داده‌ی مولکولی Binding-DB، به صنعت داروسازی کمک زیادی می‌کند. این قوانین عمومی، ماهیت لیگاندهایی را که خواص دارویی دارند، با بررسی ویژگی‌های قابل محاسبه مانند خواص فیزیکی، معرفی می‌کند. مزیت روش مورد مطالعه، عدم نیاز به مولکول‌های غیر فعال برای شناسایی مولکول‌های فعال برای هر هدف پروتئینی خاص در پایگاه‌های مولکولی می‌باشد.

روش‌ها

در این پژوهش از ریز مولکول‌های مربوط به ۴۵ دسته از پرکاربردترین پروتئین موجود در پایگاه مولکولی Binding-DB (۸) استفاده شد. داده‌ها در این پایگاه داده به صورت فایل‌های Structure data file (SDF) وجود دارند که محتوای آن‌ها به صورت مختصات اتم‌های مولکول‌ها در فضای دکارتی است.

لازم به ذکر است که در این پایگاه داده، اطلاعات فعالیت مولکول‌ها نیز به همراه مختصات فضایی اتم‌های مولکول‌ها وجود دارند. در ابتدا با استفاده از نرم‌افزار OpenBabel فایل‌های SDF به فایل‌های Hyperchem HIN chemical modeller input file (HIN) تبدیل می‌گردد. علت این تبدیل ساختار، این است که برای محاسبه‌ی توصیف کننده‌های مولکولی، لازم است بار جزیی اتم‌ها در مولکول‌ها مشخص باشد. برای محاسبه‌ی این بارهای جزیی، از نرم‌افزار HyperChem استفاده شد که ورودی این نرم‌افزار، فایل‌هایی با ساختار HIN می‌باشد. با توجه به ریز مولکول‌های مربوط به ۴۵ دسته پروتئین، ۱۶۰۳۷۲ فایل با ساختار HIN به دست آمد.

در مرحله‌ی بعدی، با استفاده از نرم‌افزار Dragon برای هر مولکول (فایل) تعداد ۱۴۹۷ توصیف کننده محاسبه شد. پس از اتمام محاسبات، یک ماتریس ۱۴۹۷ × ۱۶۰۳۷۲ به دست آمد. به طور کلی، این ماتریس متعلق به ۴۵ گروه (Class) متفاوت می‌باشد که هر گروه، با توجه به میزان فعالیت با پروتئین مربوط به دو دسته‌ی فعال و غیر فعال تفکیک شد.

در این پژوهش، با استفاده از الگوریتم Apriori (۹) و ترکیب آن با روش طبقه‌بندی تفکیکی خطی Linear discriminant analysis (LDA) برای هر پروتئین مورد مطالعه، مدلی برای شناسایی مولکول‌های فعال ارایه گردید. به این ترتیب، با استفاده از این مدل‌ها، روشی برای جستجو در پایگاه‌های مولکولی بدون نیاز به شناسایی مولکول‌های غیر فعال پیشنهاد شد.

الگوریتم Apriori، یکی از روش‌های پیش انتخاب توصیف کننده‌ها بر مبنای روش کاوش مجموعیت (Association Mining) در دسته‌ی داده‌های بسیار بزرگ است و مزیت عمده‌ی آن، سرعت این الگوریتم می‌باشد. هدف از این الگوریتم، پیدا کردن مجموعه‌ای از خواص پایگاه داده است که به طور مکرر تکرار می‌شود و همچنین، یافتن عناصری که به عنوان اصول، بقیه‌ی زیر مجموعه را تحت تأثیر قرار می‌دهد. الگوریتم Association rule mining

(ARM)، با استفاده از Apriori که در سال ۱۹۹۳ توسط Agrawal کشف شد (۹) و در سال ۲۰۰۳ بهبود یافت (۱۰).

یکی از موارد قابل بررسی در کموانفورماتیک (Chemoinformatics) انتخاب ویژگی‌ها به طور بهینه است. ابعاد بسیار بالای فضای ویژگی در مقایسه با تعداد کمتری از ورودی‌ها، نیاز به انتخاب قسمتی از فضای توصیف کننده را افزایش می‌دهد. همچنین، انتخاب ویژگی بهینه، سرعت و نیز صحت دسته‌بندی داده‌ها را افزایش می‌دهد. در انتخاب ویژگی، می‌توان زیر مجموعه‌ای از ویژگی‌ها را بدون در نظر گرفتن سایر متغیرها به عنوان توصیف کننده‌ی برتر، انتخاب کرد که می‌تواند شامل روش‌های Filtering باشد و یا فضای متغیرها را رتبه‌بندی (Feature ranking) کرد. در این روش، انتخاب متغیرهایی با بیشترین رتبه‌ی اهمیت و کمترین درصد شباهت با دیگر متغیرهای انتخاب شده، مد نظر است.

الگوریتم Apriori، به دنبال قوانینی برای کاهش ابعاد ویژگی‌ها می‌باشد؛ به گونه‌ای که اطلاعات طبقه‌بندی داده‌ها محفوظ بماند. با مشخص کردن ویژگی‌هایی که نتایج مشترک را تولید می‌کنند، می‌توان قوانین ارجحی برای دسته‌بندی داده‌ها به دست آورد. به این ترتیب، می‌توان متغییر y را به عنوان نماینده‌ای از ویژگی x انتخاب کرد؛ به نحوی که اطلاعات طبقه‌بندی با انتخاب متغیر y بهینه باشد و این انتخاب متغیر، برای ترکیب‌های دوتایی و چند تایی نیز به طور مجزا اعمال می‌شود. در حالی که در سایر روش‌های انتخاب ویژگی، نمی‌توان به صورت متمرکز برهم‌کنش‌های دوتایی و یا سه‌تایی را به صورت مجزا بررسی نمود.

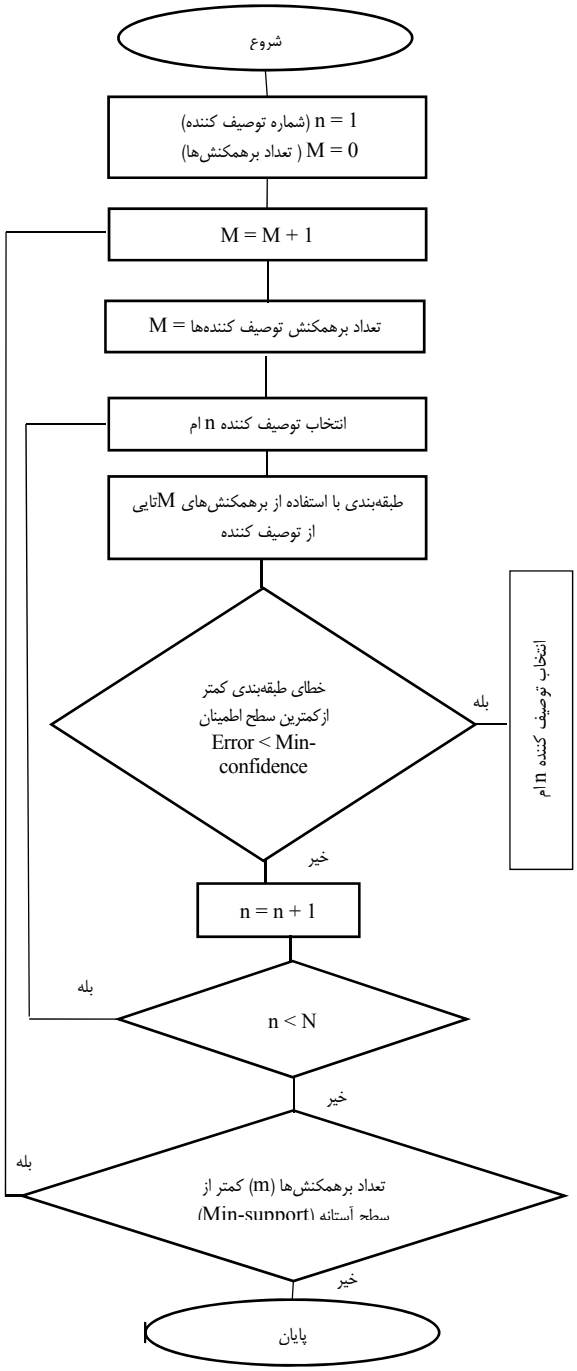
روشی که به عنوان معیار برای انتخاب توصیف کننده‌ها از آن استفاده شد، روش تفکیک متغیر خطی (LDA) بود. روش LDA یکی از روش‌های آنالیز تفکیک خطی است و در روش‌های خوشه‌بندی جزء روش‌های نظارت شده است (۱۱).

در هر مرحله، با ترکیبی از توصیف کننده‌ها، توصیف کننده‌هایی که کمترین خطای طبقه‌بندی را ایجاد می‌کنند، انتخاب شدند. در شکل ۱، روند الگوریتم نمایش داده شده است.

یافته‌ها

همان‌طور که بیان شد، ابتدا ویژگی‌های متناظر برای طبقه‌بندی مولکول‌های هر هدف دارویی استخراج گردید و سپس، برای بررسی قابلیت تفکیک لیگاندهای فعال با هر پروتئین در پایگاه مولکولی، از سطح زیر نمودار Receiver operating characteristic (ROC) استفاده شد.

ابتدا، لیگاندهای فعال هر پروتئین با پایگاه مولکولی مربوط در کنار هم در نظر گرفته شدند؛ به نحوی که نظم خاصی به منظور تفکیک مولکول‌های فعال از پایگاه مولکولی وجود نداشته باشد و سپس، ریز مولکول‌ها بر مبنای ویژگی‌های استخراج شده در مرحله‌ی طبقه بندی مرتب می‌شود و هر چه این لیگاندها در ابتدای پایگاه مولکولی مرتب شده یافت شوند، توانایی ویژگی‌های استخراج شده در مرحله ی قبل به منظور تشخیص لیگاندهای فعال در پایگاه مولکولی بیشتر می‌باشد.



شکل ۱. معرفی روش انتخاب ویژگی Apriori با استفاده از مدل "Linear Discriminant Analysis"

در پایان، با رسم نمودار ROC بر مبنای تفکیک لیگانـدهای فعـال، هر چه مقدار AUC (Area under curve) به مقـدار یـک نزدیـک‌تـر باشد، ریز مولکول‌های فعال بهتر تفکیک شـده‌انـد و توانـایی الگـوریتم بهینه است و در این جا، به دلیل حجم بالایی از اطلاعات که مربوط بـه ویژگی‌های شیمیایی فیزیکی و توپولوژیکی لیگاندهای فعـال مـی‌باشـد، تنها به ارایه‌ی مقدار AUC نمودار مربوط و تعداد ویژگی‌های اسـتخراج شده‌ی مرتبط با هر پروتئین اکتفا می‌شود (جدول ۱).

مقـــدار AUC بیشـــتر از ۰/۸ در بـــیش از ۸۰ درصـــد از ۴۵ رده‌ی مختلف، عملکـرد بهینـه‌ی روش مـورد نظـر در اسـتخراج ویژگی از لیگاندهای فعال را نشان می‌دهد.

جدول ۱. مقادیر Area under curve (AUC) غربالگری مجازی لیگاندهای مرتبط با اهداف دارویی در پایگاه مولکولی ZINC و Binding-DB با استفاده از ویژگی های استخراج شده از الگوریتم Apriori-LDA

ردیف	نام پروتئین‌های هدف	میزان AUC در پایگاه binding-DB	میزان AUC در پایگاهZinc	تعداد توصیف کننده‌ها
X1	5-HT2_3D	۰/۸۰۶۰	۰/۹۲۷۳	۲۴
X2	BETA-HSD1	۰/۹۲۳۷	۰/۹۶۹۹	۲۵
X3	ACETYLCHOLINESTERASE	۰/۸۴۱۳	۰/۹۵۷۷	۲۷
X4	ADENOSINE	۰/۹۲۰۶	۰/۶۴۹۹	۲۰
X5	ADORA	۰/۸۰۹۹	۰/۹۶۶۲	۱۹
X6	ALPHA-1B-1D_ADRENERGIC	۰/۸۶۶۰	۰/۵۱۳۹	۱۹
X7	ANGIOTENSIN	۰/۸۸۹۹	۰/۹۲۱۱	۳۱
X8	BETA-SECRETASE	۰/۹۳۳۷	۰/۸۹۴۲	۲۹
X9	BUTYRYLCHOLINESTERASE	۰/۹۳۵۹	۰/۹۵۵۵	۴۱
X10*	CANNABINOID_RECEPTOR	۰/۸۵۰۳	۰/۹۷۷۲	
X11	CARBONIC_ANHYDRASE	۰/۹۴۷۳	۰/۹۷۶۲	۵
X12	CATHEPSIN	۰/۸۸۹۷	۰/۷۶۶۲	۴۰
X13	CATHEPSIN_ PREPROPROTEIN	۰/۹۱۳۵	۰/۹۱۳۵	
X14*	C-C_CHEMOKINE_RECEPTOR	۰/۹۰۲۰	۰/۹۶۸۳	
X15*	CYCLOOXYGENASE	۰/۹۱۶۶	۰/۹۵۴۲	۴۲
X16*	CYTOCHROME	۰/۷۸۴۸	۰/۴۸۴۲	۴۴
X17*	DELTA_OPIOID_RECEPTOR	۰/۹۱۸۹	۰/۹۶۳۰	۶۴
X18	DIHYDROFOLATE_REDUCTASE	۰/۹۷۳۳	۰/۹۸۳۹	۴۵
X19	DIPEPTIDYL_PEPTIDASE	۰/۹۲۷۸	۰/۷۱۶۸	۵۳
X20	DOPAMINE	–	۰/۷۲۶۵	۲۰
X23	HTR1A_3D-HTR2A	۰/۷۳۲۱	۰/۷۱۱۹	۱۰
X24*	MAP_KINASE_P38_ALPHA			
X25	MU_OPIOID_RECEPTOR	۰/۹۰۸۱	۰/۷۸۵۸	۳۵
X26	NOREPINEPHRIN_TRANSPORTER	۰/۹۲۱۹	۰/۹۲۴۵	۲۶
X28	PEROXISOME_PROLIFERATOR	۰/۹۳۳۰	۰/۸۸۲۹	۳۱
X29	PHOSPHOINOSITIDE	۰/۸۹۸۴	۰/۷۸۰۳	۳۴
X30	PROGESTERONE_RECEPTOR	۰/۹۴۹۴	۰/۹۶۴۷	۳۴
X31	TYROSINE_PHOSPHATASE	۰/۹۵۸۷	۰/۷۸۰۹	۲۷
X32	PROTEIN_FARNESYLTRANSFE	۰/۸۷۱۲	۰/۸۸۰۵	۲۵
X33*	PROTEIN_KINASE	۰/۹۱۱۱	۰/۹۱۷۵	۲۰
X35	SEROTONIN	۰/۸۰۱۸	۰/۹۱۰۵	۱۸
X36	SIGMA_OPIOID	۰/۹۰۰۶	۰/۸۳۵۰	۵
X37	SLC6-MERGE	۰/۹۰۱۱	۰/۷۴۳۷	۳۱
X38	STREPTOKINASE_A_PRECURSOR	۰/۸۰۱۲	۰/۷۲۰۷	۳۱
X39	THROMBIN	۰/۸۶۷۶	۰/۸۷۳۷	۳۰
X4	TYROSINE-PROTEIN_KINASE	۰/۸۹۱۸	۰/۶۶۹۸	۳۱
X41*	VEGFR-2__KDR	۰/۸۵۷۹	۰/۳۵۶۳	۲۲
X42*	DOPAMINE_RECEPTOR	۰/۸۶۰۹	۰/۷۱۳۲	۲۲
X43*	DOPAMINE_TRANSPORTER	۰/۹۱۱۹	۰/۹۲۶۸	۲۷
X44*	SEROTONIN_RECEPTOR	۰/۷۷۵۶	۰/۸۵۸۷	۲۰
X45*	SEROTONIN_TRANSPORTER	۰/۸۴۵۰	۰/۹۴۲۱	۴۵

* نشانگر دسته‌ای از پروتئین‌ها که نماینده‌ی کل پروتئین‌ها از لحاظ زیستی هستند.

AUC: Area under curve

بحث

همان‌گونه که مطرح شد، با استفاده از روش ترکیبی Apriori-LDA نتایج به این صورت ارایه می‌شود که ابتدا توصیف کننده‌های متناظر با مولکول‌های فعال هر پروتئین به منظور مجزاسازی در مقابل سایر مولکول‌های فعال موجود در پایگاه Binding-DB انتخاب می‌شوند. ویژگی متمایز روش ارایه شده در مجزا کردن مولکول‌های مرتبط با یک هدف دارویی از سایر مولکول‌های موجود بر اساس برهم‌کنش‌های مختلف بین توصیف کننده‌ها می‌باشد و به این ترتیب، توصیف کننده‌هایی را که به تنهایی قابلیت تفکیک دو رده‌ی مختلف ندارند، اما با برهم‌کنش‌های دوتایی و سه‌تایی این تفکیک را انجام می‌دهند، انتخاب می‌گردند و از این توصیف کننده‌ها، به عنوان مشخصه‌ی لیگاندهای فعال یک پروتئین خاص برای جستجو در پایگاه‌های اطلاعاتی مولکولی بزرگ شبیه ZINC و پایگاه مولکولی Binding-DB استفاده می‌شود.

لازم به ذکر است طبقه‌بندی هر دسته از لیگاندها در برابر همه‌ی مولکول‌های فعال، به زمان زیاد نیازمند است و همچنین، عدم تقارن زیاد بین دو رده (لیگاندهای فعال پروتئین مورد نظر- لیگاندهای فعال سایر پروتئین‌ها) طبقه‌بندی مناسبی ارایه نمی‌دهد. در نتیجه، از دسته‌ای از پروتئین‌ها که نماینده‌ی کل پروتئین‌ها از لحاظ زیستی هستند (در جدول ۱ با علامت * نمایش داده شده است)، استفاده می‌شود. می‌توان از ویژگی‌های استخراج شده برای هر هدف دارویی به عنوان توصیف کننده‌ایی برای شناسایی ریز مولکول‌های دارویی در مراحل اولیه‌ی داروسازی استفاده کرد و به این ترتیب، نتایج این پژوهش به عنوان روش مؤثر در سرعت بخشیدن به شناسایی مواد دارویی کاربرد دارد.

تشکر و قدردانی

از مرکز پردازش سیگنال و سنسور به خاطر حمایت مالی از اجرای این مطالعه، قدردانی می‌گردد. همچنین، لازم است از کمیته‌ی پژوهشی دانشکده‌ی فن‌آوری‌های نوین علوم پزشکی سپاسگزاری شود.

References

1. Palanisam SK. Association rule based classi cation [MSc Thesis]. Worcester, MA: Worcester Polytechnic Institute; 2006.
2. Vaidya J, Clifton C. Privacy preserving association rule mining in vertically partitioned data. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2002 Jul 23-25; Edmonton, AB, Canada. New York, NY; ACM; p. 639-44.
3. Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases; 1994 Sep 12-15; Santiego, Chile. San Francisco, CA: Morgan Kaufmann Publishers Inc; p. 487-99.
4. Chen X, Lin Y, Liu M, Gilson MK. The Binding Database: data management and interface design. Bioinformatics 2002; 18(1): 130-9.
5. Chuprina A, Lukin O, Demoiseaux R, Buzko A, Shivanyuk A. Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. J Chem Inf Model 2010; 50(4): 470-9.
6. Li AP. Preclinical in vitro screening assays for drug-like properties. Drug Discov Today Technol 2005; 2(2): 179-85.
7. Camp D, Davis RA, Campitelli M, Ebdon J, Quinn RJ. Drug-like properties: guiding principles for the design of natural product libraries. J Nat Prod 2012; 75(1): 72-81.
8. Hou T, Wang J, Li Y. ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. J Chem Inf Model 2007; 47(6): 2408-15.
9. Ghose AK, Viswanadhan VN, Wendoloski JJ. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. J Comb Chem 1999; 1(1): 55-68.
10. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 2001; 46(1-3): 3-26.
11. Mani-Varnosfaderani A, Valadkhani A, Jalali-Heravi M. CS-MINER: A tool for association mining in binding-database. Mol Inform 2015; 34(4): 185-96.

Journal of Isfahan Medical School
Received: 18.01.2017
Vol. 35, No. 426, 2nd Week, June 2017
Accepted: 01.02.2017

Classification and Similarity Analysis of Binding-Database: A Survey on Application of Multi-Class Classifiers for Deriving General Rules from Large Compound Databases

Marzieh Mokhtari[1], Ahmad Mani-Varnosfaderani[2]

Original Article

Abstract

Background: In this research, we extracted and modified features of active ligands related to specific biological targets with combination of data mining and classification methods to aid medicinal chemists in their drug discovery projects. Preparing an inactive ligand is the major problem for development of multi-class classifiers. Therefore, our models were developed based on only active ligands found in Binding-database (DB) without any needs for preparing inactive molecules.

Methods: Our database consisted of 160372 ligands in 45 classes of common proteins and 1497 different features (topological, chemistry, physical, etc.) were calculated for each molecule. Then, the specific features of active ligands of any target were extracted based on combination of linear discriminate analysis and Apriori algorithm.

Findings: Receiver operating characteristic (ROC) was a useful operator to analysis the accuracy and sensitivity of classification models and retrieving molecules from ZINC and Binding-DB databases. Area under curve (AUC) of this diagram was evaluated for analysis of each target in Zinc and Binding-DB and their results were 0.8341 ± 0.1495 and 0.8615 ± 0.1502, respectively.

Conclusion: Specific features of active ligands could be found using the methodology described in this work and with these features, we can sort each database based on corresponding target. AUC shows that the present method is useful for virtual screening in big databases without survey on inactive ligands.

Keywords: Virtual systems, Multiple classification, Data mining, Database management systems, Ligands

Citation: Mokhtari M, Mani-Varnosfaderani A. **Classification and Similarity Analysis of Binding-Database: A Survey on Application of Multi-Class Classifiers for Deriving General Rules from Large Compound Databases.** J Isfahan Med Sch 2017; 35(426): 400-5.

1- MSc Student, Department of Biomedical Engineering AND Student Research Committee, School of Advanced Medical Technology, Isfahan University of Medical Sciences, Isfahan, Iran
2- Assistant Professor, Department of Chemistry, School of Basic Sciences, Tarbiat Modares University, Tehran, Iran
Corresponding Author: Marzieh Mokhtari, Email: mokhtarymarzieh@yahoo.com