JARC

روش‌های خطی و غیرخطی ارتباط کمی‌ساختار- فعالیت جهت پیش‌بینی فعالیت دارویی برخی از مشتقات آمینواسیدها

مهدی نکویی[1و*]، مجید محمدحسینی[2]، مهدی رحیمی[3] و عبدالرضا علوی‌قره‌باغ[4]

۱- استادیار شیمی تجزیه، گروه شیمی، دانشگاه آزاد اسلامی، واحد شاهرود، شاهرود، ایران
۲- دانشیار شیمی تجزیه، گروه شیمی، دانشگاه آزاد اسلامی، واحد شاهرود، شاهرود، ایران
۳- کارشناس ارشد شیمی فیزیک، گروه شیمی، دانشگاه آزاد اسلامی، واحد شاهرود، شاهرود، ایران
۴- مربی، دانشجوی دکترای برق، گروه برق، دانشگاه آزاد اسلامی، واحد شاهرود، شاهرود، ایران

دریافت: شهریور ۱۳۹۱، بازنگری: آبان ۱۳۹۱، پذیرش: آذر ۱۳۹۱

چکیده: این پژوهش به پیش‌بینی فعالیت دارویی ۳۸ مشتق آمینواسید به عنوان بازدارنده‌های هیستون دی استیلاز (HDAC) جهت درمان سرطان و برخی از بیماری‌ها اختصاص دارد. آنزیم‌های HDAC موجب تسریع روند حذف گروه‌های استیل از باقیمانده‌های لیزین از پروتیین‌های شامل هیستون (Histone) می‌شوند. پس از محاسبه‌ی توصیف‌کننده‌های مولکولی مستقل، با استفاده از روش مرحله‌ای انتخاب متغیر و گزینش ۴ توصیف‌کننده، جهت مدل‌سازی از رگرسیون خطی چندگانه (MLR) و شبکه‌ی عصبی مصنوعی (ANN) استفاده شد. سری‌های آموزش و آزمون جهت ساخت مدل و ارزیابی قدرت پیش‌بینی روش‌های MLR و ANN به ترتیب شامل ۳۰ و ۸ ترکیب بودند. افزون‌بر آن، از روش‌های متفاوت جهت ارزیابی مدل‌ها استفاده شد. نتیجه‌ها حاکی از آن است که روش غیرخطی شبکه‌ی عصبی مصنوعی در مجموع دارای توانمندی پیش‌بینی مناسب‌تر در مقایسه با روش MLR است. شاخص‌های آماری مرتبط با مدل مبتنی بر شبکه‌ی عصبی مصنوعی دلالت بر این حقیقت دارد که مدل ارایه شده می‌تواند جهت پیش‌بینی فعالیت دارویی ترکیب‌های مشابه مورد استفاده قرار گیرد.

واژه‌های کلیدی: ارتباط کمی‌ساختار- فعالیت، مشتقات آمینواسیدها، رگرسیون خطی چندگانه، شبکه عصبی مصنوعی، هیستون دی استیلاز (HDAC)

مقدمه

یکی از مشکلاتی که جامعه بشری همیشه با آن روبه‌رو بوده، مقابله با انواع بیماری‌هایی است که سلامت انسان‌ها را به مخاطره انداخته و همواره یکی از مهم‌ترین دغدغه‌های پژوهشگران، یافتن داروهای مؤثر، برای رفع این معضل و یا 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 به خود جلب کرده‌اند [۱]. روند کشف و توسعه داروهای جدید مبتنی بر روش آزمون و خطا، وقت گیر، طاقت‌فرسا و هزینه‌بر است. مشکل دیگری که در این راه دانشمندان را آزار می‌دهد، عدم اطلاع آن‌ها از فعالیت دارویی ترکیب‌ها، قبل از سنتز و بررسی تجربی آن‌ها بوده و به همین دلیل یکی از مهم‌ترین هدف‌های شیمیدان‌ها و پژوهشگران دارویی ارزیابی مقدماتی فعالیت داروها، قبل از ساخت آن‌ها

*عهده‌دار مکاتبات: m_nekoei1356@yahoo.com

است. از این رو نیاز به استفاده از روش‌های نظری و محاسباتی که بدون انجام آزمایش بتوانند ویژگی و یا فعالیت ترکیب‌های دارویی را پیش‌بینی کنند اجتناب‌ناپذیر به نظر می‌رسد. ظهور علم کمومتریکس توانسته راه حلی برای رفع این مشکلات باشد [۲ تا ۵].

یکی از مهم‌ترین زمینه‌های کاربرد روش‌های کمومتریکس مطالعه ارتباط بین ویژگی‌های مولکول‌ها با ویژگی‌های ساختاری آن‌هاست. این نوع مطالعات که با عنوان ارتباط کمی‌ساختار- فعالیت (QSAR)[1] معروف شده‌اند، به بررسی نحوه‌ی ارتباط بین ویژگی‌های متفاوت مولکول‌ها با مشخصات ساختاری و ذاتی آن‌ها می‌پردازند. بررسی ساختار شیمیایی و فعالیت ترکیب‌ها، پیش‌بینی فعالیت ترکیب‌های جدید را بر اساس اطلاعات مرتبط به ساختار شیمیایی آن‌ها امکان‌پذیر می‌سازد.

در سالیان اخیر، روند رو به افزایش مقالات منتشره در منابع علمی‌بر اساس QSAR، دلالت بر جایگاه منحصر به فرد این دیدگاه در شیمی‌نظری و تعمیق بینش دانشمندان در فهم و توجیه سازوکارهای دخیل در فعالیت گستره‌ی وسیعی از ترکیب‌ها دارد [۶ تا ۸]. از جمله روش‌هایی که در مطالعات QSAR جهت مدل‌سازی و پیش‌بینی فعالیت ترکیب‌های دارویی مورد استفاده قرار می‌گیرند می‌توان به روش‌های خطی از جمله رگرسیون خطی چندگانه (MLR)[2] و روش‌های غیرخطی مانند شبکه‌ی عصبی مصنوعی اشاره کرد. رگرسیون خطی چندگانه روشی توانمند است که جهت ایجاد رابطه‌ای خطی بین یک متغیر وابسته مانند Y و مجموعه‌ای از متغیرهای مستقل نظیر X_1, X_2, X_3, …, X_n به کار می‌رود. رابطه‌ی خطی این رابطه به صورت زیر است [۹].

$$Y= b_0+b_1X_1+ b_2X_2+ \ldots\ldots +b_nX_n \qquad (۱)$$

که در آن n، تعداد متغیرها و b_i، i امین ضریب رگرسیون و X_i، مقدار عددی i امین متغیر مستقل است. شبکه‌های عصبی مصنوعی (ANN)[3] ساختارهای محاسباتی جدیدی هستند که قادر به پردازش اطلاعات و یادگیری براساس اصول الهام گرفته شده از سامانه عصبی موجودات زنده هستند.

این شبکه‌های مصنوعی متشکل از مجموعه‌ای از نرون‌ها[4] با ارتباطات داخلی بین یکدیگر هستند که قادرند بر اساس اطلاعات و داده‌های ورودی، جواب‌های خروجی را ایجاد کنند. شبکه‌های عصبی به‌طور معمول به صورت لایه لایه و منظم ایجاد می‌شوند. نخستین لایه که اطلاعات و داده‌های ورودی به آن وارد می‌شود، لایه‌ی ورودی و آخرین لایه که جواب‌های خروجی را ایجاد می‌کند لایه‌ی خروجی نامیده می‌شود. بین این دو لایه یک یا چند لایه دیگر نیز وجود دارد که به آن‌ها لایه‌های مخفی گفته می‌شود [۱۰]. ساده‌ترین و معمول‌ترین نوع شبکه‌ی عصبی که بسیار از آن استفاده می‌شود، شبکه‌ی عصبی چند لایه پیشخور (MLP)[5] همراه با ناظر است که در آن معمولاً از روش پس‌انتشار خطا جهت آموزش شبکه استفاده می‌شود.

در این شبکه تعداد نرون‌های لایه ورودی برابر با تعداد عناصر بردار ورودی (تعداد توصیف‌کننده‌ها)[6] و تعداد نرون‌های لایه‌ی خروجی برابر با تعداد عناصر بردار خروجی (فعالیت ترکیب‌ها) است. آنالیز دقیق و واقعی برای پیدا کردن تعداد نرون‌های لایه‌ی میانی در کل بسیار پیچیده است. اما می‌توان گفت که تعداد نرون‌های لایه میانی تابعی از تعداد عناصر بردار ورودی و هم‌چنین حداکثر تعداد نواحی از فضای ورودی که به طور خطی از هم جدا پذیرند، است. از این رو تعداد نرون‌های لایه مخفی به‌طور معمول به طور تجربی به‌دست می‌آید. هر نرون توسط خروجی‌اش به نرون‌های لایه‌ی بعد متصل است، ولی با نرون‌های لایه خودش ارتباط ندارد. خروجی هر نرون توسط رابطه‌ی زیر تعریف می‌شود:

$$\alpha_j = f(\sum_{i=1}^{n} P_i W_{j,i} + b_j) \qquad (۲)$$

که در این رابطه $W_{j,i}$ برابر با مقدار وزن اتصال بین نرون jام لایه مذکور با نرون iام لایه قبل است که بیانگر اهمیت ارتباط بین دو نرون در دو لایه‌ی متوالی است، b_j برابر است با وزن مربوط به بایاس[7] برای نرون jام، P_i برابر است با مقدار خروجی از نرون iام لایه قبل، α_j برابر است با مقدار خروجی از نرون jام و f برابر با تابع آستانه نرون jام است. توابع زیادی در شبکه‌های عصبی مصنوعی مورد استفاده قرار می‌گیرند. از جمله‌ی آن‌ها می‌توان به

1. Quantitative structure-activity relationship (QSAR)
2. Multiple linear regression (MLR)
3. Artificial neural network
4. Neuron
5. Multi-layer perceptron (MLP)
6. Descriptors
7. Biass

توابع سیگموید، گوسی، تانژانت هیپربولیک و غیره اشاره کرد. ولی تابع سیگموئیدی بیشترین استفاده را در علوم متفاوت دارد. این تابع به صورت زیر است:

$$f(z)=\frac{1}{1+\exp(-z)} \qquad (۳)$$

شبکه‌های عصبی مصنوعی شامل دو مرحله هستند. مرحله‌ی اول آموزش شبکه است که در طی آن مقدارهای وزن‌ها و بایاس‌های شبکه به نحوی تنظیم می‌شوند که به ازای یک دسته ورودی معین پاسخ مطلوبی ایجاد شود. مرحله‌ی بعدی فاز عملیاتی است که در طی آن از شبکه‌ی آموزش یافته برای حل مسایل مربوط استفاده می‌شود [۱۱ تا ۱۷].

هدف اصلی این پژوهش ارایه‌ی مدل‌های مناسب جهت پیش‌بینی فعالیت دارویی برخی از مشتقات آمینواسیدها با استفاده از روش‌های MLR و ANN است. هم‌چنین براساس روش‌های معمول در مطالعات با استفاده از الگوهای ارزشیابی تقاطعی و خارجی قدرت پیش‌بینی مدل‌ها بررسی شده است.

محاسبات

انتخاب سری داده‌ها

در این پژوهش، تعداد ۳۸ ترکیب از مشتقات آمینواسیدها که به عنوان بازدارنده‌های آنزیم هیستون دی استیلاز جهت اهداف درمانی در معالجه سرطان در سال ۲۰۰۸ در گزارش هوبز و همکارانش ارایه شده [۱۸] توسط روش‌های کمومتریکس مورد بررسی قرار گرفته است. در فرایند مدل‌سازی، قدرت بازدارندگی این ترکیب‌ها بر حسب IP[1] به مقیاس لگاریتمی ($pIP=-\log(IP)$) تبدیل و مورد استفاده قرار گرفته است. در این کار پس از تقسیم تصادفی ترکیب‌ها سری داده به دو گروه سری آموزش (۳۰ مولکول = ۸۰٪ ترکیب‌ها) و سری پیش‌بینی (۸ مولکول = ۲۰٪ ترکیب‌ها)، مقدارهای pIP به عنوان متغیر وابسته و توصیف‌کننده‌ها به عنوان متغیر مستقل انتخاب شد. سری آموزش جهت ایجاد یک مدل مناسب و سری پیش‌بینی جهت ارزیابی مدل مورد استفاده قرار گرفت.

محاسبه و کاهش تعداد توصیف‌کننده‌ها

در ابتدا برای محاسبه‌ی توصیف‌کننده‌ها، ساختار ترکیب‌ها به کمک نرم‌افزار هایپرکم[2] [۱۹] رسم شد. سپس ساختارهای مولکولی ترسیم شده به وسیله‌ی الگوریتم AM1 بهینه شدند. ساختارهای بهینه شده به نرم‌افزار دراگون[3] [۲۰] منتقل و توصیف‌کننده‌ها به تعداد ۱۴۹۷ مورد به کمک این نرم‌افزار محاسبه شدند. این نرم‌افزار برای محاسبه‌ی هجده دسته از توصیف‌کننده‌های مولکولی مورد استفاده قرار می‌گیرد.

یکی از مشکلاتی که در هنگام ایجاد مدل‌های QSAR با آن مواجه می‌شویم، تعداد زیاد متغیرهای مستقل است. در بیشتر موردها تعداد توصیف‌کننده‌ها از تعداد مولکول‌ها بسیار بیشتر است. در این صورت استفاده از روش‌های حداقل مربعات باعث ایجاد مشکلاتی نظیر انتخاب شانسی و هم‌بستگی[4] تصادفی می‌شود. با توجه به این که بعضی از متغیرهای مستقل، ثابت بوده و هم‌چنین برخی دیگر با یکدیگر هم‌بستگی نشان می‌دهند، بنابراین، به روش زیر برخی از متغیرها حذف شدند.

۱- توصیف‌کننده‌هایی که مقدارهای ثابت و یا به تقریب ثابت دارند (بیش از ۹۰٪ داده‌های ثابت دارند)، حذف شدند. در این مرحله تعداد ۳۴۶ توصیف کننده حذف و بدین ترتیب ۱۱۵۱ توصیف‌کننده باقی ماند.

۲- توصیف‌کننده‌هایی که با یکدیگر همبستگی بالای ۰/۹ دارند مورد بررسی قرار گرفتند و بین آن‌ها، توصیف‌کننده‌ای که همبستگی کمتری با متغیر مستقل داشت حذف شد. بدین ترتیب تعداد ۶۹۰ توصیف کننده، حذف و تعداد ۴۶۱ توصیف‌کننده باقی ماند.

انتخاب توصیف‌کننده‌های مؤثر

مهم‌ترین بخش در ایجاد یک مدل کار آمد، انتخاب توصیف‌کننده‌های مناسب است. پس از محاسبه‌ی توصیف‌کننده‌های متفاوت، تعدادی از آن‌ها به عنوان توصیف‌کننده‌های مناسب برای ساخت مدل انتخاب می‌شوند. این مرحله، شامل یافتن

1. Inflection point 2. Hyperchem 3. Dragon 4. Correlation

توصیف‌کننده‌های حاوی اطلاعات مفید است به طوری که قدرت پیش‌بینی مدل در سطح قابل قبولی باشد [۷ و ۸].

در این پژوهش از روش افزایش مرحله‌ای برای انتخاب توصیف‌کننده‌های مناسب استفاده شد. با آزمایش همه توصیف‌کننده‌ها، فرایند انتخاب تا زمانی ادامه می‌یابد که مدلی با ضریب همبستگی بالا (به‌طور معمول در حدود ۰/۷ تا ۰/۹) به دست آید. اگر ضریب همبستگی به مقدار بالای ۰/۹۵ برسد، می‌توان گفت که مدل خطی، مدل مناسبی برای توصیف سامانه مورد بررسی است و با استفاده از این مدل می‌توان پیش‌بینی را به نحو مطلوب انجام داد. اما در اغلب سامانه‌های موجود، متغیرهای مؤثر بر عامل مورد بررسی، رفتاری غیر خطی از خود نشان می‌دهند، در چنین مواردی، برای ارایه مدلی که شرایط غیر خطی سامانه را نیز در نظر بگیرد، از مدلی غیر خطی مانند شبکه‌ی عصبی مصنوعی استفاده می‌شود.

با روش مرحله‌ای، از بین ۴۶۱ توصیف‌کننده باقی مانده تعداد ۴ توصیف کننده به عنوان مناسب‌ترین آن‌ها انتخاب و با روش‌های MLR و ANN مدل‌سازی انجام شد. لیست توصیف‌کننده‌های انتخاب شده به‌وسیله‌ی روش افزایش مرحله‌ای به همراه توصیف مختصری از آن‌ها در جدول ۱ آورده شده است.

جدول ۱ توصیف‌کننده‌های انتخاب شده با روش افزایش مرحله‌ای و توصیف آن‌ها

نوع توصیف کننده	توصیف کننده
RDF	RDF020m [1]
RDF	RDF070m [2]
WHIM	E1m [3]
GETAWAY	R4m+ [4]

1. Radial Distribution Function - 2.0 / weighted by atomic masses
2. Radial Distribution Function - 7.0 / weighted by atomic masses
3. 1st component accessibility directional WHIM index / weighted by atomic masses
4. R maximal autocorrelation of lag 4 / weighted by atomic masses

ارزشیابی توصیف‌کننده‌های انتخاب شده

یکی از شرایط مهم انتخاب توصیف‌کننده‌ها، لزوم رفتار مستقل آن‌ها از یکدیگر است. چرا که در صورت وابستگی بالای آن‌ها، تنها توصیف‌کننده‌ای در محاسبات نهایی وارد خواهد شد که همبستگی بیشتری با متغیر وابسته (فعالیت) داشته باشد. این کار به سادگی از طریق گزینه‌ی Bivariate analysis توسط نرم‌افزار SPSS[1] انجام شده و میزان همبستگی دو به دوی توصیف‌کننده‌ها در جدول ۲ آورده شده است. با توجه به جدول ۲ بیشترین ضریب همبستگی بین توصیف کننده‌های RDF020m و E1m مقدار عددی ۰/۴۹۳- است. این نتیجه‌ها دلالت بر عدم همبستگی قابل ملاحظه و رفتار مستقل توصیف‌کننده‌های انتخاب شده دارد.

جدول ۲ ماتریس ضرایب همبستگی توصیف‌کننده‌های انتخاب شده.

	RDF020m	RDF070m	E1m	R4m+
RDF020m	۱	۰	۰	۰
RDF070m	-۰/۰۵۴	۱	۰	۰
E1m	-۰/۴۹۳	-۰/۲۷۶	۱	۰
R4m+	-۰/۱۸۵	۰/۱۱۸	۰/۱۳۴	۱

ایجاد مدل با استفاده از روش رگرسیون خطی چندگانه (MLR)

پس از انتخاب مناسب‌ترین توصیف‌کننده‌ها توسط روش افزایش مرحله‌ای، برقراری ارتباط مناسب بین توصیف‌کننده‌های منتخب و فعالیت ترکیب‌های دارویی با استفاده از نرم‌افزار SPSS صورت گرفت. معادله‌ی خطی ساخته شده (معادله‌ی ۴) جهت پیش‌بینی فعالیت‌های بازدارندگی مشتقات آمینواسیدها به صورت زیر است.

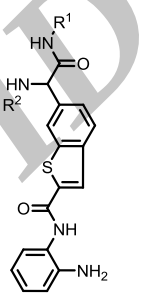
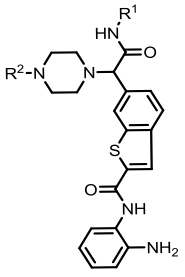
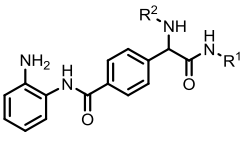
$$\begin{aligned} pIP = {} & 13.837(\pm 0.673) \\ & -1.334(\pm 0.266)\ \mathbf{RDF020m} \\ & -0.026(\pm 0.009)\ \mathbf{RDF070m} \\ & -6.521(\pm 0.857)\ \mathbf{E1m} \\ & -42.680\ (\pm 3.724)\ \mathbf{R4m+} \end{aligned} \qquad (۴)$$

سپس از معادله‌ی به‌دست آمده برای پیش‌بینی فعالیت سری آزمون استفاده شد. مقدارهای واقعی و پیش‌بینی‌شده‌ی فعالیت‌ها برای تمام ترکیب‌ها در جدول ۳ آورده شده است. هم‌چنین نمودار مقدارهای فعالیت‌های پیش‌بینی شده بر حسب مقدارهای تجربی برای دو مجموعه‌ی آموزش و آزمون در شکل ۱ نشان داده شده است.

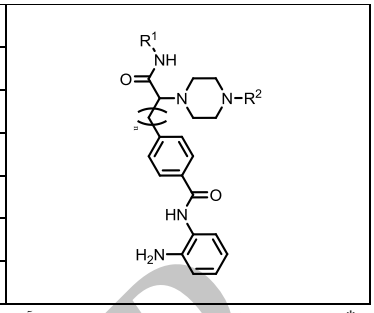
شکل ۲ مقدارهای باقیمانده‌ی خطاها را نسبت به مقدارهای

1. Statistical package for the social sciences (SPSS)

جدول ۳ مقدارهای تجربی و محاسبه شده‌ی pIP برای ترکیب‌های متفاوت برای مجموعه‌های آموزشی و پیش‌بینی در مدل‌های MLR و ANN

شماره ترکیب	R^1	R^2	Exp.	MLR	ANN	شکل ترکیب
۱	PMP	PMB	۷٫۰۵	۷٫۱۲	۷٫۱۹	
۲*	PMP	e-C_6H_{11}	۷٫۱۳	۷٫۱۳	۷٫۲۴	
۳	PMP	CH_2CH_2Ph	۷٫۱۵	۷٫۰۷	۷٫۲۳	
۴	PMP	Bn	۷٫۱۰	۷٫۱۹	۷٫۱۴	
۵	PMP	CH_2CH_2-4-pyridyl	۷٫۴۹	۷٫۴۳	۷٫۳۵	
۶	CH_2-CH_2-1-morpholion	CH_2CH_2Ph	۶٫۶۴	۶٫۵۴	۶٫۴۸	
۷	Bn	CH_2CH_2Ph	۷٫۱۵	۷٫۲۱	۷٫۰۵	
۸	Bn	CH_2CH_2OMe	۶٫۷۷	۶٫۷۴	۶٫۷۰	
۹	Bn	CH_2CH_2OH	۶٫۸۲	۶٫۸۶	۶٫۷۸	
۱۰	Bn	$CH_2CH_2NMe_2$	۶٫۸۲	۷٫۰۷	۶٫۹۷	
۱۱	Bn	CH_2CH_2-2-imidazolyl	۶٫۲۸	۶٫۴۶	۶٫۳۲	
۱۲	Bn	CH_2-3-pyridyl	۶٫۷۴	۷٫۱۰	۶٫۹۵	
۱۳	Bn	CH_2CH_2-4-pyridyl	۷٫۴۱	۷٫۲۳	۷٫۲۱	
۱۴*	Bn	CH_2CH_2-2-pyridyl	۶٫۹۶	۷٫۲۴	۷٫۱۹	
۱۵	Bn	Me	۶٫۹۶	۶٫۹۷	۷٫۱۶	
۱۶	Bn	Et	۷٫۱۰	۶٫۸۰	۷٫۰۵	
۱۷	Bn	$CH_2CHCHPh$	۶٫۶۴	۶٫۸۶	۶٫۷۱	
۱۸	Bn	Ph	۶٫۸۲	۷٫۰۵	۷٫۰۱	
۱۹*	Bn	CH_2CH_2Ph	۶٫۷۴	۶٫۸۶	۶٫۸۳	
۲۰	Bn	Boe	۶٫۵۹	۶٫۶۸	۶٫۶۹	
۲۱	Bn	Cbz	۶٫۳۳	۶٫۲۲	۶٫۲۴	
۲۲*	4-ClPh	Me	۷٫۲۱	۷٫۰۱	۷٫۱۶	
۲۳	2-Naphthyl	Me	۷٫۶۰	۷٫۳۹	۷٫۴۰	
۲۴	4-MeOPh	Me	۷٫۳۷	۷٫۲۹	۷٫۲۸	
۲۵	4-MePh (±)	Me	۷٫۶۰	۷٫۵۶	۷٫۴۶	
۲۶	4-MePh	$COCH_2Ph$	۶٫۱۶	۶٫۱۶	۶٫۰۵	
۲۷	4-MeOPh	$COCH_2Ph$	۶٫۰۶	۶٫۰۹	۶٫۰۳	
۲۸	2-Naphthyl	$COCH_2Ph$	۶٫۷۷	۶٫۴۱	۶٫۶۶	
۲۹	4-MePh	$COCH_2CH_2Ph$	۶٫۴۷	۶٫۴۹	۶٫۶۰	
۳۰*	4-MeOPh	$COCH_2CH_2Ph$	۶٫۵۹	۶٫۲۶	۶٫۴۳	
۳۱	2-Naphthyl	$COCH_2CH_2Ph$	۶٫۶۰	۶٫۴۶	۶٫۶۹	

ادامه‌ی جدول ۳

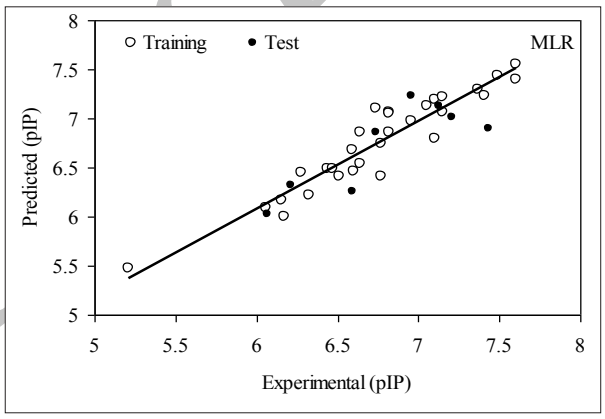
۳۲	4-ClPh	Me	۵٫۲۱	۵٫۴۷	۵٫۳۹	
۳۳*	Bn	Me	۶٫۰۷	۶٫۰۳	۶٫۰۰	
۳۴	2-Naphthyl	Me	۶٫۴۴	۶٫۴۹	۶٫۳۴	
۳۵*	2-Naphthyl	Ph	۶٫۲۱	۶٫۳۲	۶٫۱۶	
۳۶	2-Naphthyl	CH_2CH_2Ph	۶٫۱۷	۶٫۰۰	۶٫۰۸	
۳۷	4-ClPh	Me	۶٫۵۱	۶٫۴۱	۶٫۵۵	
۳۸*	2-Naphthyl	Me	۷٫۴۳	۶٫۹۰	۷٫۲۰	

*: ترکیباتی که دارای ستاره هستند در مجموعه آزمون و ترکیب‌های بدون ستاره در مجموعه آموزش قرار دارند.

PMP: p-methoxyphenyl
Bn: benzyl amide

تجربی نشان می‌دهد. نحوه‌ی پراکندگی خطاها در اطراف محور نشان‌دهنده‌ی این است که پیش‌بینی با استفاده از این روش مناسب ناست. میزان پراکندگی به‌ویژه برای سری آزمون بیشتر است و به احتمال یک رابطه‌ی غیرخطی بین توصیف‌کننده‌ها و فعالیت‌های این ترکیب‌ها موجود است.

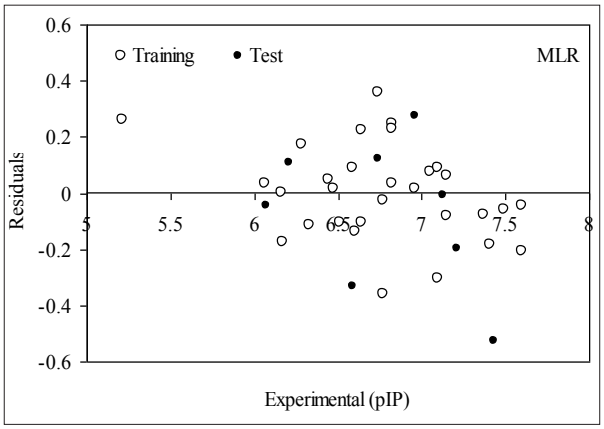
ضریب تعیین (R^2)[1] برای سری آموزش و آزمون به ترتیب برابر ۰٫۸۹۴ و ۰٫۷۱۰ به‌دست آمد. نتیجه‌های به‌دست آمده به ویژه برای سری آزمون نتیجه چندان مطلوبی نیست. بنابراین، دستیابی به نتیجه‌های بهتر از روش غیرخطی شبکه‌ی عصبی مصنوعی استفاده شد.



شکل ۱ مقدارهای pIP محاسبه شده برای مشتقات آمینواسیدها بر اساس مدل MLR در دو مجموعه‌ی آموزشی و آزمون بر حسب مقدارهای تجربی

ایجاد مدل با استفاده از شبکه‌ی عصبی مصنوعی (ANN)

در این روش، توصیف‌کننده‌های انتخاب شده وارد شبکه عصبی مصنوعی می‌شوند. یک شبکه‌ی سه لایه با تابع انتقال سیگموییدی برای نرون‌ها طراحی شد. مقدارهای اولیه وزن‌ها به طور تصادفی از بازه [۱ و ۰] بوده و قبل از عمل آموزش، مقدارهای ورودی و خروجی در فاصله [۰٫۹ و ۰٫۱] نرمال شده است. بهینه‌سازی و به هنگام کردن وزن‌ها و بایاس‌ها به وسیله‌ی الگوریتم پس- انتشار (BP)[2] انجام شد. این الگوریتم یکی از ساده‌ترین اعضای خانواده‌ی الگوریتم‌های آموزشی جهت آموزش شبکه‌های عصبی مصنوعی بوده و برای نخستین بار توسط پاول ورباس (Paul werbos)



شکل ۲ نمودار تغییرهای باقی‌مانده‌ها بر حسب مقدارهای تجربی برای مقدارهای pIP محاسبه شده برای مشتقات آمینواسیدها بر اساس مدل MLR در دو مجموعه‌ی آموزشی و آزمون

1. Statistical package for the social sciences (SPSS)
2. Back propagation (BP)

در سال ۱۹۷۴ ارایه شد [۲۱]. الگوریتم پس- انتشار در اصل بر مبنای روش حداقل‌سازی مربع خطای خروجی شبکه با تنظیم مقدارهای وزن‌ها و بایاس‌ها طرح شده است. سری داده‌ها به طور تصادفی به مجموعه‌های آموزش و آزمون تقسیم شده است. مجموعه‌ی آموزش با استفاده از روش بیرون گذاشتن یک نمونه برای ایجاد مدل و مجموعه پیش‌بینی برای آزمون مدل به کار رفت.

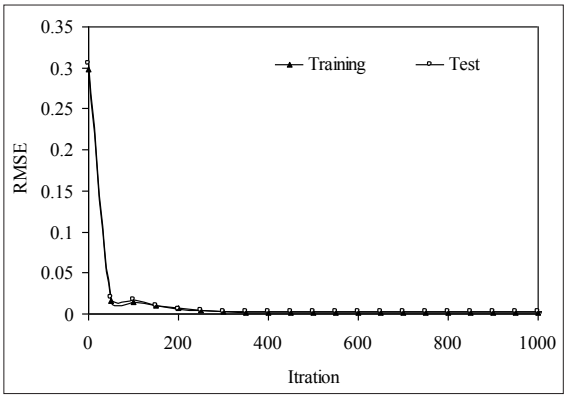
تعداد نرون‌ها در لایه‌ی ورودی با تعداد توصیف‌کننده‌های وارد شده به شبکه‌ی عصبی مصنوعی برابر است. به ازای هر تعداد توصیف‌کننده وارد شده به شبکه، تعداد نرون‌ها در لایه‌ی مخفی بهینه می‌شوند. بدین ترتیب که به ازای هر مدل ANN، تعداد نرون‌ها در لایه‌ی مخفی از ۱ تا ۱۰ تغییر داده شد و مقدارهای ریشه متوسط مربعات خطا (RMSE)[1] برای مجموعه‌های آموزشی و پیش‌بینی محاسبه شد. از رسم مقدارهای RMSE بر حسب تعداد نرون‌ها در لایه‌ی مخفی، تعداد مناسب نرون‌های لایه‌ی مخفی بهینه شده انتخاب می‌شود.

برای ایجاد شبکه‌ی عصبی مصنوعی برای پیش‌بینی فعالیت‌های دارویی ترکیب‌های مورد نظر باید افزون بر تعداد نرون‌ها در لایه‌ی مخفی، وزن‌ها، بایاس‌ها، سرعت یادگیری، مومنتوم و تعداد دورها نیز بهینه شوند که این کار انجام و مقدارهای بهینه شده‌ی این عامل‌ها در جدول ۴ آورده شده است.

جدول ۴ ساختار و مشخصات ANN تولید شده

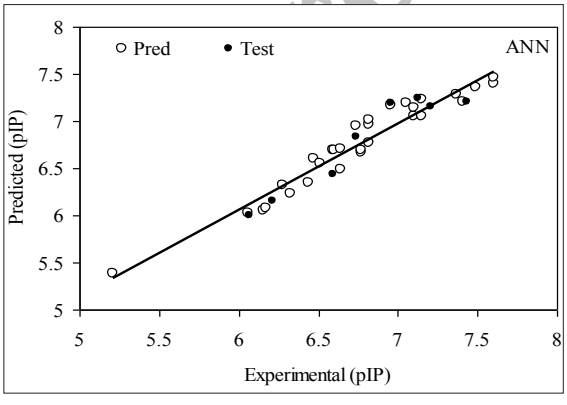
تعداد نرون‌ها در لایه‌ی ورودی	۴
تعداد نرون‌ها در لایه‌ی مخفی	۳
تعداد نرون‌ها در لایه‌ی خروجی	۱
سرعت یادگیری	۰/۶
مومنتوم	۰/۵
تعداد دورها	۱۰۰۰

برای جلوگیری از Overfitting در طول آموزش مقدارهای RMSE بعد از هر ۵۰ سیکل محاسبه و ثبت شد. شکل ۳ نمودار آموزش برای این داده‌ها را نشان می‌دهد که بعد از ۴۰۰ سیکل مقدارهای RMSE برای دو سری آموزشی و آزمون تغییر قابل ملاحظه‌ای نداشته است. بنابراین، آموزش شبکه در این نقطه متوقف می‌شود.



شکل ۳ مقدارهای RMSE برای مجموعه‌های آموزشی و آزمون بر حسب تعداد چرخه‌های آموزش

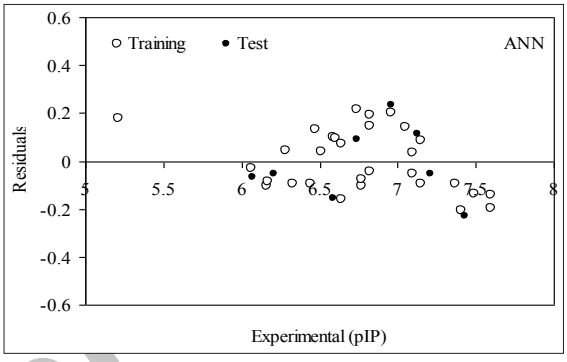
با استفاده از مدل ANN بهینه شده مقدارهای فعالیت‌های بازدارندگی ترکیب‌های مورد نظر در مجموعه‌های آموزشی و پیش‌بینی محاسبه و نتیجه‌های به‌دست آمده در جدول ۳ نشان داده شده است. در شکل ۴، مقدارهای محاسبه شده pIP ترکیب‌های موردنظر در مجموعه‌های متفاوت بر حسب مقدارهای تجربی رسم شده‌اند. ضریب تعیین (R^2) برای سری آموزش و آزمون به ترتیب برابر ۰/۹۴۰ و ۰/۹۱۱ به‌دست آمد.



شکل ۴- مقدارهای pIP محاسبه شده برای مشتقات آمینواسیدها بر اساس مدل ANN در دو مجموعه آموزشی و آزمون بر حسب مقدارهای تجربی

1. Root-mean-square error (RMSE)

به منظور بررسی مقبولیت مدل ارایه شده برای مقدارهای pIP محاسبه شده، نمودار تغییرات مقدار خطا بر حسب مقدارهای تجربی pIP در شکل ۵ رسم شده است. همان‌گونه که مشاهده می‌شود توزیع خطا به‌طور کامل تصادفی بوده و مدل ارایه شده از نظر آماری قابل قبول است.



شکل ۵ نمودار تغییرات باقی‌مانده‌ها بر حسب مقدارهای تجربی برای مقدارهای pIP محاسبه شده برای مشتقات آمینواسیدها بر اساس مدل ANN در دو مجموعه آموزشی و آزمون.

مقایسه‌ی روش‌های MLR و ANN

جدول ۵ عامل‌های آماری متفاوت برای روش‌های MLR و ANN را نشان می‌دهد. مقدار R^2 در واقع نشانگر میزان تطابق مقدارهای پیش‌بینی شده با نتیجه‌های تجربی است. به عبارت دیگر، R^2 معیاری از قدرت پیش‌بینی مدل رگرسیون است.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2} \qquad (۵)$$

در معادله‌ی بالا y_i، $\hat{y}$ و $\bar{y}$ به ترتیب نشان‌دهنده مقدارهای تجربی، مقدارهای پیش‌بینی شده و میانگین مقدارهای تجربی هستند. مقدار R^2 برای سری آموزش و پیش‌بینی در روش MLR به ترتیب برابر ۰٫۸۹۴ و ۰٫۷۱۰ و در روش ANN برابر ۰٫۹۴۰ و ۰٫۹۱۱ به‌دست آمد. یکی دیگر از راه‌های ارزیابی قدرت پیش‌بینی مدل، عامل RMSE است که از رابطه‌ی زیر محاسبه می‌شود:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \qquad (۶)$$

عامل RMSE مقدار پراکندگی مقدارهای تجربی حول خط رگرسیون یا به عبارت دیگر میزان خطا را مشخص می‌کند. پایین بودن مقدار RMSE نشان‌دهنده‌ی قدرت پیش‌بینی و اعتبار مدل است. نتیجه‌های به‌دست آمده نشان می‌دهد که روش ANN نسبت به روش MLR قدرت بیشتری جهت پیش‌بینی فعالیت دارویی مشتقات آمینواسیدها دارد.

جدول ۵ عامل‌های آماری به‌دست آمده برای مدل‌های MLR و ANN

	سری آموزش			سری آزمون		
	R^2	RMSE	F	R^2	RMSE	F
MLR	۰٫۸۹۴	۰٫۱۶۶	۵۲٫۷۱۵	۰٫۷۱۰	۰٫۲۵۸	۲٫۰۷۶
ANN	۰٫۹۴۰	۰٫۱۲۵	۵۶٫۵۳۴	۰٫۹۱۱	۰٫۰۶۶	۱٫۸۵۴

آماره‌ی F برای ارزیابی دقت مدل‌های پیشنهادی به وسیله‌ی رابطه‌ی زیر قابل تعیین است:

$$F = \frac{MSR}{MSE} \qquad (۷)$$

در این رابطه، MSR و MSE به ترتیب دلالت بر میانگین مربع‌های رگرسیون و میانگین مربع‌های خطا دارند. این دو عامل به با رابطه‌های زیر قابل تعیین هستند.

$$MSR = \frac{SSR}{m} \qquad (۸)$$

$$MSE = \frac{SSE}{n-m-1} \qquad (۹)$$

در معادله‌های ۸ و ۹، SSR و SSE نشانگر مجموع مربع‌های رگرسیون و مجموع مربع‌های خطا هستند. m و n هم به ترتیب بیانگر تعداد توصیف‌کننده‌ها (۴) و تعداد ترکیب‌ها در سری آموزش (۳۱) می‌باشند.

نتیجه‌گیری

مطالعه‌های QSAR ابزاری قوی و توانمند برای پژوهش و بررسی ساختار با فعالیت ترکیب‌های شیمیایی است که به طور وسیع در شیمی‌دارویی به‌منظور بررسی و پیش بینی فعالیت بازدارندگی داروها و طراحی داروهای جدید به‌کار می‌رود

در این پژوهش تعداد ۳۸ مولکول از مشتقات آمینواسیدها با استفاده از روش‌های خطی و غیرخطی مورد ارزیابی قرار گرفت. تعداد ۱۴۹۷ توصیف‌کننده برای هر مولکول محاسبه و از بین آن‌ها ۴ توصیف‌کننده‌ای که بیشترین ارتباط را با فعالیت ترکیب‌ها داشتند انتخاب شدند. این ۴ توصیف‌کننده عبارتند از: +R4m، E1m، RDF070m، RDF020m. سپس با روش‌های MLR و ANN مدل‌سازی و پیش‌بینی انجام شد. نتیجه‌های به‌دست آمده نشان از برتری روش ANN نسبت به روش MLR دارد. توصیف‌کننده‌های بکار رفته به نوعی به جرم اتم‌های تشکیل دهنده و یا به عبارتی به جرم مولکولی ترکیب‌های مورد مطالعه بستگی دارد و از آن‌جایی‌که در معادله‌ی به‌دست آمده مقدارهای ضریب همه این توصیف‌کننده‌ها علامت منفی دارد نشان‌دهنده‌ی آن است که مقدارهای pIP این ترکیب‌ها با مقدار جرم مولکولی آن‌ها رابطه عکس دارد و با زیاد شدن جرم مولکولی مقدار pIP این ترکیب‌ها کم می‌شود. پس می‌توان نتیجه‌گیری کرد که در ترکیب‌هایی که هنوز سنتز نشده‌اند داشتن جرم مولکولی بالا در نهایت باعث تأثیر عکس در مقدار pIP مربوط شده و این مقدار در مولکول‌های سبک بیشتر خواهد بود.

مراجع

[1] Melagraki, G.; Afantitis, A.; Makridima, K.; Sarimveis, H.; Igglessi-Markopoulou, O.; J. Mol. Model., 12, 297-305, 2006.

[2] Habibi-Yangjeh, A.; Pourbasheer, E.; Danandeh-Jenagharad, M.; B. Korean. Chem. Soc., 29, 833-841, 2008.

[3] Habibi-Yangjeh, A.; Pourbasheer, E.; Danandeh-Jenagharad, M.; Monatsh. Chem.; 140, 15-27, 2009.

[4] Beheshti, A.; Pourbasheer, E.; Nekoei, M.; Banaei, A.R.; Anal.Chem. Let., 2, 33-43, 2012.

[5] Adimi, M.; Salimi, M.; Nekoei, M.; Pourbasheer, E.; Beheshti, A.; J. Serb. Chem. Soc., 77, 639-650, 2012.

[6] Dolatabadi, M.; Nekoei, M.; Banaei, A.R.; Monatsh Chem., 141, 577-588, 2010.

[7] Nekoei, M.; Salimi, M.; Dolatabadi, M.; Mohammadhosseini, M.; Monatsh Chem., 142, 943-948, 2011.

[8] Nekoei, M.; Salimi, M.; Dolatabadi, M.; Mohammadhosseini, M.; J. Serb. Chem. Soc., 76, 1117-1127, 2011.

[9] Todeschini, R.; Consonni, V.; Mannhold, R.; Kubinyi, H.; Timmerman, H(Eds.), Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.

[10] Menhaj, M.; Foundations of Neural Networks, Amir Kabir University Press, Iran, 2000.

[11] Beals, R.; Jackson, T.; Neural computing: An introduction, IOP Publishing Ltd., NewYork, 1998.

[12] Pham, D.T.; Liu, X.; Neural networks for identification. predication and control, Springer-Velag, London, 1999.

[13] Meiler, J.; Meusinger, R.; Will, M.; J. Chem. Inf. Comput. Sci., 40, 1169-1176, 2000.

[14] Habibi-Yangjeh, A.; Nooshyar, M.; Phys. Chem. Liq., 43, 239-247, 2005.

[15] Nekoei, M.; Mohammadhosseini, M.; J. Chin. Chem. Soc., 54, 383-390, 2007.

[16] Zarei, K.; Atabati, M.; Nekoei, M.; Annali. Di. Chimica., 97, 723-731, 2007.

[17] Nekoei, M.; Mohammadhosseini, M.; Zarei, K.; J. Chin.Chem. Soc., 55, 362-368, 2008.

[18] Hubbs, J.L.; Zhou, H.; Kral, A.M.; Fleming, J.C.; Dahlberg, W.K.; Hughes, B.L.; Middleton, R.E.; Szewczak, A.A.; Secrist, J.P.; Miller, T.A.; Bio. & Med. Chem. Let., 18, 34-38, 2008.

[19] HyperChem Release 7, HyperCube, Inc.: http://www.hyper.com.

[20] Todeschini, R.; Milano Chemometrics and QSPR Group; http://www.disat.unimib.it/vhm.

[۲۱] منهاج، م.ب.؛ "مبانی شبکه‌های عصبی"، انتشارات دانشگاه امیرکبیر، ۱۳۷۹.