مجله علوم پزشکی دانشگاه آزاد اسلامی
دوره ۲۸، شماره ۳، پاییز ۹۷، صفحات ۱۸۱ تا ۱۹۴

Original Article

پیش بینی فعالیت ضدسرطانی مشتقات ۱و۸- نفتیریدین توسط روش الگوریتم ژنتیک-رگرسیون خطی چندگانه

شهین احمدی[۱]، روح الله خانی[۲]، مریم مقدس[۳]

[۱] استادیار دانشگاه گروه شیمی، واحد کرمانشاه، دانشگاه آزاد اسلامی، کرمانشاه، ایران
[۲] دانشجوی ارشد دانشگاه - گروه شیمی دارویی، علوم پزشکی تهران ، دانشگاه آزاداسلامی، تهران، ایران
[۳] استادیار دانشگاه - گروه شیمی، واحد صفادشت، دانشگاه آزاد اسلامی، صفادشت، ایران

چکیده

سابقه و هدف: *این مطالعه به مقایسه مدل سازی QSAR فعالیت ضد سرطانی ترکیبات ۱و۴-دی هیدرو-۴-اکسو-۱-(۲-تیازولیل)-۱و۸-نفتیریدین و مشتقات آن با روش رگرسیون خطی چندگانه مرحله ای (S-MLR) و روش الگوریتم ژنتیک-رگرسیون خطی چندگانه (GA-MLR) پرداخت.*

روش بررسی: *مجموعه ای از ۱۰۰ ترکیب با فعالیت ضد سرطانی مشخص از مقاله معتبر بین المللی انتخاب شد و روش میدان نیروی آلینجر MM2 برای کمینه کردن انرژی مولکول‌ها استفاده شد. ساختار هندسی مولکول‌ها از طریق روش کوانتوم نیمه تجربی روش آوستین با استفاده از الگوریتم پلاک-ریبایر (Polak-Ribiere) با استفاده از نرم افزار موپک بهینه سازی شدند. تعداد زیادی از توصیفگرهای تئوری برای هر مولکول با استفاده از نرم افزار دراگون محاسبه شد. به منظور انتخاب بهترین دسته از توصیفگرها برای مدل سازی QSAR از دو روش انتخاب متغیر ترکیب الگوریتم ژنتیک-رگرسیون خطی چندگانه و رگرسیون خطی چندگانه مرحله‌ای استفاده شد. برای مدل‌سازی ابتدا نمونه برداری تصادفی دسته آموزش (۸۰ درصد از داده‌ها) ۲۰ بار به صورت تصادفی صورت گرفته و مولکول‌های باقیمانده (۲۰ درصد باقیمانده از داده‌ها) به عنوان دسته پیشگویی برای اعتبارسنجی خارجی استفاده شدند. در میان نمونه‌های تصادفی، یکی از نمونه‌ها با بالاترین Q^2CV، Q^2cal و Q^2test به عنوان بهترین دسته یادگیری و آموزش انتخاب شد. با استفاده از این دسته یادگیری هر بار مدل به دو روش S-MLR و GA-MLR ایجاد شد.*

یافته‌ها: *مدل‌های QSAR به دست آمده با GA-MLR مجذور ضریب همبستگی اعتبارسنجی بزرگتری نسبت به روش S-MLR داشتند.*

نتیجه‌گیری: *نتایج این مقایسه نشان می‌دهد که می‌توان با استفاده از مدل حاصل، فعالیت ترکیبات ضد سرطانی مشابه را پیشگویی کرد.*

واژگان کلیدی: *مدل‌سازی QSAR، فعالیت ضدسرطان، انتخاب متغیر، GA-MLR، Stepwise-MLR*

مقدمه

یکی از مشکلاتی که جامعه بشری همیشه با آن روبه‌رو است، مقابله با انواع بیماری‌هایی است که سلامت انسان‌ها را به مخاطره انداخته و همواره یکی از مهم‌ترین دغدغه‌های محققان یافتن داروهای مؤثر، برای رفع این معضل و کاهش عوارض این بیماری‌ها بوده است. بروز انواع بیماری‌ها از قبیل سرطان، بیماری هولناک ایدز، بیماری‌های مشترک انسان و دام و مقاوم شدن ویروس‌ها در برابر آنتی‌بیوتیک‌ها همه از جمله مواردی هستند که ذهن دانشمندان را در جهت یافتن داروهای مؤثر و کارآمد برای مقابله با این بیماری‌ها معطوف خود کرده‌اند (۱).

روندی که در گذشته منجر به کشف داروهای جدید می‌شد به روش آزمون و خطا صورت می‌گرفت که روش وقت‌گیر و هزینه‌بر است. مشکل دیگری که در این راه دانشمندان را آزار می‌دهد عدم اطلاع آن‌ها از فعالیت دارویی ترکیبات، قبل از

آدرس نویسنده مسئول: کرمانشاه، دانشگاه آزاد کرمانشاه، گروه شیمی تجزیه، **شهین احمدی**
(**email**: ahmadi.chemometrics@gmail.com)
تاریخ دریافت مقاله: ۹۶/۹/۲۷
تاریخ پذیرش مقاله: ۹۶/۱۱/۲۹

انجام سنتز و بررسی تجربی آن‌ها است و به‌همین دلیل یکی از مهم‌ترین اهداف شیمیدان‌ها و محققان دارویی پیش‌بینی فعالیت ترکیبات، قبل از سنتز و یا انجام آزمایش بر روی آن‌ها است. چرا که انجام بسیاری از آزمایشات مستلزم صرف زمان و هزینه‌های زیادی است. از این‌رو نیاز به استفاده از روش‌های تئوری و محاسباتی که بدون انجام آزمایش بتواند ویژگی و یا فعالیت ترکیبات را پیشگویی کند ضروری به نظر می‌رسد. ظهور علم کمومتری توانسته راه‌حلی برای رفع این مشکلات باشد (۴-۲).

روابط کمی ساختار – فعالیت (QSAR) برای چندین دهه در توسعه روابط بین خواص فیزیکی ترکیبات شیمیایی و فعالیت بیولوژیکی آنها برای به دست آوردن یک مدل آماری قابل اعتماد برای پیش بینی فعالیت ترکیبات جدید استفاده شده است. یکی از مهم‌ترین کاربردهای QSAR استفاده از این روش در پیش بینی فعالیت داروها و کمک به طراحی و سنتز ترکیبات دارویی جدید با احتمال تاثیر بالاتر برای درمان بسیاری از بیماری‌های صعب العلاج است. توسعه داروهای ضدسرطان از چهار دهه پیش، با کشف فعالیت ضدسرطانی داروها و استفاده موفقیت آمیزشان در درمان سلول‌های سرطانی گوناگون شروع شده است. از آن پس ترکیبات بی‌شماری سنتز شده و به عنوان نامزدهای بالقوه برای داروهای ضد سرطان استفاده شده‌اند، اما تنها تعداد انگشت شماری از آنها به عنوان داروهای بالینی موثر عمل می‌کنند (۵).

در حال حاضر، هزاران توصیفگر مولکولی برای مطالعات QSAR/QSPR وجود دارد. وارد کردن همه توصیفگرها در مدل‌های QSAR/QSPR می‌تواند منجر به بیش برازش (Overfitting)، پیچیدگی مدل و کاهش تفسیرپذیری آن شود و بنابراین عملکرد مدل و قابلیت پیشگویی آن را کاهش می‌یابد. روش های انتخاب متغیر می‌توانند با حذف تعدادی از توصیفگرها و انتخاب توصیفگرهای مرتبط با ویژگی یا فعالیت بر تعدادی از این ضعفها غلبه کنند. تاکنون چندین روش انتخاب متغیر، از قبیل الگوریتم ژنتیک، رگرسیون مرحله‌ای و شبیه سازی تبرید تدریجی (Simulated annealing) به طور گسترده استفاده شده اند. روش‌های بهینه سازی هوش ازدحامی (Swarm intelligence optimizations)، از قبیل بهینه سازی ازدحامی جزیی (Partial swarm optimization)، از جمله روشهای انتخاب متغیر هستند که معمولاً بر اساس رفتار زیستی حیوان و حشره به منظور پیدا کردن کوتاه ترین مسیر بین یک منبع غذا و لانه هایشان شبیه سازی شده‌اند، اخیراً در روشهای QSAR/QSPR به کار می‌روند (۶، ۷). یکی از بهترین روش‌های جستجو برای پیدا کردن با اهمیت‌ترین توصیفگرها روش الگوریتم ژنتیک است که بر اساس تکامل سیستم‌های بیولوژیکی است. روش الگوریتم ژنتیک برای پیدا کردن مینیمم های جهانی (Global minima) برای مسأله‌های چند بعدی مانند انتخاب توصیفگر در QSAR/QSPR زمانی که رویه ی پاسخ چندین بهینه موضعی (Local optima) دارد بسیار سودمند است (۸). لوکاسیوس (Lucasius) و همکارانش ثابت کردند که روش الگوریتم ژنتیک به طور کلی عملکرد بهتری نسبت به رگرسیون مرحله ای و تبرید تدریجی دارد (۹).

از جدیدترین تحقیقات در زمینه مدلسازی فعالیت ضد سرطانی، مطالعات مدلسازی QSAR ارائه شده توسط شایانفر و همکارانش در سال ۲۰۱۳، برای مهارکننده فارنسیل ترانسفراز (Farnesyltranseferase) به عنوان دسته‌ای از داروهای ضد سرطان به کمک روشهای مختلف اشاره کرد (۱۰). در سال ۲۰۱۱، بوهاری و همکارانش مدل‌های QSAR، مربوط به ۲۶۶ ترکیب برای ۲۹ سلول سرطانی مختلف با تعداد توصیف گرهای متفاوت بسط دادند (۱۱). در سال ۲۰۱۰، برتوسا و همکارانش تجزیه و تحلیل آماری مدل‌های QSAR، مربوط به فعالیت ضد تومور ۵۹ آمید و کینولین از سری تیوفن را مطالعه کردند (۱۲).

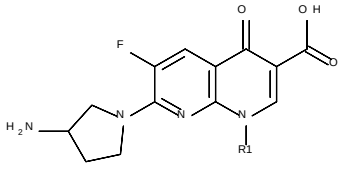
هدف از انجام این پروژه پیش بینی فعالیت ضد سرطانی مشتقات ۱و ۴-دی هیدرو-۴-اکسو-۱-(۲-تیازول)-۱و۸-نفتیریدین با استفاده از الگوریتم ژنتیک-رگرسیون خطی چندگانه و رگرسیون خطی چندگانه مرحله‌ای و در نهایت مقایسه عملکرد این دو روش در انتخاب توصیفگرهای مولکولی بود.

مواد و روشها

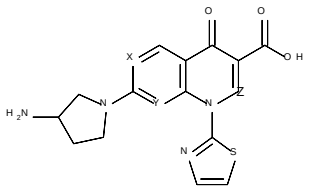
دسته داده‌ها

در این مطالعه یک سری صدتایی از مشتقات ۳ و ۷ استخلاف شده از ۱و ۴-دی هیدرو-۴-اکسو-۱-(۲-تیازول)-۱و۸-نفتیریدین ها مورد تحقیق قرار گرفتند. این ترکیبات قبلاً توسط تومیتا و همکارانش سنتز و آزمایش شدند (۱۵-۱۳). جدول ۱ ساختارهای شیمیایی ترکیبات مورد مطالعه را نشان می‌دهد. از IC_{50} ترکیبات در مقابل رده سلولی P380 که مربوط به لوسمی موش است در مطالعه استفاده شد. در نهایت، از مقادیر $\log(1/IC_{50})$ ترکیبات در جدول ۱، به عنوان متغیر وابسته در این تحقیق استفاده شد.

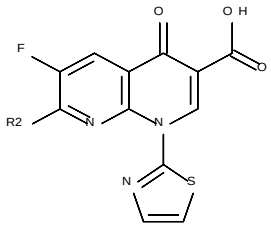
جدول ۱. ساختارهای شیمیایی ترکیبات ۱و ۴-دی هیدرو-۴-اکسو-۱-(۲-تیازول)-۱و۸-نفتیریدین و مشتقات آن



شماره	R1	شماره	R1
۱		۱۱	
۲		۱۲	
۳		۱۳	
۴		۱۴	
۵		۱۵	
۶		۱۶	
۷		۱۷	
۸		۱۸	
۹		۱۹	
۱۰			



شماره	X	Y	Z		
۲۰	N	N	CH		
۲۱	CF	N	N		
۲۲	N	CH	CH		
۲۳	CF	CH	CH		
۲۴	CF	CF	CH		

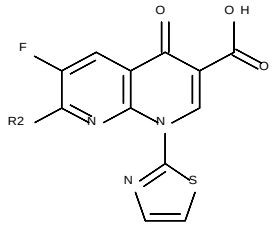


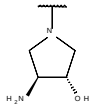
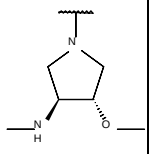
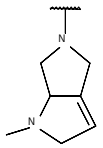
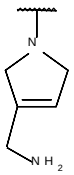
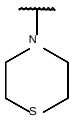
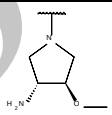
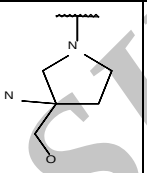
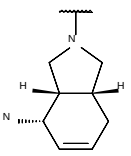
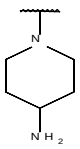
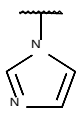
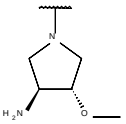
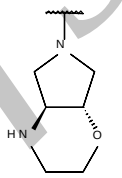
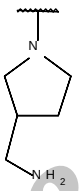
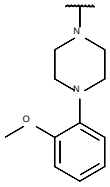
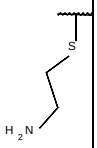
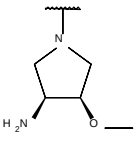
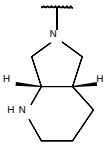
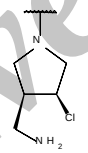
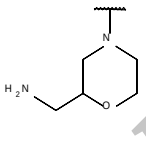
شماره	R2	شماره	R2
۲۵		۳۰	
۲۶		۳۱	
۲۷		۳۲	
۲۸		۳۳	
۲۹		۳۴	

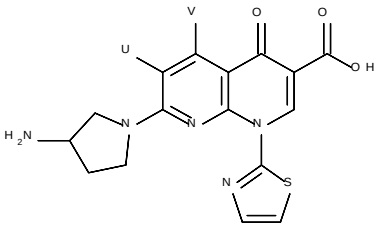
بهینه سازی مولکولی و محاسبه توصیفگرها

تمامی محاسبات توسط لپ تاپ ۷ هسته‌ای با ویندوز ۷ به عنوان سیستم عامل انجام شد. ابتدا با استفاده از نرم افزارکم درا (ChemDraw Ultra Version 8.0) ساختار مولکول‌ها رسم شد و ساختار هندسی مولکول‌ها از طریق روش مکانیک مولکولی میدان نیروی آلینجر MM2 و روش کوانتوم نیمه تجربی روش آوستین (AM1) با استفاده از الگوریتم پلاک-ریبایر (Polak-Ribier) تا گرادیان ریشه مربع میانگین ۰/۰۱ با استفاده از نرم افزار موپک (MOPAC Version 6.00) بهینه سازی شدند. مولکول‌های بهینه شده به نرم افزار دراگون (Dragon Version 5.4) انتقال داده شد و توصیفگرهای

ادامه جدول ۱. ساختارهای شیمیایی ترکیبات ۱و ۴-دی هیدرو-۴-اکسو-۱-(۲-تیازول)-۱و۸-نفتیریدین و مشتقات آن



شماره	R2	شماره	R2	شماره	R2	شماره	R2	شماره	R2
۳۵		۳۹		۴۳		۴۷		۵۱	
۳۶		۴۰		۴۴		۴۸		۵۲	
۳۷		۴۱		۴۵		۴۹		۵۳	
۳۸		۴۲		۴۶		۵۰		۵۴	



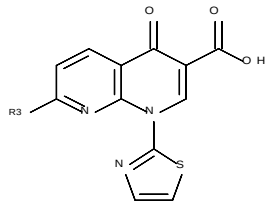
شماره	U	V		شماره	U	V
۵۵	H	H		۵۹	Cl	H
۵۶	NO_2	H		۶۰	H	Cl
۵۷	NH_2	H		۶۱	H	CF_3
۵۸	OH	H		۶۲	H	NH_2

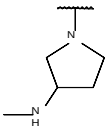
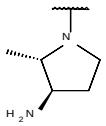
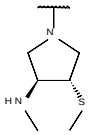
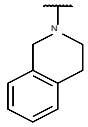
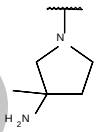
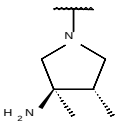
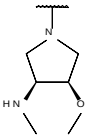
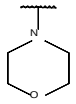
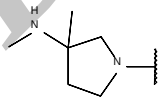
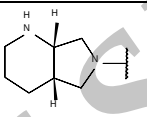
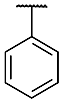
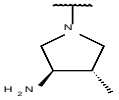
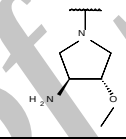
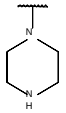
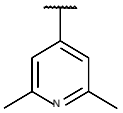
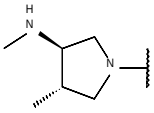
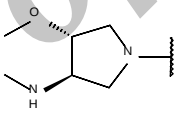
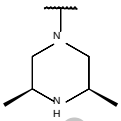
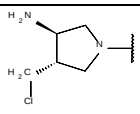
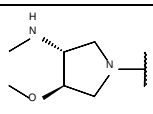
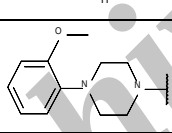
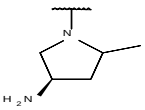
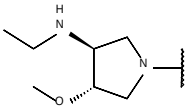
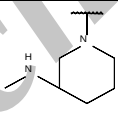
مولکولی محاسبه شدند. تمامی توصیفگرهای مولکولی به دست آمده از اطلاعات ساختارهای مولکولی به دست آمدند؛ بنابراین نیازی به داده های آزمایشگاهی برای محاسبه نبود. حدود ۱۲۱۲ توصیفگر مولکولی از قبیل GETAWAY، RDF، WHIM، توصیفگرهای شمارشی گروه عاملی و توصیفگرهای 3D-MoRSE برای مشخص کردن ساختارهای مولکولی مشتق شدند. این توصیفگرهای مولکولی به حدود ۴۶۸ توصیفگر با حذف متغیرهایی که همبستگی کمی با فعالیت دارند، متغیرهای ثابت و توصیفگرهای همبسته کاهش پیدا کردند.

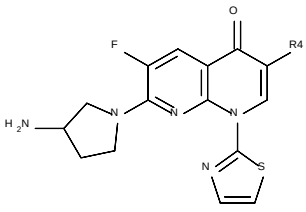
شکافت داده‌ها (Data splitting)

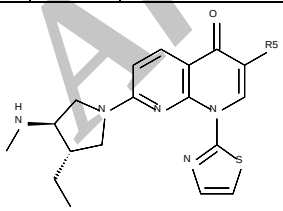
در قدم بعدی سری داده‌ها به دو دسته آموزشی و آزمون به صورت تصادفی توسط نرم افزار اکسل (Microsoft Excel

ادامه جدول ۱. ساختارهای شیمیایی ترکیبات ۱و ۴-دی هیدرو-۴-اکسو-۱-(۲-تیازول)-۱و۸-نفتیریدین و مشتقات آن



شماره	R3	شماره	R3	شماره	R3	شماره	R3
۶۳		۷۰		۷۷		۸۴	
۶۴		۷۱		۷۸		۸۵	
۶۵		۷۲		۷۹		۸۶	
۶۶		۷۳		۸۰		۸۷	
۶۷		۷۴		۸۱		۸۸	
۶۸		۷۵		۸۲		۸۹	
۶۹		۷۶		۸۳		۹۰	





شماره	R_4	شماره	R_4	شماره	R_4	شماره	R_5	شماره	R_5
۹۱	COOEt	۹۳	CH_2Ph	۹۵	$CONH_2$	۹۷	CHO	۹۹	$COCH_3$
۹۲	H	۹۴	CHO	۹۶	OH	۹۸	CH_2OH	۱۰۰	$CH_2CH=CHCOOCH_3$

2013) تقسیم می‌شوند. سری آموزشی که حاوی بیش‌ترین تعداد مولکول‌ها است باید نماینده مناسبی از کل مولکول‌ها باشد، چون این دسته از مولکول‌ها برای مدل‌سازی استفاده می‌شوند. مولکول‌های سری آزمون طوری انتخاب می‌شوند که نماینده مولکول‌های موجود در سری آموزش باشند و معمولاً ۲۰٪ از کل داده‌ها را به‌عنوان سری آزمون انتخاب می‌کنند. سری آزمون در هیچ یک از مراحل مدل‌سازی شرکت نمی‌کند و در واقع همان‌طور که از نامش پیداست به منظور ارزیابی قدرت پیش‌بینی و اعتبار مدل‌های به دست آمده مورد استفاده قرار می‌گیرد. شایان ذکر است که هر قدر مجموعه داده‌ها بزرگ‌تر و متنوع‌تر باشد مدل حاصل از آن معتبر بوده و قدرت پیش‌بینی آن بالا خواهد بود. در روش انتخاب تصادفی،

جدول۲. نتایج اعتبار سنجی مدل ترکیبات تصادفی انتخاب شده برای مجموعه آزمون

آزمایه	دسته آزمون تصادفی	S-MLR			GA-MLR		
		Q^2_{CV}	Q^2_{Cal}	Q^2_{Test}	Q^2_{CV}	Q^2_{Cal}	Q^2_{Test}
۱	۳,۱۲,۱۳, ۲۰, ۲۳, ۲۶, ۳۲, ۳۳, ۳۹, ۴۶, ۴۹, ۵۶, ۶۳, ۶۹, ۷۰, ۷۶, ۷۹, ۸۳, ۸۶, ۸۸	۰/۷۰۲۵	۰/۷۶۱۹	۰/۳۲۱۲	۰/۷۶۸۹	۰/۸۰۸۴	۰/۲۳۴۷
۲	۷, ۸, ۱۵, ۲۶, ۳۱, ۳۸, ۳۹, ۴۱, ۴۴, ۴۵, ۵۱, ۵۴, ۵۹, ۶۱, ۶۳, ۶۴, ۶۵, ۶۷, ۷۸, ۸۹	۰/۷۱۶۵	۰/۷۷۱۶	-۰/۳۰۲۹	۰/۷۰۷۹	۰/۷۶۱۲	۰/۳۹۶۵
۳	۵, ۶, ۱۲, ۱۶, ۲۰, ۲۳, ۳۲, ۵۰, ۵۶, ۶۱, ۶۴, ۷۲, ۷۳, ۸۲, ۸۵, ۸۷, ۹۰, ۹۱, ۹۴, ۹۵	۰/۶۶۷۷	۰/۷۱۷۸	۰/۶۶۴۲	۰/۶۸۳۱	۰/۷۳۴۲	۰/۶۲۷۸
۴	۲, ۱۹, ۲۷, ۳۲, ۳۳, ۴۱, ۴۷, ۴۸, ۵۳, ۵۵, ۶۵, ۶۶, ۶۷, ۷۰, ۷۸, ۸۷, ۸۹, ۹۰, ۹۲, ۹۷	۰/۶۶۰۳	۰/۷۲۶۰	۰/۵۹۷۱	۰/۶۹۲۰	۰/۷۴۸۷	۰/۵۹۴۲
۵	۲, ۷, ۱۵, ۳۰, ۳۳, ۳۸, ۴۴, ۴۶, ۵۳, ۵۶, ۶۳, ۶۶, ۶۹, ۷۵, ۷۶, ۸۰, ۸۲, ۸۴, ۸۹,۹۲	۰/۶۲۴۱	۰/۶۹۵۲	۰/۲۶۲۵	۰/۶۸۱۷	۰/۷۳۲۵	۰/۵۵۶۵
۶	۳, ۴, ۱۳, ۱۴, ۲۰, ۲۱, ۲۶, ۲۷, ۳۷, ۴۴, ۴۵, ۴۹, ۵۲, ۵۳, ۵۴, ۶۶, ۷۶, ۸۸, ۸۹, ۹۹	۰/۶۴۷	۰/۷۱۱۶	۰/۲۸۲۲	۰/۷۱۳۷	۰/۷۷۲۸	-۰/۰۲۵۴
۷	۱۶, ۲۱, ۲۲, ۳۳, ۴۱, ۴۲, ۴۴, ۴۶, ۵۶, ۵۷, ۵۹, ۶۵, ۶۶, ۷۲, ۷۷, ۸۹, ۹۱, ۹۵, ۹۷, ۱۰۰	۰/۶۷۵۱	۰/۷۳۱۸	-۰/۰۰۴۵	۰/۷۳۰۳	۰/۷۸۳۱	۰/۱۹۰۹
۸	۲, ۶, ۱۴, ۱۶, ۱۷, ۲۷, ۳۱, ۳۲, ۳۵, ۳۶, ۴۲, ۴۸, ۵۲, ۵۵, ۵۹, ۷۵, ۷۹, ۸۲, ۸۳, ۹۵	۰/۷۱۱۳	۰/۷۶۲۸	-۰/۷۶۱۰	۰/۷۰۴۲	۰/۷۷۱۶	-۱/۹۳۴۱
۹	۲, ۱۱, ۲۱, ۲۵, ۲۶, ۳۸, ۴۱, ۴۴, ۵۱, ۵۳, ۵۷, ۶۲, ۶۷, ۸۱, ۸۴, ۸۷, ۸۸, ۹۲, ۹۴, ۹۵	۰/۶۹۲۵	۰/۷۴۵۹	۰/۳۵۸۹	۰/۷۱۳۸	۰/۷۶۶۳	۰/۳۵۸۸
۱۰	۱, ۲, ۶, ۹, ۱۰, ۱۹, ۲۳, ۲۵, ۳۰, ۳۹, ۴۱, ۴۵, ۵۲, ۵۴, ۵۶, ۶۴, ۶۶, ۷۹, ۹۲, ۹۸	۰/۶۶۴۳	۰/۷۳۷۸	۰/۴۷۱۸	۰/۷۲۴۸	۰/۷۸۰۲	۰/۴۲۷۳
۱۱	۲, ۱۲, ۱۴, ۱۶, ۱۷, ۱۹, ۲۰, ۲۹, ۳۴, ۳۸, ۳۹, ۴۰, ۵۸, ۶۰, ۶۷, ۷۶, ۸۵, ۹۴, ۹۶, ۹۹	۰/۷۰۷۹	۰/۷۶۱۰	-۰/۴۷۶۲	۰/۷۱۰۵	۰/۷۶۲۳	-۰/۱۹۷۱
۱۲	۹, ۱۱, ۱۷, ۲۱, ۲۴, ۲۶, ۳۰, ۳۸, ۳۹, ۴۰, ۵۰, ۵۲, ۵۳, ۵۹, ۶۱, ۶۷, ۶۸, ۸۱, ۹۷, ۹۸	۰/۵۸۰۲	۰/۶۶۵۰	۰/۵۲۶۰	۰/۶۵۱۴	۰/۷۰۹۲	۰/۴۷۹۰
۱۳	۳, ۷, ۱۳, ۱۶, ۲۸, ۳۰, ۳۴, ۳۹, ۴۴, ۴۵, ۵۰, ۵۳, ۶۲, ۶۷, ۷۸, ۸۰, ۸۵, ۸۸, ۹۱, ۹۲	۰/۶۱۷۹	۰/۶۹۰۷	۰/۳۲۸۳	۰/۶۸۹۵	۰/۷۴۸۹	۰/۵۱۳۳
۱۴	۶, ۷, ۸, ۱۱, ۱۲, ۱۵, ۱۶, ۲۲, ۲۳, ۴۱, ۴۸, ۴۹, ۶۰, ۶۳, ۶۷, ۷۰, ۷۶, ۸۵, ۹۳, ۹۹	۰/۶۶۶۱	۰/۷۲۹۷	۰/۴۶۱۶	۰/۷۱۴۴	۰/۷۶۴۱	۰/۳۲۹۶
۱۵	۲, ۴, ۲۰, ۲۷, ۳۳, ۳۴, ۳۵, ۳۷, ۴۰, ۴۷, ۵۲, ۵۳, ۶۶, ۶۹, ۷۱, ۷۷, ۸۳, ۸۷, ۸۹, ۹۷	۰/۶۱۰۸	۰/۶۸۸۹	۰/۴۱۱۰	۰/۶۶۵۰	۰/۷۱۸۳	۰/۶۷۳۴
۱۶	۳, ۱۰, ۱۸, ۲۰, ۲۱, ۳۱, ۴۰, ۴۲, ۴۹, ۵۵, ۶۷, ۷۲, ۷۳, ۷۶, ۷۹, ۸۰, ۸۵, ۹۰, ۹۷, ۱۰۰	۰/۶۷۵۶	۰/۷۴۳۱	۰/۵۳۲۹	۰/۶۹۵۶	۰/۷۵۵۶	۰/۳۰۴۸
۱۷	۳, ۱۰, ۳۱, ۳۴, ۳۵, ۳۷, ۴۴, ۴۵, ۴۹, ۵۵, ۶۱, ۶۲, ۶۴, ۷۲, ۷۶, ۸۴, ۸۶, ۹۰, ۹۳, ۹۶	۰/۶۹۳۶	۰/۷۶۲۱	-۰/۳۷۶۹	۰/۷۴۹۶	۰/۷۹۴۰	۰/۴۶۵۵
۱۸	۱, ۱۰, ۲۰, ۲۳, ۲۵, ۴۱, ۵۶, ۵۸, ۶۰, ۶۶, ۶۷, ۶۸, ۷۰, ۷۱, ۸۴, ۸۷, ۸۹, ۹۶, ۹۷, ۹۹	۰/۷۴۴۴	۰/۷۹۱۲	۰/۳۱۳۷	۰/۷۵۳۹	۰/۸۰۱۳	۰/۴۱۰۹
۱۹	۲, ۶, ۱۸, ۱۹, ۳۸, ۴۳, ۴۴, ۵۰, ۵۱, ۵۵, ۶۲, ۷۰, ۷۱, ۷۲, ۷۴, ۷۸, ۸۶, ۸۹, ۹۳, ۹۶	۰/۶۶۳۳	۰/۷۳۰۹	-۰/۰۱۴۲	۰/۶۹۲۹	۰/۷۳۷۸	۰/۵۸۲۱
۲۰	۱, ۳, ۵, ۶, ۱۰, ۱۱, ۱۲, ۲۱, ۳۰, ۳۱, ۳۲, ۳۳, ۳۷, ۴۰, ۴۱, ۴۶, ۴۷, ۹۳, ۹۸, ۱۰۰	۰/۷۶۰۷	۰/۸۱۲۹	-۰/۰۴۴۰	۰/۷۶۴۸	۰/۸۱۷۶	-۰/۱۲۱۳

مولکولها را برحسب فعالیتشان بصورت صعودی و یا نزولی مرتب می کنیم و سپس از بین آنها نسبت های ذکر شده در بالا را طوری انتخاب می کنیم که از لحاظ آماری دارای توزیع نرمال باشد. در این مطالعه ۲۰ بار انتخاب تصادفی انجام شد و دسته آموزش و آزمون به ترتیب شامل ۸۰ و ۲۰ مولکول است. جدول ۲ تمامی ترکیبات انتخاب شده برای دسته آزمون و نتایج مدلسازی آنها را نشان می‌دهد.

انتخاب متغیر توسط الگوریتم ژنتیک

روش ترکیبی الگوریتم ژنتیک- رگرسیون خطی چندگانه (GA-MLR) یکی از روش‌هایی است که در این مطالعه به عنوان روش انتخاب متغیر استفاده شده است. الگوریتم ژنتیک یک روش جستجوی قدرتمند براساس تکامل سیستم های بیولوژیکی است. در مطالعات QSAR/QSPR، دستیابی به مدلی با تعداد کمی از توصیفگرهای ساختاری دارای اهمیت است، زیرا این مسأله منجر به ایجاد مدلی ساده و قابل پیشگویی خواهد شد. در حقیقت روش الگوریتم ژنتیک یکی از رایج‌ترین روشهای انتخاب متغیر است که به میزان گسترده‌ای برای مشکلات بهینه سازی پیچیده در زمینه‌های گوناگون از جمله روش‌های QSAR/QSPR استفاده می‌شود (۲۴-۱۶). در اینجا از جعبه ابزار الگوریتم ژنتیک موجود در نرم افزار MATLAB 2013a (gatool) برای انتخاب توصیفگرهای مناسب استفاده شده است. همان طور که در بخش گذشته اشاره کردیم متغیرها از ۱۲۱۲ متغیر به حدود ۴۶۸ متغیر با حذف متغیرهایی که همبستگی کمی با فعالیت دارند، متغیرهای ثابت و توصیفگرهای همبسته کاهش پیدا کردند.

روش الگوریتم ژنتیک به کار گرفته در این مطالعه از یک نمایش دوتایی به عنوان روش رمزگذاری استفاده می کند؛ حضور و یا عدم حضور یک توصیفگر در یک کروموزوم به وسیله ۱ یا ۰ رمزگذاری می‌شود. در الگوریتم ژنتیک بهینه سازی از طریق تغییر و انتخاب و به واسطه ارزیابی تابع شایستگی η انجام می‌شود. تابع شایستگی مورد استفاده در این مطالعه متوسط ریشه میانگین مربع خطای واسنجی و

جدول۳. نتایج مدل S-MLR برای مشتقات ۱و ۴-دی هیدرو-۴-اکسو-۱-(۲-تیازول)-۱و۸-نفتیریدین برای بهترین دسته تصادفی

متغیر	نوع توصیفگر	تعریف	[1]b	[2]S_b	[3]b_s	[4]VIF
عرض از مبدا	—	—	13/046	2/567	—	—
RDF020p	توصیفگرهای RDF	تابع توزیع شعاعی -۲۰ / توزین شده توسط الکترونگاتیویته ساندرسون	0/514	0/096	0/408	1/315
C-002	قطعات اتم محور	CH2R2	0/564	0/161	0/268	1/334
Mor16m	توصیفگرهای 3D-MoRSE	سیگنال ۱۶/ توزین شده توسط جرم	-1/418	0/416	-0/246	1/185
R4m	توصیفگرهای GETAWAY	همبستگی R لایه ۴/ توزین شده توسط جرم	-4/582	0/983	-0/326	1/116
Mor21m	توصیفگرهای 3D-MoRSE	سیگنال ۲۱/ توزین شده توسط جرم	-1/797	0/495	-0/274	1/300
G2e	توصیفگرهای WHIM	دومین جز تقارن جهتی ضریب WHIM / توزین شده توسط الکترونگاتیویته ی ساندرسون	65/981	14/679	0/355	1/422
nRNR2	شمارش های گروه عاملی	تعداد آمین های نوع سوم (آلیفاتیک)	۰/۶۸۰	0/235	۰/۲۳۰	1/437
Mor22m	توصیفگرهای 3D-MoRSE	سیگنال ۲۲/ توزین شده توسط جرم	1/040	0/406	0/182	1/149

[1] b ضریب رگرسیونی غیراستاندارد توصیفگرها؛ [2] S_b خطای استاندارد رگرسیونی؛ [3] b_s ضریب رگرسیونی استاندارد شده؛ [4] VIF فاکتور تورم واریانس برای سنجش همبستگی خطی چندگانه بین توصیف گرها در مدل مورد نظر

جدول ۴. نتایج مدل GA-MLR برای مشتقات ۱و۴-دی هیدرو-۴-اکسو-۱-(۲-تیازولیل)-۱و۸-نفتیریدین برای بهترین دسته تصادفی

متغیر	نوع توصیفگر	تعریف	[1]b	[2]S_b	[3]b_s	[4]VIF
عرض از مبدا	—	—	25/437	8/325	—	—
RDF055e	توصیفگرهای RDF	تابع توزیع شعاعی -۵۵ / توزین شده توسط الکترونگاتیویته ی ساندرسون	-0/111	0/016	-0/555	1/562
nRNR2	شمارش های گروه عاملی	تعداد آمین های نوع سوم (آلیفاتیک)	0/556	0/۲۰۰	0/187	1/144
Mor03p	توصیفگرهای 3D-MoRSE	سیگنال ۳/ توزین شده توسط قطبش پذیری	-0/777	0/144	-0/4۰۰	1/389
nCp	شمارش های گروه عاملی	تعداد C(sp3) نوع اول انتهایی	0/439	0/094	0/383	1/713
nRSR	شمارش های گروه عاملی	تعداد سولفیدها	0/84۰	0/243	0/240	1/223
Mor07e	توصیفگرهای 3D-MoRSE	سیگنال ۷/ توزین شده توسط الکترونگاتیویتی ساندرسون	0/289	0/096	0/212	1/247
ISH	توصیفگرهایGETAWAY	مقدار اطلاعات استاندارد شده روی اهرم برابری	21/324	8/322	0/164	1/035
Mor16p	توصیفگرهای 3D-MoRSE	سیگنال ۱۶/ توزین شده توسط قطبش پذیری	-1/654	0/479	-0/251	1/337

[1] b ضریب رگرسیونی غیراستاندارد توصیفگرها؛ [2] S_b خطای استاندارد رگرسیونی؛ [3] b_s ضریب رگرسیونی استاندارد شده؛ [4] VIF فاکتور تورم واریانس برای سنجش همبستگی خطی چندگانه بین توصیف گرها در مدل مورد نظر

اعتبارسنجی متقاطع،۲/(RMSEC+RMSECV)، است. اندازه جمعیت در این بررسی ۷۰۰۰ بود و سایر گزینه ها از جمله هم گذری، جهش و مهاجرت، مطابق با پیش فرض بخش مربوط به الگوریتم ژنتیک در نرم افزار MATLAB 2012 است. پس از ارزیابی مدل‌های متفاوت مشاهده می‌شود که بهترین مدل خطی چندمتغیره‌ای دارای هشت توصیفگر است.

انتخاب متغیر توسط رگرسیون خطی چندگانه گام به گام

رگرسیون خطی چندگانه (MLR) روش گسترش یافته رگرسیون کلاسیک است که بیش از یک بعد را در محاسبات در نظر می‌گیرد. روش MLR به واسطه انجام محاسبات رگرسیونی چند متغیره استاندارد و از طریق استفاده از چندین پارامتر در یک معادله واحد، معادله QSAR را محاسبه می‌کند (۲۵، ۲۶). به منظور انجام محاسبات در روش S-MLR، نرم افزار IBM SPSS مورد استفاده قرار گرفت. مجموعه داده‌ها به مجموعه‌های کاهش یافته ترکیبات ضد سرطان به مجموعه آموزش و آزمون، در مرحله نخست، باید مجموعه آزمون از کل داده ها حذف گردیده و سپس محاسبات با استفاده از روش رگرسیونی خطی گام به گام انجام شوند. پس از بررسی، مدل‌های دارای چهار، پنج، شش و ... بهترین مدل خطی چندمتغیره دارای ۸ توصیفگر بود.

در نهایت باید اعتبار مدل از لحاظ آماری مورد ارزیابی قرار گیرد و مدل مناسب انتخاب شود. پارامتر آماری Q2 برای ارزیابی مدل حاصل از اعتبار سنجی همگذری به کار می‌رود و

جدول ۵. ماتریس ضریب همبستگی برای توصیف کننده ها و ثابت ($Log(1/IC_{50})$ ترکیبات ۱و ۴-دی هیدرو-۴-اکسو-۱-(۲-تیازول)-۱و۸-نفتیریدین و مشتقات آن با اثر ضد سرطان در مدل S-MLR

	$Log(1/IC_{50})$	RDF020p	C-002	Mor16m	R4m	Mor21m	G2e	nNR2	Mor22m
$Log(1/IC_{50})$	۱								
RDF020p	۰/۴۹۲	۱							
C-002	۰/۳۰۷	-۰/۱۲۸	۱						
Mor16m	-۰/۴۲۴	-۰/۲۴۷	-۰/۰۶۵	۱					
R4m	-۰/۲۸۵	۰/۰۵۲	-۰/۲۲۶	۰/۰۰۵	۱				
Mor21m	-۰/۳۰۶	-۰/۱۶۳	۰/۰۵۹	۰/۳۱۱	-۰/۱۱۸	۱			
G2e	-۰/۰۶۸	-۰/۱۸۱	-۰/۲۲۴	۰/۱۴۴	۰/۱۹۶	۰/۳۱۶	۱		
nNR2	۰/۳۶۳	۰/۱۴۷	۰/۳۸۲	-۰/۰۸۶	-۰/۰۸۷	-۰/۲۱۳	-۰/۴۳۲	۱	
Mor22m	۰/۳۶۱	۰/۳۳۲	۰/۰۵۵	-۰/۰۲۴	۰/۰۰۰	-۰/۰۳۶	-۰/۰۲۳	۰/۰۹۴	۱

جدول ۶. ماتریس ضریب همبستگی برای توصیف کننده ها و ثابت (Log(1/IC50 ترکیبات ۱و ۴-دی هیدرو-۴-اکسو-۱-(۲-تیازول)-۱و۸-نفتیریدین و مشتقات آن با اثر ضد سرطان در مدل GA-MLR

	$Log(1/IC_{50})$	RDF055e	nNR2	Mor03p	nCp	nRSR	Mor07e	ISH	Mor16p
Log(1/IC50	۱								
RDF055e	-۰/۰۷۸	۱							
nNR2	۰/۳۶۳	۰/۲۱۴	۱						
Mor03p	-۰/۴۱۹	-۰/۳۲۳	-۰/۲۲۳	۱					
nCp	۰/۳۶۶	۰/۴۸۵	۰/۲۵۱	-۰/۰۳۴	۱				
nRSR	۰/۴۳۳	۰/۰۷۸	۰/۰۹۸	-۰/۳۳۴	۰/۱۰۸	۱			
Mor07e	۰/۲۸۱	۰/۳۴۶	۰/۱۷۲	-۰/۲۱۴	۰/۳۱۹	-۰/۰۶۲	۱		
ISH	۰/۱۲۷	-۰/۰۹۲	-۰/۰۴۷	۰/۰۲۵	-۰/۰۶۹	۰/۰۱۷	-۰/۰۳۴	۱	
Mor16p	-۰/۴۵۸	-۰/۱۷۷	-۰/۲۲۸	۰/۰۹	-۰/۴۳۵	-۰/۲۱۱	-۰/۱۶۷	۰/۱۵۸	۱

مقدار Q^2 آزمون خوبی برای داده‌هایی است که به طور یکنواخت توزیع شده‌اند. در این بخش با نگاهی گذرا بر پارامترهای آماری خواهیم پرداخت. اعتبار سنجی همگذری تک به تک (LOO-CV: Leave One Out –Cross Validation) یکی از اعتبار سنجی‌های داخلی مدل QSAR است.توانایی پیشگویی مدل QSAR با استفاده از روش LOO-CV تعیین می‌شود. واریانس اعتبار سنجی همگذری Q^2_{CV} توسط معادله زیر محاسبه می‌شود.

(۱)

$$Q^2_{CV} = 1 - \frac{\sum_{i-1}^{Cal}(y_i - \hat{y}_i)^2}{\sum_{i-1}^{Cal}(y_i - \bar{y})^2}$$

y_i، $\hat{y}_i$ و $\bar{y}$ به ترتیب مقادیراندازه گیری، پیشگویی و میانگین(کل داده ها ی کالیبراسیون) متغییر وابسته هستند؛ عمل جمع تمام ترکیبات در دسته واسنجی را در بر می‌گیرد. روش LOO-CV برای تعیین استحکام و قدرت پیشگویی کافی نیست. مدل QSAR توسط دسته واسنجی ترکیبات شیمیایی به تنهایی ایجاد شده سپس برای اعتبار سنجی برونی دسته ترکیبات شیمیایی و تحقیق اعتبار سنجی بیشتر قدرت پیشگویی مدل، به کار گرفته می‌شود. سرانجام مدل با استفاده از دسته اعتبارسنجی برونی(دسته آزمون) اعتبارسنجی می‌شود و نتایج اعتبار سنجی مدل در جدول۲ نشان داده شده است.

فرمول برای محاسبه Q^2_{ext} به صورت زیر است:

(۲)

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{test}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{test}(y_i - \bar{y}_{test})^2}$$

که $\hat{y}i$ و yi به ترتیب مقادیر پیشگویی و اندازه گیری فعالیت ترکیب هستند، و $\bar{y}$test مقدار میانگین فعالیت برای دسته آزمون است؛ جمع تمام ترکیبات دسته ی آزمون را پوشش می‌دهد.

جدول ۷. مقادیر توصیفگرهای مولکولی، مقادیر $\log(1/IC_{50})$ آزمایشی و پیشگویی شده توسط مدل GA-MLR

شماره	RDF055e	nNR2	Mor03p	nCp	nRSR	Mor07e	ISH	Mor16p	$Log(1/IC_{50})_{EXP}$	$\log(1/IC_{50})_{Prd}$	دسته
۱	۱۸/۴۸۰	۲	-۳/۴۱۸	۰	۰	۳/۵۱۸	۱	-۰/۰۱۸	-۰/۸۱۳۹	-۱/۳۵۰۰	واسنجی
۲	۱۴/۵۰۴	۲	-۳/۲۶۵	۰	۰	۳/۹۵۴	۱	-۰/۱۲۹	-۰/۷۳۵۲	-۰/۷۱۸۰	آزمون
۳	۱۴/۵۰۴	۲	-۳/۲۶۵	۰	۰	۳/۹۵۴	۱	-۰/۱۲۹	-۱/۴۳۲۶	-۰/۷۱۸۰	واسنجی
۴	۱۷/۶۸۶	۲	-۲/۸۳۰	۰	۰	۳/۵۵۷	۰/۹۸۲	-۰/۱۸۸	-۱/۳۷۶۱	-۱/۸۱۰۱	آزمون
۵	۱۱/۵۸۸	۲	-۳/۶۱۹	۰	۱	۲/۹۰۵	۱	-۰/۰۴۹	-۰/۵۰۵۹	۰/۲۸۵۳	واسنجی
۶	۱۳/۸۱۸	۲	-۳/۰۱۳	۰	۰	۳/۳۰۹	۰/۹۸۴	-۰/۲۳۳	-۱/۴۱۷۷	-۱/۱۹۳۲	واسنجی
۷	۱۴/۱۴۵	۲	-۳/۰۹۷	۱	۰	۲/۹۱۲	۱	-۰/۲۶۷	-۰/۱۹۱۳	-۰/۴۴۲۵	واسنجی
۸	۱۱/۰۹۷	۲	-۳/۴۸۱	۰	۰	۳/۱۳۴	۰/۹۸۴	-۰/۱۵۲	-۱/۱۶۴۸	-۰/۷۱۲۱	واسنجی
۹	۱۹/۴۹۱	۲	-۲/۵۲۴	۱	۰	۳/۵۰۱	۱	-۰/۰۲۳	-۱/۴۲۹۰	-۱/۷۱۴۵	واسنجی
۱۰	۱۱/۶۰۵	۲	-۳/۶۴۴	۰	۱	۳/۰۶۱	۱	-۰/۱۸۲	-۱/۴۰۹۶	۰/۵۶۷۹	واسنجی
۱۱	۱۱/۱۱۰	۲	-۳/۴۲۴	۱	۱	۲/۸۶۰	۱	-۰/۲۵۴	۰/۸۸۲۷	۰/۹۵۱۹	واسنجی
۱۲	۱۰/۸۲۵	۲	-۳/۳۵۴	۱	۱	۲/۵۸۸	۱	-۰/۲۰۵	۰/۱۷۵۴	۰/۷۶۹۵	واسنجی
۱۳	۱۶/۴۱۷	۲	-۳/۰۳۶	۲	۱	۳/۱۰۸	۱	-۰/۰۳۴	۰/۳۵۰۵	۰/۲۰۸۱	واسنجی
۱۴	۳۲/۱۶۲	۲	-۴/۰۴۵	۳	۱	۲/۱۰۹	۰/۹۸۶	-۰/۴۳۲	-۰/۳۰۹۵	-۰/۲۴۵۵	واسنجی
۱۵	۱۸/۰۲۳	۲	-۴/۱۱۷	۰	۱	۴/۲۳۷	۱	-۰/۱۹۴	۰/۹۳۸۵	۰/۵۸۲۷	واسنجی
۱۶	۱۱/۴۰۹	۲	-۳/۴۵۶	۰	۱	۲/۴۳۱	۱	-۰/۱۷۴	۰/۶۳۹۴	۰/۲۴۸۳	واسنجی
۱۷	۱۰/۴۷۶	۲	-۳/۴۶۵	۰	۱	۲/۵۶۶	۱	-۰/۱۶۵	۰/۷۲۷۹	۰/۳۸۳۰	واسنجی
۱۸	۱۹/۴۲۳	۲	-۳/۳۵۱	۱	۱	۳/۵۱۵	۱	-۰/۱۶۷	۰/۰۶۳۸	۰/۰۱۷۸	واسنجی
۱۹	۱۹/۷۷۰	۲	-۳/۹۷۳	۰	۱	۴/۷۲۹	۱	-۰/۳۲۵	۰/۹۳۸۵	۰/۶۳۵۸	واسنجی
۲۰	۱۰/۶۹۳	۲	-۳/۵۹۲	۰	۱	۳/۳۴۱	۱	-۰/۰۱۸	۰/۸۸۲۴	۰/۴۳۸۴	آزمون
۲۱	۱۹/۶۳۱	۱	-۳/۴۴۷	۰	۱	۳/۳۱۱	۱	-۰/۱۱۳	-۱/۴۲۴۴	-۱/۰۷۳۹	واسنجی
۲۲	۱۳/۸۲۲	۲	-۳/۵۹۰	۰	۱	۲/۹۸۳	۱	-۰/۱۳۸	-۰/۴۴۶۹	۰/۱۸۴۵	واسنجی
۲۳	۱۶/۵۶۴	۲	-۲/۵۴۹	۰	۱	۲/۵۳۸	۱	-۰/۲۶۲	-۱/۴۲۶۷	-۰/۸۵۲۲	واسنجی
۲۴	۲۰/۲۵۰	۲	-۲/۵۲۰	۰	۱	۲/۲۱۷	۰/۹۸۴	-۰/۳۰۸	-۲/۱۴۶۷	-۱/۶۴۱۷	واسنجی
۲۵	۱۴/۱۶۶	۱	-۲/۹۷۴	۰	۱	۱/۳۹۷	۱	-۰/۰۲۴	-۱/۴۵۶۸	-۱/۵۳۵۲	واسنجی
۲۶	۱۲/۹۷۱	۲	-۲/۹۰۲	۰	۱	۳/۱۳۸	۱	-۰/۱۹۳	۱/۰۱۹۱	-۰/۱۱۹۸	واسنجی
۲۷	۱۳/۹۵۷	۲	-۳/۹۰۹	۰	۱	۳/۰۴۳	۱	-۰/۲۵۴	۰/۷۶۲۷	۰/۶۲۶۶	آزمون
۲۸	۱۲/۶۳۱	۲	-۴/۰۰۱	۰	۱	۳/۰۲۸	۱	-۰/۳۷۲	۱/۲۷۴۹	۱/۰۳۶۱	واسنجی
۲۹	۱۴/۶۹۳	۲	-۴/۰۰۲	۰	۱	۲/۸۹۸	۱	-۰/۳۱۶	۱/۳۷۰۶	۰/۶۷۷۸	واسنجی
۳۰	۱۴/۵۲۷	۲	-۳/۹۲۵	۰	۱	۲/۹۸۶	۱	-۰/۲۹۵	۱/۱۲۷۳	۰/۶۲۷۱	واسنجی
۳۱	۱۵/۹۵۷	۳	-۳/۰۵۹	۲	۱	۳/۱۴۳	۱	-۰/۲۰۱	۰/۸۴۹۹	۱/۱۱۹۴	واسنجی
۳۲	۱۳/۶۲۳	۲	-۴/۲۰۰	۱	۱	۲/۹۷۱	۱	-۰/۱۸۶	۱/۱۷۵۲	۱/۱۹۵۵	واسنجی
۳۳	۱۵/۲۷۳	۲	-۴/۰۵۵	۱	۱	۲/۸۲۶	۱	-۰/۳۲۳	۱/۵۴۹۸	۱/۰۸۴۴	آزمون
۳۴	۱۲/۰۴۶	۲	-۳/۹۲۶	۱	۱	۲/۵۹۹	۱	-۰/۲۸۲	۰/۸۳۴۵	۱/۲۰۸۹	آزمون

به علاوه، آنالیز فاکتر تورم واریانس (VIF= Variance Inflation Factor) انجام می شود تا ببینیم آیا هم بستگی خطی چند گانه بین توصیف کننده‌ها در مدل وجود دارد. مقدار VIF از رابطه $1/1-r^2$ محاسبه می‌شود، که r^2 ضریب هم بستگی چند گانه یک توصیف گر بر روی بقیه توصیف کننده های مولکولی است. اگر مدل‌ها دارای توصیف کننده های مولکولی با مقدار VIF بزرگتر از ۱۰ باشند پذیرفته نخواهند شد و تضمین می‌کند که مربع ضریب هم بستگی چند گانه برای هر توصیف کننده در مدل از ۰/۹ تجاوز نمی‌کند (۲۷،۲۸).

یافته‌ها

مدلسازی S-MLR و GA-MLR

بعد از شکافت داده‌ها به دو دسته واسنجی و آزمون، به منظور انتخاب متغیرهای مهم برای توصیف فعالیت ترکیبات

ادامه جدول ۷. مقادیر توصیفگرهای مولکولی، مقادیر $\log(1/IC_{50})$ آزمایشی و پیشگویی شده توسط مدل GA-MLR

شماره	RDF055e	nNR2	Mor03p	nCp	nRSR	Mor07e	ISH	Mor16p	$Log(1/IC_{50})_{EXP}$	$\log(1/IC_{50})_{Prd}$	دسته
۳۵	۱۳/۷۵۹	۲	-۴/۱۰۵	۰	۱	۳/۲۱۹	۰/۹۸۴	-۰/۲۴۲	-۰/۷۲۷۵	۰/۴۹۰۷	آزمون
۳۶	۱۵/۵۴۰	۲	-۳/۴۴۴	۱	۱	۳/۴۳۰	۱	-۰/۴۷۰	۰/۷۷۵۵	۰/۹۹۷۷	واسنجی
۳۷	۱۸/۵۱۲	۲	-۳/۹۶۵	۱	۱	۳/۳۷۴	۱	-۰/۳۱۳	۱/۲۴۶۴	۰/۷۹۶۸	آزمون
۳۸	۱۵/۹۵۳	۲	-۳/۴۱۰	۱	۱	۳/۲۵۶	۱	-۰/۴۳۱	۰/۶۸۳۶	۰/۸۱۰۶	واسنجی
۳۹	۱۹/۶۵۵	۲	-۴/۰۷۴	۲	۱	۳/۶۰۶	۰/۹۸۶	-۰/۴۶۱	۱/۳۶۷۵	۱/۲۰۶۹	واسنجی
۴۰	۱۴/۰۰۴	۲	-۴/۳۴۶	۰	۱	۳/۰۰۹	۱	-۰/۱۳۸	۰/۶۸۳۶	۰/۷۵۹۳	آزمون
۴۱	۱۸/۹۳۳	۲	-۴/۶۵۰	۰	۱	۲/۳۷۵	۱	-۰/۲۹۹	-۰/۳۷۹۴	۰/۵۳۱۴	واسنجی
۴۲	۱۸/۵۳۲	۲	-۴/۲۶۶	۰	۱	۳/۴۲۳	۰/۹۸۶	-۰/۲۲۸	۰/۱۳۵۶	۰/۱۶۴۵	واسنجی
۴۳	۱۶/۷۱۷	۳	-۳/۵۳۳	۱	۱	۳/۶۲۴	۰/۹۸۶	-۰/۳۵۹	۱/۰۳۶۷	۱/۰۶۶۱	واسنجی
۴۴	۲۰/۰۲۵	۲	-۴/۵۹۲	۰	۱	۳/۸۰۳	۱	-۰/۱۸۳	۰/۴۵۴۸	۰/۵۸۶۰	واسنجی
۴۵	۱۱/۷۴۰	۲	-۳/۹۸۰	۰	۱	۲/۵۰۰	۰/۹۶۹	-۰/۲۵۶	۰/۵۸۶۰	۰/۱۱۳۲	واسنجی
۴۶	۱۳/۰۰۷	۲	-۴/۱۸۹	۰	۱	۱/۵۰۶	۱	-۰/۲۶۰	۰/۴۳۴۰	۰/۵۱۵۴	واسنجی
۴۷	۱۲/۸۸۱	۲	-۳/۷۵۲	۰	۱	۲/۰۰۲	۰/۹۸۴	-۰/۱۴۰	۰/۶۶۳۹	-۰/۲۰۶۵	آزمون
۴۸	۲۰/۳۷۳	۲	-۳/۸۹۲	۰	۱	۲/۰۱۹	۱	-۰/۰۵۶	-۰/۶۲۱۸	-۰/۷۲۲۲	واسنجی
۴۹	۲۸/۷۹۰	۲	-۴/۴۲۸	۱	۱	۳/۹۳۷	۰/۹۸۸	-۰/۱۱۶	-۰/۳۱۷۴	-۰/۴۰۳۴	واسنجی
۵۰	۱۶/۸۰۷	۲	-۳/۸۲۲	۰	۱	۱/۵۴۰	۰/۹۸۵	-۰/۱۵۵	-۰/۶۳۰۲	-۰/۶۷۵۳	واسنجی
۵۱	۱۸/۹۲۳	۲	-۴/۰۴۴	۰	۱	۲/۶۷۱	۱	-۰/۰۷۴	-۰/۶۴۶۸	-۰/۲۲۴۹	واسنجی
۵۲	۱۴/۷۳۲	۲	-۳/۱۵۰	۰	۱	۳/۲۸۳	۱	-۰/۲۹۷	-۱/۴۴۶۹	۰/۰۹۱۳	واسنجی
۵۳	۲۲/۶۸۱	۱	-۳/۷۴۱	۰	۱	۲/۴۸۶	۰/۹۸۲	-۰/۱۵۵	-۱/۴۳۶۱	-۱/۷۳۶۸	واسنجی
۵۴	۱۰/۶۲۱	۱	-۲/۸۳۷	۰	۰	۲/۳۶۸	۱	-۰/۰۰۸	-۱/۵۱۲۵	-۱/۸۳۴۰	آزمون
۵۵	۱۰/۷۶۸	۲	-۳/۷۳۳	۰	۱	۳/۳۳۱	۱	-۰/۱۴۷	۱/۵۵۲۸	۰/۷۵۰۱	آزمون
۵۶	۲۴/۱۶۵	۲	-۳/۳۶۵	۰	۱	۲/۶۴۹	۰/۹۸۵	-۰/۵۰۴	-۰/۷۵۵۲	-۰/۹۴۹۴	واسنجی
۵۷	۱۲/۲۱۸	۲	-۴/۰۰۲	۰	۱	۳/۱۶۴	۰/۹۸۴	-۰/۱۵۴	-۰/۴۳۳۳	۰/۴۲۰۳	واسنجی
۵۸	۱۳/۰۸۸	۲	-۳/۶۸۸	۰	۱	۳/۳۲۷	۱	-۰/۱۲۸	-۱/۰۶۲۳	۰/۴۲۵۰	واسنجی
۵۹	۱۴/۵۹۷	۲	-۳/۵۲۷	۰	۱	۲/۱۰۲	۱	-۰/۳۴۶	۰/۲۳۷۱	۰/۱۳۹۰	واسنجی
۶۰	۱۵/۲۶۹	۲	-۴/۲۶۱	۰	۱	۳/۳۴۶	۱	-۰/۲۶۹	۰/۹۰۲۷	۰/۸۶۶۹	واسنجی
۶۱	۱۶/۰۱۴	۲	-۴/۳۰۲	۰	۱	۳/۵۵۵	۱	-۰/۲۰۵	۰/۶۳۷۵	۰/۷۷۰۶	واسنجی
۶۲	۱۴/۷۶۱	۲	-۴/۰۳۶	۰	۱	۳/۲۴۹	۱	-۰/۱۸۴	۱/۶۶۱۵	۰/۵۷۹۸	واسنجی
۶۳	۱۳/۱۸۴	۲	-۳/۷۷۳	۱	۱	۳/۳۳۷	۱	-۰/۱۷۷	۱/۵۷۰۲	۱/۰۰۳۳	واسنجی
۶۴	۱۲/۲۹۶	۲	-۴/۲۳۷	۱	۱	۳/۳۸۶	۱	-۰/۰۲۸	۱/۲۹۰۷	۱/۲۳۰۲	واسنجی
۶۵	۱۶/۳۵۶	۲	-۴/۱۰۲	۲	۱	۲/۹۳۱	۱	-۰/۱۶۹	۱/۳۳۰۷	۱/۲۱۵۳	واسنجی
۶۶	۱۸/۰۰۳	۲	-۴/۰۰۰	۱	۱	۳/۲۳۴	۱	-۰/۱۴۳	۱/۵۸۸۴	۰/۵۵۸۸	آزمون
۶۷	۱۸/۰۵۰	۲	-۳/۸۴۳	۲	۱	۱/۱۹۵	۱	-۰/۳۶۷	۱/۵۰۷۲	۱/۳۲۹۵	واسنجی
۶۸	۱۵/۱۰۸	۲	-۴/۵۳۳	۰	۱	۲/۹۳۳	۱	-۰/۰۹۵	۱/۲۲۸۴	۰/۶۸۸۹	واسنجی

ضدسرطانی، روشهای S-MLR و GA-MLR بر روی دسته واسنجی اعمال شد. برای دستیابی به بهترین مدلهای QSAR، ابتدا بهترین مدل های دارای یک، دو، سه، چهار تا ده متغیره ساخته شدند و از بین این مدلها، مدل هشت متغیره به عنوان بهترین مدل از نظر شاخص‌های آماری برای هر دو روش برگزیده شد. باید توجه داشت که انتخاب مدل‌های با تعداد بالای توصیفگرها ضروری نیست، زیرا افزایش در تعداد توصیفگرهای مولکولی هیچ گونه اثر معناداری بر دقت و صحت بهترین مدل نخواهد داشت. همان طور که جدول ۲ نشان می‌دهد، در این تحقیق برای روش GA-MLR و مدل‌های ۸ توصیفگر دارای مجذور همبستگی اعتبار- سنجی خارجی بین ۱/۹۳۴۱- تا ۰/۶۷۳۴ و برای روش S-MLR بین ۰/۷۶۱۰- تا ۰/۶۶۴۲ است. بهترین ارتباط معنی‌دار موجود برای $\log(1/IC_{50})$ این دسته از ترکیبات ضدسرطان در مدل‌های به دست آمده از روش های S-MLR و GA-MLR برای دسته

ادامه جدول ۷. مقادیر توصیفگرهای مولکولی، مقادیر $\log(1/IC_{50})$ آزمایشی و پیشگویی شده توسط مدل GA-MLR

شماره	RDF055e	nNR2	Mor03p	nCp	nRSR	Mor07e	ISH	Mor16p	$\log(1/IC_{50})_{EXP}$	$\log(1/IC_{50})_{Prd}$	دسته
۶۹	۱۲/۵۴۶	۲	-۴/۱۰۳	۱	۱	۲/۷۷۷	۰/۹۸۴	-۰/۰۰۸	۰/۵۲۸۴	۰/۵۴۸۰	آزمون
۷۰	۱۱/۷۳۳	۲	-۳/۷۹۰	۱	۱	۳/۰۷۹	۰/۹۸۴	-۰/۱۷۲	۰/۵۹۶۷	۰/۷۵۳۶	واسنجی
۷۱	۱۴/۲۱۹	۲	-۴/۰۸۵	۲	۱	۲/۷۰۶	۱	-۰/۲۴۶	۱/۲۸۴۸	۱/۵۰۱۷	آزمون
۷۲	۱۶/۲۹۲	۲	-۴/۲۰۳	۰	۱	۳/۷۲۵	۱	-۰/۲۷۰	۱/۳۶۸۶	۰/۸۱۹۴	واسنجی
۷۳	۱۵/۴۲۴	۲	-۳/۶۶۳	۱	۱	۳/۴۱۴	۱	-۰/۱۵۷	۱/۳۳۲۵	۰/۶۵۸۴	واسنجی
۷۴	۱۳/۸۶۱	۲	-۳/۱۹۳	۲	۱	۳/۹۳۶	۰/۹۸۵	-۰/۵۲۵	۱/۷۲۸۲	۱/۳۴۵۴	واسنجی
۷۵	۱۲/۲۵۱	۲	-۳/۳۸۴	۲	۱	۳/۷۶۱	۱	-۰/۶۰۵	۱/۲۸۱۵	۲/۰۷۴۱	واسنجی
۷۶	۱۸/۵۴۶	۲	-۳/۶۸۳	۲	۱	۳/۸۲۵	۱	-۰/۶۱۰	۱/۲۲۰۴	۱/۶۳۴۵	واسنجی
۷۷	۱۵/۹۶۹	۲	-۳/۹۱۵	۲	۲	۲/۴۸۲	۰/۹۷	-۰/۴۲۱	۱/۳۱۹۷	۱/۶۰۰۳	آزمون
۷۸	۱۴/۹۶۳	۲	-۳/۱۵۱	۲	۱	۳/۷۵۸	۰/۹۷	-۰/۳۹۸	۱/۳۰۲۸	۰/۶۰۹۱	واسنجی
۷۹	۹/۸۱۰	۲	-۳/۴۸۷	۰	۱	۳/۱۵۱	۱	-۰/۱۳۲	۰/۷۱۴۹	۰/۵۸۸۵	واسنجی
۸۰	۱۲/۳۴۵	۲	-۳/۶۰۹	۰	۱	۱/۶۴۸	۱	-۰/۱۶۱	۰/۷۹۷۲	۰/۰۱۵۵	واسنجی
۸۱	۱۶/۶۵۸	۲	-۳/۹۶۱	۲	۱	۱/۶۵۵	۱	-۰/۳۰۲	۰/۶۷۷۶	۰/۹۲۳۵	واسنجی
۸۲	۲۳/۴۹۴	۳	-۳/۴۲۶	۱	۱	۳/۴۰۶	۱	۰/۱۴۰	-۰/۳۳۴۰	-۰/۳۵۹۱	واسنجی
۸۳	۱۷/۲۷۹	۲	-۳/۸۰۶	۱	۱	۲/۱۱۰	۰/۹۸۴	-۰/۰۹۲	-۰/۱۳۰۰	-۰/۲۶۱۹	آزمون
۸۴	۲۰/۳۸۷	۲	-۳/۹۲۲	۰	۱	۲/۹۶۷	۰/۹۸۶	۰/۰۵۷	-۱/۳۹۳۱	-۰/۹۱۱۹	واسنجی
۸۵	۱۲/۹۳۳	۲	-۳/۳۵۶	۰	۱	۱/۸۶۹	۱	-۰/۱۸۸	۰/۴۷۵۲	-۰/۱۳۷۹	واسنجی
۸۶	۱۱/۶۵۰	۱	-۳/۴۸۷	۰	۱	۳/۴۴۰	۰/۹۸۳	۰/۱۴۸	-۱/۲۵۸۱	-۰/۹۱۳۹	واسنجی
۸۷	۱۵/۳۳۳	۱	-۳/۶۶۶	۲	۱	۲/۸۵۶	۰/۹۶۹	۰/۰۷۴	-۱/۲۸۷۲	-۰/۶۵۰۵	آزمون
۸۸	۸/۳۶۲	۱	-۲/۱۶۸	۰	۱	۲/۷۹۴	۱	-۰/۰۵۰	-۰/۵۲۳۹	-۱/۰۷۰۵	واسنجی
۸۹	۹/۴۱۹	۱	-۲/۳۵۱	۰	۱	۱/۴۹۲	۱	-۰/۱۴۷	-۰/۸۸۸۶	-۱/۲۶۱۵	آزمون
۹۰	۱۱/۱۴۵	۱	-۳/۰۰۶	۰	۱	۲/۸۴۶	۱	-۰/۱۲۵	-۱/۱۰۶۹	-۰/۵۸۹۲	واسنجی
۹۱	۱۹/۵۲۵	۲	-۳/۳۵۷	۱	۱	۳/۵۸۴	۱	-۰/۳۱۱	۰/۴۷۲۲	۰/۲۶۹۳	واسنجی
۹۲	۸/۹۷۰	۲	-۳/۳۰۴	۰	۱	۲/۶۸۵	۰/۹۸۱	-۰/۳۰۳	۰/۲۷۷۳	۰/۲۸۲۵	واسنجی
۹۳	۱۷/۷۸۲	۲	-۴/۴۳۳	۰	۱	۴/۵۸۸	۰/۹۸۶	-۰/۲۳۵	۰/۷۶۸۸	۰/۷۲۵۷	واسنجی
۹۴	۱۱/۹۸۲	۲	-۳/۵۴۶	۰	۱	۲/۷۰۱	۱	-۰/۲۶۶	۱/۲۲۹۱	۰/۴۸۴۸	واسنجی
۹۵	۱۱/۲۷۲	۲	-۳/۳۷۷	۰	۱	۳/۲۹۹	۱	-۰/۱۶۶	۰/۱۴۳۶	۰/۴۳۹۷	واسنجی
۹۶	۱۰/۱۳۷	۲	-۳/۴۵۰	۰	۱	۲/۵۲۶	۱	-۰/۳۴۹	۰/۳۱۵۵	۰/۷۰۱۷	واسنجی
۹۷	۲۲/۴۷۵	۲	-۳/۷۰۹	۲	۱	۳/۸۴۵	۱	-۰/۳۷۴	۱/۰۹۵۸	۰/۸۳۴۰	آزمون
۹۸	۲۲/۸۲۴	۲	-۴/۱۱۱	۲	۱	۳/۶۲۶	۰/۹۸۵	-۰/۴۱۴	۰/۸۴۷۱	۰/۷۹۰۶	واسنجی
۹۹	۲۳/۵۰۱	۲	-۳/۴۵۹	۳	۱	۴/۰۶۸	۱	-۰/۴۵۹	۰/۸۲۸۶	۱/۱۶۹۹	واسنجی
۱۰۰	۳۶/۶۸۶	۲	-۴/۴۸۳	۳	۱	۵/۴۴۸	۱	-۰/۲۸۷	۰/۶۸۲۱	۰/۶۱۶۳	واسنجی

تصادفی ۱۵ (به عنوان بهترین دسته تصادفی) به صورت زیر است.

برای روش S-MLR معادله حاصل به صورت زیر است:

$$Log\left(\frac{1}{IC_{50}}\right) = 13.046\ (\pm 2.567) + 0.514\ (\pm 0.096)RDF020p + 0.564(\pm 0.161)C - 002 - 1.418(\pm 0.416)Mor16m - 4.582(\pm 0.983)R4m - 1.797(\pm 0.495)Mor21m + 65.981(\pm 14.679)G2e + 0.680(\pm 0.235)nRNR2 + 1.040(\pm 0.406)Mor22m \quad (۳)$$

Q^2_{CV} =0.6108; Q^2_{Cal}=0.6889; Q^2_{Test}=0.4110

$$Log\left(\frac{1}{IC_{50}}\right) = 25.437\ (\pm 8.325) - 0.111\ (\pm 0.016)RDF055e + 0.556(\pm 0.200)nRNR2 - 0.777(\pm 0.144)Mor03p + 0.439(\pm 0.094)nCp + 0.840(\pm 0.243)nRSR + 0.289(\pm 0.096)Mor07e + 21.324(\pm 8.322)ISH - 1.654(\pm 0.479)Mor16p \quad (۴)$$

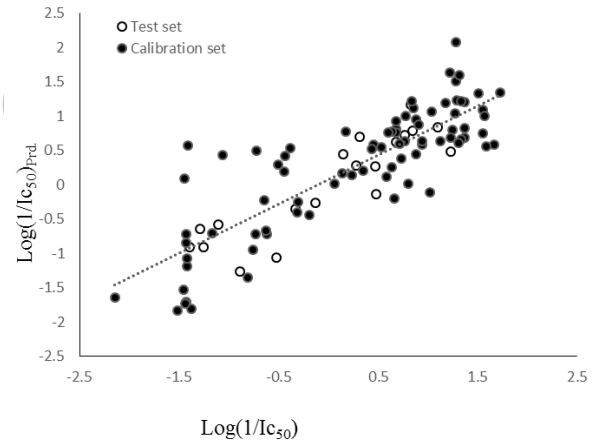
Q^2_{CV} =0.6650; Q^2_{Cal}=0.7183; Q^2_{Test}=0.6734

نتایج مربوط به روش‌های S-MLR و GA-MLR در جدول‌های ۳ و ۴ نشان داده شده‌اند. جدول‌های ۵ و ۶ نشان دهنده ماتریس ضرایب همبستگی خطی برای مقادیر $\log(1/IC_{50})$ و هشت متغیر موجود در مدل‌های MLR هستند. نتایج ماتریس ضریب همبستگی بیانگر این موضوع است که بین توصیفگرهای انتخاب شده همبستگی وجود نداشته و توصیف گرها مستقل از هم

هستند و نتایج به دست آمده از مدلسازی دال بر وابستگی توصیف کننده‌ها نیست. همان طور که در جدول ۴ و ۶ مشاهده می‌شود ضریب همبستگی هر جفت توصیف کننده در روش GA-MLR کمتر از ۰/۴۸۵ است و برای مدل S-MLR ضریب همبستگی هر جفت توصیفگرها کمتر از ۰/۴۳۲ است.

ضریب رگرسیونی استاندارد حاصل از مدلسازی GA-MLR اهمیت نسبی توصیفگرها را نشان می دهد. مهمترین توصیفگر RDF055e است و سایر توصیفگرها به ترتیب اهمیت عبارتند از: Mor03p>nCp>Mor16p>nRSR>Mor07e>nRNR2>ISH

مقادیر توصیفگرهای مولکولی، مقادیر $\log(1/IC_{50})$ آزمایشی و پیشگویی شده توسط مدل GA-MLR برای دسته واسنجی و دسته آزمون در جدول ۷ می توان مشاهده کرد؛ به علاوه نمودار ۱، نمودار مقادیر $\log(1/IC_{50})$ پیشگویی شده در برابر آزمایشی مربوط به مدل GA-MLR برای دسته واسنجی و آزمون ترکیبات ضدسرطان را نشان می دهد.



نمودار ۱. نمودار مقادیر log(IC50) پیشگویی شده در برابر آزمایشی مربوط به مدل GA-MLR برای دسته واسنجی و آزمون ترکیبات ضدسرطان

بحث

با تفسیر و ارزیابی توصیفگرهای بهترین مدل GA-MLR این امکان برای بدست آوردن بینش شیمیایی مفید برای طراحی دارو وجود دارد. به همین دلیل، تفسیر قابل قبول از نتایج QSAR در زیر ارائه شده است.

توصیفگر RDF055e متعلق به دسته توصیفگرهایRDF، توصیف گر تابع توزیع شعاعی است که توسط الکترونگاتیویته ساندرسون وزن شده است (۲۹). ضریب منفی این توصیفگر در معادله QSAR نشان دهنده تاثیر منفی بر روی $Log(1/IC_{50})$ و در نتیجه تاثیر مثبت بر روی IC_{50} دارد و با توجه به آنکه IC_{50} رابطه معکوس با فعالیت دارد با افزایش آن این توصیفگر، فعالیت ترکیبات کاهش می یابد. توصیف‌گرهای nRNR2، nCp و nRSR متعلق به توصیفگرهای شمارش گروه عاملی هستند که به ترتیب با شمارش تعداد آمین نوع سوم، تعداد کربن نوع اول انتهایی sp^3 و تعداد سولفیدها محاسبه می‌شوند. این گروه جزو ساده‌ترین توصیفگرهای دراگون هستند. علامت مثبت این ضرایب بیانگر اثر مثبت این متغیرها بر روی $\log(1/IC_{50})$ و درنتیجه با افزایش این توصیفگرها فعالیت این ترکیبات افزایش می‌یابد.

توصیفگرهای Mor03p، Mor16p و Mor07e متعلق به دسته توصیفگرهای سه بعدی مورس (3D-MoRSE) هستند که به ترتیب توسط قطبش و الکترونگاتیویته ساندرسون توزین شده است. توصیفگرهای سه بعدی مورس (نمایش سه بعدی مولکول از سازه های مبتنی بر پراش الکترونی) از شبیه سازی طیف‌های مادون قرمز با استفاده از یک تابع پراکندگی عمومی مشتق شده است (۳۰). توصیفگرهای Mor03p و Mor16p به ترتیب نمایانگر سیگنال ۳ و ۱۶ از توصیفگرهای 3D-MoRSE هستند که با استفاده از قطبش پذیری توزین شده است و ضرایب منفی این دو توصیفگر نشان دهنده اثر منفی بر روی $\log(1/IC_{50})$ و در نتیجه اثر منفی بر روی فعالیت دارو است. توصیفگر Mor07e نمایانگر سیگنال ۷ از توصیفگرهای 3D-MoRSE است که با استفاده از الکترونگاتیویته ی ساندرسون توزین شده است و ضرایب مثبت آن نشان دهنده اثر منفی بر روی فعالیت دارو است. توصیفگر ISH متعلق به توصیفگرهای GETAWAY است. توصیفگرهای GETAWAY از جمله توصیفگرهایی هستند که اخیراً ارایه شده و مربوط به ساختار شیمیایی مولکول‌ها هستند. این گروه از توصیفگرها به منظور جور نمودن هندسه سه بعدی مولکول طراحی شده‌اند و از عناصر h_{ij} از ماتریس نفوذ مولکولی (H) به دست آمده از طریق مختصات کارتزین اتمی، مشتق شده‌اند (۳۱،۳۲). توصیفگر ISH دارای ضریب مثبت است که نشان دهنده تاثیر مثبت این متغیر بر روی $\log(1/IC_{50})$ و در نتیجه تاثیر منفی بر روی IC_{50} است و با افزایش این توصیفگر فعالیت ترکیب افزایش می‌یابد.

در این تحقیق مدلسازی QSAR برای رابطه ساختار- فعایت ترکیبات ۱و ۴-دی هیدرو-۴-اکسو-۱-(۲-تیازول)-۱و۸-نفتیریدین و مشتقات آن به دست آمد. داده‌ها به روش تصادفی به دو دسته واسنجی و آزمون شکافته شدند. تاثیر انتخاب دسته‌های تصادفی متفاوت به روش تصادفی بر روی میزان پیشگویی مدل مورد بررسی قرار گرفت. انتخاب بهترین توصیف گرها از میان توصیفگرهای موجود توسط روش های S-MLR و GA-MLR انجام شد. آماره اعتبارسنجی خارجی گزارش شده برای هر مدل،

مبنایی برای مقایسات نهایی است. نتایج این مقایسه نشان می‌دهد که مدل های QSAR به دست آمده با روش GA-MLR دارای مجذور ضریب همبستگی اعتبارسنجی بزرگتری نسبت به روش S-MLR می باشند. نتایج نشان می دهند که مقادیر Q^2_{test} برای روش‌های S-MLR و GA-MLR به ترتیب ۰/۴۱۱۰ و ۰/۶۷۳۴ است. مطابق نتایج حاصل از هر دو روش می توان نتیجه گرفت که روش GA-MLR نسبت به S-MLR روش قدرتمندتری است و نتایج به دست آمده می تواند در طراحی و سنتز داروهای ضد سرطان بسیار سودمند باشد.

REFERENCES

1. Funatsu K, Miyao T, Arakawa M, Systematic generation of chemical structures for rational drug design based on QSAR models. Curr Comput Aided Drug Des 2011; 7: 1-9.
2. Ahmadi S, Habibpour E, Application of GA-MLR for QSAR modeling of the arylthioindole class of tubulin polymerization inhibitors as anticancer agents. Anticancer Agents Med Chem 2017; 17: 552-65.
3. Ahmadi S, Ganji S, Genetic algorithm and self-organizing maps for QSPR study of some N-aryl derivatives as butyrylcholinesterase inhibitors. Curr Drug Discov Technol 2017; 13: 232-53.
4. Ahmadi S, Khazaei MR, Abdolmaleki A, Quantitative structure-property relationship study on the intercalation of anticancer drugs with ct-DNA. Med Chem Res 2014; 23:1148-61.
5. Spiegel K, Magistrato A, Modeling anticancer drug–DNA interactions via mixed QM/MM molecular dynamics simulations. Org Biomol Chem 2006; 4: 2507-17.
6. Goodarzi M, Dejaegher B, Vander Heyden Y, Feature selection methods in QSAR studies. J AOAC Int 2012; 95: 636–651.
7. Gonzalez MP, Teran C, Saiz-Urra L, Teijeira M, Variable selection methods in QSAR: an overview. Curr Topics Med Chem 2008; 8: 1606–27.
8. Leardi R, Genetic algorithms in chemometrics and chemistry: a review. J Chemometr 2001; 15: 559-69.
9. Lucasius CB, Beckers MLM, Kateman G, Genetic algorithms in wavelength selection: a comparative study. Anal Chim Acta 1994; 286: 135-53.
10. Shayanfar A, Ghasemi S, Soltani S, Asadpour-Zeynali K, Doerksen RJ, Jouyban A, Quantitative structure-activity relationships of imidazole-containing farnesyltransferase inhibitors using different chemometric methods. Med Chem 2013; 9: 434-48.
11. Bohari MH, Srivastava HK, Sastry GN, Analogue-based approaches in anti-cancer compound modelling: the relevance of QSAR models. Org Med Chem Lett 2011; 1: 1-12.
12. Bertoša B, Aleksić M, Karminiski-Zamola G, Tomić S, QSAR analysis of antitumor active amides and quinolones from thiophene series. Int J Pharm 2010; 394: 106-14.
13. Tomita K, Tsuzuki Y, Shibamori K, Tashima M, Kajikawa F, Sato Y, et al. Synthesis and structure-activity relationships of novel 7-substituted 1,4-dihydro-4-oxo-1-(2thiazolyl)-1,8-naphthyridine-3-carboxylic acids as antitumor agents. Part1. J Med Chem 2002; 45: 5564-75.
14. Tsuzuki Y, Tomita K, Shibamori K, Sato Y, Kashimoto S, Chiba K, Synthesis and structure-activity relationships of novel 7-substituted 1,4-dihydro-4-oxo-1-(2-thiazolyl)-1,8-naphthyridine-3-carboxylic acids as antitumor agents. Part 2. J Med Chem 2004; 47: 2097-109.
15. Tsuzuki Y, Tomita K, Sato Y, Shibamori K, Kashimoto S, Chiba K, Synthesis and structure-activity relationships of 3- substituted 1,4-dihydro-4-oxo-1-(2-thiazolyl)-1,8-naphtyridines as novel antitumor agents. Bioorg Med Chem Lett 2004; 14: 3189-93.
16. Ahmadi S, Babaee E, khazaei MR, Application of self organizing maps and GA-MLR for the estimation of stability constant of 18-crown-6 ether derivatives with sodium cation. J Incl Phenom Macrocycl Chem 2014; 79: 141-49.
17. Ahmadi S, Deligeorgiev TG, Vasilev A, Kubista M, The dimerizationtion study of some cationic monomethine cyanine dyes by chemometrics method. Russ J Phys Chem A 2012; 86: 1974-81.
18. Ahmadi S, Application of GA-MLR method in QSPR modeling of stability constants of diverse 15-Crown-5 complexes with sodium cation. J Incl Phenom Macrocycl Chem 2012; 74: 57-66.
19. Ahmadi S, A QSPR study of association constants of macrocycles toward sodium cation. Macroheterocycles 2012; 5: 23-31.
20. Ghasemi JB, Ahmadi S, Brown SD, A quantitative structure-retention relationship study for prediction of chromatographic relative retention time of chlorinated monoterpenes. Environ Chem Lett 2011; 9: 87-96

21. Ghasemi JB, Ahmadi S, Ayati M, QSPR modeling of stability constants of the Li-hemispherands complexes using MLR: a theoretical host-guest study. Macroheterocycles 2010; 3: 234-42.

22. Ghasemi JB, Ahmadi S, Combination of genetic algorithm and partial least squares for cloud point prediction of nonionic surfactants from molecular structures. Ann Chim Rome 2007; 97: 69-83.

23. Rogers D, Hopfinger AJ, Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships J Chem Inf Comput Sci 1994; 34: 854-66.

24. Cho SJ, Hermsmeier MA, Genetci algorithm guided selection: variable selection and subset selection. J Chem Inf Comput Sci 2002; 42: 927-36.

25. Myers RH, Classical and modern regression with applications. Boston: PWS-KENT Publishing company; 1990.

26. Hintze J, NCSS – Number Cruncher Statistical System. UT: NCSS: Kaysville; 2001.

27. Jaiswal M, Khadikar PV, Scozzafava A, Supuran CT, Carbonic anhydrase inhibitors: the first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides. Bioorg Med Chem Lett 2004; 14: 3283-90.

28. Shapiro S, Guggenheim B, Inhibition of oral bacteria by phenolic compounds part1 QSAR analysis using molecular connectivity Quant Struct Act Relat 1998; 17: 327-37.

29. Hemmer MC, Steinhauer V, Gasteiger J, Driving the 3D structure of organic molecules from their infrared spectra. Vib Spectrosc 1999; 19: 151-64.

30. Gasteiger J, Sadowski J, Schuur J, Selzer P, Steinhauer L, Steinhauer V, Chemical information in 3D space. J Chem Inf Comput Sci 1996; 36: 1030-37.

31. Consonni V, Todeschini R, M. Pavan, Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Theory of the novel 3D molecular descriptors. J Chem Inf Comput Sci 2002; 42: 682-92.

32. Consonni V, Todeschini R, Pavan M, Gramatica P, Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. J Chem Inf Comput Sci 2002; 42: 693-705.