مجله شیمی کوانتومی و اسپکتروسکوپی
۱۲:٤٠ (١٤٠١) ٧٤-٩١
Journal of Quantum Chemistry and Spectroscopy (JQCS)

کاربرد رگرسیون خطی چندگانه و شبکه های عصبی مصنوعی جهت مطالعه ارتباط کمی ساختار-فعالیت دسته ایی از مشتقات کموکین ها

محمدرضا کیانسب[۱]، مهدی نکوئی[۱*]، مجید محمدحسینی[۱]، بهنام مهدوی[۲]، تهمینه باهری[۳]

[۱]گروه شیمی، واحد شاهرود، دانشگاه آزاد اسلامی، شاهرود، ایران

[۲]گروه شیمی، دانشکده علوم، دانشگاه حکیم سبزواری، سبزوار، ایران

[۳]گروه مبارزه با مواد مخدر، دانشگاه علوم انتظامی امین، تهران، ایران

تاریخ ثبت اولیه:۱۴۰۱/۰۸/۰۳، تاریخ دریافت نسخه اصلاح شده:۱۴۰۱/۱۱/۲۲ ، تاریخ پذیرش قطعی:۱۴۰۱/۱۲/۱۳

چکیده

مطالعه ارتباط کمی ساختار-فعالیت (QSAR) جهت پیش بینی فعالیت دارویی برخی از مشتقات کموکین ها با استفاده از رگرسیون خطی چندگانه و شبکه های عصبی مصنوعی (ANN) انجام شد. در ابتدا ساختار ترکیبات دارویی به کمک نرم افزار هایپرکم رسم و بهینه گردیدند. سپس دسته وسیعی از توصیف کننده‌های مولکولی توسط نرم افزار دراگون محاسبه شدند. بعد از کاهش تعدادی از توصیف کننده ها که همبستگی بالای۰/۹ داشتند و توصیف کننده هایی که بیش از ۹۰٪ آنها مشابه بود از رگرسیون مرحله ایی برای بدست آوردن بهترین توصیف کننده‌ها که بیشترین ارتباط را با فعالیت دارویی ترکیبات مورد نظر داشتند استفاده گردید. با این کار تعداد ۷ توصیف کننده شامل MATS2p، PCWTe، RDF045m، RDF065m، RDF115m، C-003 و C-040 انتخاب شدند سپس از روشهای رگرسیون خطی چندگانه (MLR) و شبکه های عصبی مصنوعی (ANN) برای مدلسازی و پیش بینی فعالیت ترکیبات سری تست استفاده گردید. نتایج بدست آمده نشان می دهد که هر دو روش نتایج قابل قبولی ارایه می دهند که می توان از آنها برای پیش بینی ترکیبات دارویی جدید استفاده کرد.

واژه های کلیدی: ارتباط کمی ساختار- فعالیت، ، مشتقات کموکین ها، رگرسیون خطی چند گانه، شبکه های عصبی مصنوعی.

۱. مقدمه

کموکین‌ها خانواده‌ای از سیتوکین‌های با اندازه کوچک یا پروتئینهای تولید شده توسط سلول‌ها هستند. نام این خانواده از خاصیت آن‌ها برای جذب سلول‌های پاسخ دهنده الهام گرفته شده‌است. پروتئین‌هایی که به عنوان کموکین طبقه بندی می‌شوند دارای ویژگی‌های ساختاری مشابهی همچون اندازه کوچک (۸-۱۰ کیلودالتون) و حضور چهار عدد مولکول سیستئین در

***عهده دار مکاتبات:** مهدی نکوئی

نشانی: گروه شیمی، واحد شاهرود، دانشگاه آزاد اسلامی، شاهرود، ایران

تلفن:۰۲۳۳۲۳۹۴۲۸۹ **پست الکترونیک:**E-mail:m_nekoei1356@yahoo.com

جایگاه‌های کلیدی ساختاری هستند. برخی از کموکین‌ها دارای نقش پیش-التهاب هستند و در هنگام التهاب، سلول‌های ایمنی را به مکان عفونت فرا می‌خوانند. دیگر کموکین‌ها در نگهداری همئوستاز نقش دارند و باعث حرکات سلول‌ها در طی فرایند طبیعی رشد و گسترش سلول می‌شوند. کموکین‌ها در همه مهره‌داران و برخی باکتری‌ها یافت می‌شوند.کموکین‌ها در چهار گروه CXC , CC CX3Cو XC تقسیم‌بندی می‌شوند و از طریق یک پروتئین غشایی بنام جی پروتئین اثر می‌کنند [۵-۱].

جی‌پروتئین پروتئینی است که می‌تواند گوانوزین تری فسفات (GTP) را به گوانوزین دی‌فسفات (GDP) تبدیل کند. جی‌پروتئین‌ها گیرنده‌هایی پروتئینی هستند که یک سمت آن‌ها خارج از سلول و سمت دیگرشان داخل سلول قرار دارد (عبورکننده از عرض غشا). این سیستم نوعی پیامرسان برای مواد شیمیایی خاصی مثل هورمون‌ها و پیامرسان عصبی در سطح غشای سلول است.

گیرنده های کموکین C-C نوع ۲ (CCR2) متعلق به خانواده گیرنده های جفت شده با پروتئین روی مونوسیت ها، ماکروفاژها، بازوفیل ها، ماست سل ها و لنفوسیت های T بیان می شوند. دو شکل جایگزین از گیرنده‌های CCR2 به نام‌های CCR2a و CCR2b وجود دارد که فقط در انتهای کربوکسیل‌شان متفاوت هستند. این گیرنده ها نقش مهمی در استخدام مونوسیت ها/ماکروفاژها، سلول های T ایفا می کنند و به طور مستقیم با بسیاری از بیماری ها مانند التهاب، HIV و فیبروز ریوی مرتبط هستند. گیرنده‌های CCR2 با برهمکنش با گیرنده‌های کموکین واقع در سطح سلولی لکوسیت‌ها و به دنبال آن کموتاکسی و نفوذ به بافت مجاور، در تنوعی از پاسخ‌های التهابی نقش دارند[۱۰-۶].

استفاده از روش آزمون و خطا برای ارزیابی فعالیت و خواص ترکیبات شیمیایی و داروها فرآیندی زمان‌بر و پرهزینه است. روش‌های ارتباط کمی ساختار-فعالیت (QSAR) روش‌های قدرتمندی برای غلبه بر این محدودیت هستند. این تکنیک‌ها معمولاً بر اساس مدل‌های خطی یا غیرخطی تعیین‌شده آماری هستند که رفتار شیمیایی ترکیبات و توصیفگرهای آنها را به هم مرتبط می‌کنند. علاقه اصلی در توسعه مدل‌های پیش‌بینی‌کننده QSAR به دلیل توانایی بالای آن‌ها در پیش‌بینی فعالیت‌ها و/یا خواص ترکیبات است، به‌ویژه برای آن‌هایی که به دلایل زیادی از نظر تجربی اندازه‌گیری آنها پرهزینه و زمان‌برهستند. هدف از مطالعات QSAR پیداکردن رابطه ای است که بین رفتار فیزیکوشیمیایی یک مولکول با پارامترهای ساختاری آن وجود دارد. نتایج این مطالعات علاوه بر شفاف سازی نحوه ارتباط بین خواص مولکول ها و ویژگی های ساختمانی آنها به پژوهشگران در پیش بینی رفتار مولکول های جدید براساس رفتار مولکول های مشابه کمک می کند. در واقع، QSAR روابط ریاضی بین فعالیت های شیمیایی، فیزیکی، بیولوژیکی و... با توصیفگرهای فیزیکوشیمیایی، استریوشیمیایی و توپولوژیکی ایجاد می کنند. روشهای مختلفی از جمله رگرسیون خطی چندگانه(MLR)، کمترین مربعات جزیی(PLS)، شبکه های عصبی مصنوعی(ANN) و ماشین بردار پشتیبان(SVM) در مدلسازی های QSAR مورد استفاده قرار گرفته است[۱۹-۱۱].

هدف از این تحقیق، استفاده از روشهای رگرسیون خطی چندگانه (MLR) و شبکه های عصبی مصنوعی (ANN) برای مدل سازی و پیش بینی فعالیت‌های بازداری (IC_{50}) گیرنده‌های CCR2b می باشد.

۲. روش‌های محاسباتی

۲-۱. انتخاب سری داده‌ها

در این کار تعداد ۱۰۳ ترکیب از مشتقات کموکین ها (CCR2b) توسط روشهای کمومتریکس مورد بررسی قرار گرفت [۲۰-۲۲]. در این مقاله قدرت بازدارندگی این ترکیبات به صورت IC_{50} گزارش شده است. IC_{50} عبارتست از مینیمم غلظتی از ترکیب دارویی که باعث ۵۰٪ اثر بازدارندگی می‌شود. این مقادیر به مقیاس لگاریتمی تبدیل شده $-\log(IC_{50})=pIC_{50}$ و مورد استفاده قرار گرفته است. در مدلسازی به روش MLR ترکیبات به صورت تصادفی به دو سری آموزش و پیش‌بینی تقسیم شدند، سری آموزش شامل ۸۳ مولکول و سری پیش‌بینی شامل ۲۰ مولکول می‌باشد. جهت مدلسازی به روش ANN برای جلوگیری از برازش اضافی، ترکیبات به سه سری آموزش شامل ۶۲ مولکول، ارزیابی شامل ۲۱ مولکول و سری پیش بینی شامل ۲۰ مولکول تقسیم شدند. لازم به ذکر است جهت مقایسه، سری پیش بینی (تست) در هر دو روش دارای ترکیبات یکسان می باشد. مقادیر pIC_{50} به عنوان متغیر وابسته و توصیف کننده‌ها به عنوان متغیر مستقل انتخاب شدند. سری آموزش جهت آموزش و ایجاد یک مدل مناسب و سری پیش‌بینی جهت ارزیابی مدل مورد استفاده قرار گرفت (در ANN سری ارزیابی برای بررسی و جلوگیری از برازش اضافی استفاده گردید).

۲-۲. رسم و بهینه سازی ساختار مولکول‌ها

در این مرحله از مطالعه، ساختار مولکولی هر ترکیب ابتدا در نرم افزار HyperChem07 ترسیم شد. سپس با احتساب اتم های هیدروژن، ساختار سه بعدی ترکیبات با استفاده از روش های نیمه تجربی کوانتمیAM1 بهینه گردید و این بهینه سازی تا زمانی ادامه یافت که جذر میانگین مربعات گرادیان انرژی به ۰/۰۰۱ کیلوکالری بر مول رسید. با استفاده از این نرم افزار می توان اطلاعات فراوانی نظیر زوایای پیوندی، طول پیوندها، زوایای پیچش، بار اتم ها، انرژی تشکیل مولکول و... را بدست آورد. برخی دیگر از قابلیت های این نرم افزار عبارتند از: توانایی نمایش ساختار مولکولی با قابلیت کنترل آن (از جمله انتخاب، چرخش، تبدیل و تغییر اندازه ساختار مولکولی)، دارای ابزارها و متدهای محاسباتی مختلف (از جمله تعیین تراز انرژی) و امکان تعریف نوع اتم، جرم اتمی و سایر ویژگی ها و ....

۲-۳. محاسبه توصیف کننده های مولکولی

توصیف کننده ها مقادیر عددی هستند که ویژگیهای مختلف مولکول را نشان می دهند. توصیف‌کننده‌های مولکولی نتیجه نهایی یک استدلال و روش ریاضی است که اطلاعات شیمیایی را به رمز تبدیل می‌کند و آنها را به صورت یک نماد نشان می‌دهد که ارائه دهنده یک ویژگی مولکول به صورت یک عدد مفید می‌باشد. هر یک از این توصیف کننده‌ها، اطلاعات خاصی از مولکول را در اختیار می‌گذارد. توصیف‌کننده‌های مولکولی مختلفی برای اهداف گوناگون به کار برده شده‌اند. اختلاف این توصیف‌کننده‌ها در پیچیدگی اطلاعات رمزگزاری شده و زمان مورد نیاز برای محاسبه می‌باشد. اولین نوع توصیف کننده ها، توصیف کننده های

توپولوژی می باشند. این توصیف‌کننده‌ها از روی گراف‌های مولکولی بدست می‌آیند و جزء ساده ترین نوع توصیف‌کننده‌ها می‌باشند و به ساختار فضایی مولکول ارتباطی نداشته و تنها به نوع اتم، نوع پیوندها و نحوه ارتباط اتم‌ها به یکدیگر وابسته است. از جمله این توصیف کننده ها می توان به تعداد اتم‌ها، شاخص‌های ارتباطی مولکولی و وزن مولکولی و ... اشاره کرد. دومین نوع توصیف کننده ها، توصیف کننده های هندسی است. این توصیف‌کننده‌ها با ساختار سه بعدی مولکول‌ها در ارتباط می‌باشند. برای محاسبه این توصیف‌کننده‌ها ابتدا می‌بایست ساختار فضایی مولکول‌ها بهینه شود. برخی از این توصیف کننده ها عبارتند از: حجم مولکولی، مساحت سطح و مساحت سطح در دسترس حلال. توصیف‌کننده‌های شیمی کوانتومی از جمله توصیف کننده های دیگری هستند که با استفاده از بهینه‌سازی نیمه تجربی ساختار مولکول‌ها در نرم‌افزارهای مختلف بدست می‌آیند. از جمله این توصیف کننده ها می توان به انرژی بالاترین تراز اشغال شد، انرژی پایین ترین تراز اشغال نشده، بار و الکترونگاتیویته اتم‌ها و ... اشاره کرد. توصیف‌کننده‌های دیگر، توصیف کننده های فیزیکوشیمیایی هستند که این توصیف‌کننده‌ها بیانگر بعضی از خواص فیزیکوشیمیایی مولکول‌ها می‌باشند که به ساختار مولکول وابستگی شدیدی نشان می‌دهند. از قبیل: ضریب تقسیم آب-اکتانول، ویسکوزیته، میزان حلالیت ترکیبات در آب، شکست مولکولی، نقطه ذوب و نقطه جوش. توصیف‌کننده‌های ارتباطی مولکولی نیز از جمله توصیف کننده های مهم دیگری هستند که اطلاعاتی از جمله اندازه و ساختار مولکول، مرتبه شاخه‌دار شدن و نحوه ارتباط اتم‌ها در مولکول را بیان می‌کنند[۲۳].

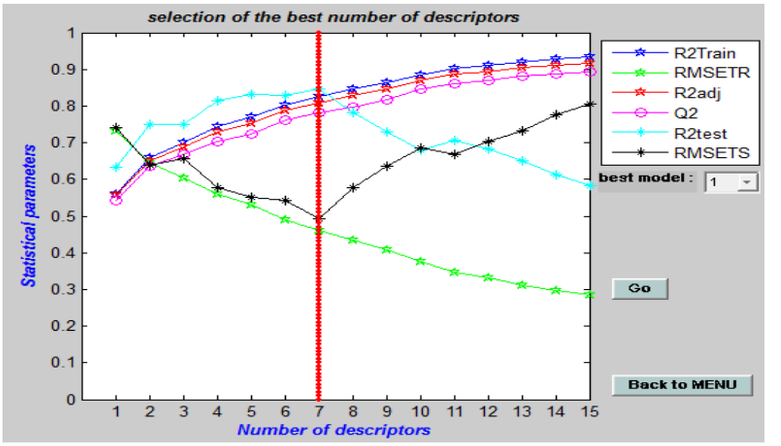
برای محاسبه توصیف کننده ها، بعد از رسم ساختارهای مولکولی به کمک نرم افزار Hyper Chem و بهینه سازی ساختار آنها، این ساختارها به نرم افزارDragon وارد شده و توصیف کننده های مولکولی به تعداد ۱۴۸۱ مورد به وسیله این نرم افزار محاسبه می شوند.

۲-٤. انتخاب مناسب‌ترین توصیف‌کننده‌ها

جهت انتخاب مناسب ترین توصیف کننده ها از روش رگرسیون مرحله ایی[1] استفاده شد. در روش رگرسیون مرحله ای، متغیرها یکی پس از دیگری وارد مدل شدند در این حالت، ابتدا متغیری وارد مدل می‌شود که بالاترین میزان همبستگی را با متغیر وابسته دارد. با ورود هر متغیر جدید، کلیه متغیرهای موجود در معادله بررسی شده و اگر هر کدام از آنها سطح معناداری خود را از دست بدهند، قبل از ورود متغیر جدید از مدل خارج می‌شود. به این ترتیب داده‌های pIC_{50} به عنوان متغیر وابسته و توصیفگرها به عنوان متغیر مستقل در نظر گرفته شده و تکنیک رگرسیون مرحله‌ای انجام شد. همانطور که می‌دانیم روش رگرسیون مرحله ایی تعداد زیادی مدل ارائه می‌کند. که مدل اول شامل یک توصیفگر، مدل دوم شامل دو توصیفگر و ... می‌باشد. با افزایش تعداد توصیفگرها بالطبع مقدار R^2 افزایش و RMSE[2](خطای جذر میانگین مربعات)کاهش می‌یابد. اما بدلیل پیچیدگی مدل، نمی‌توانیم تعداد زیادی

[1] Stepwise
[2]Root-mean-square-error

توصیف کننده را جهت مدلسازی انتخاب کنیم. بدین منظور و جهت انتخاب تعداد توصیف کننده‌های مناسب، نمودار پارامترهای مختلف آماری از جمله $RMSE_{train}$, $RMSE_{test}$،R^2_{train}, R^2_{test} برحسب تعداد توصیف کننده‌ها رسم گردید که در شکل ۱ نشان داده شده است. بر طبق این شکل، تعداد ۷ توصیف کننده به عنوان توصیف کننده‌هایی که بیشترین ارتباط را با فعالیت دارویی(pIC_{50}) دارند، انتخاب شدند. این ۷ توصیف کننده به همراه مفهوم و نوع آنها در جدول ۱ ارائه شده است.



شکل ۱. نمودار پارامتر های آماری از جمله (R^2train, R^2test, RMSEtrain, RMSEtest) برحسب تعداد توصیف کننده‌ها

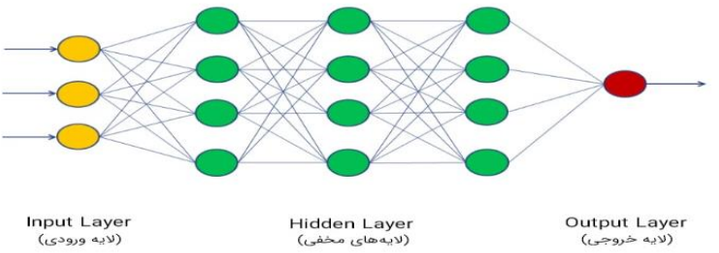
جدول ۱. توصیفگرهای انتخاب شده توسط رگرسیون خطی چندگانه مرحله به مرحله

نشانه توصیف کننده	مفهوم توصیف کننده	نوع توصیف کننده
MATS2p	Moran autocorrelation of lag 2 weighted by polarizability	2D autocorrelations
PCWTe	partial charge weighted topological electronic	Charge descriptors
RDF045m	Radial Distribution Function – 045 / weighted by mass	RDF descriptors
RDF065m	Radial Distribution Function – 065 / weighted by mass	RDF descriptors
RDF115m	Radial Distribution Function – 115 / weighted by mass	RDF descriptors
C-003	CHR3	Atom-centred fragments
C-040	R-C(=X)-X / R-C#X / X=C=X	Atom-centred fragments

۲-۵. شبکه های عصبی مصنوعی

شبکه‌های عصبی مصنوعی، سیستم‌ها و روش‌های محاسباتی نوین برای یادگیری، نمایش دانش و در انتها اعمال دانش به دست آمده در جهت بیش‌بینی پاسخ‌های خروجی از سامانه‌های پیچیده هستند. ایده اصلی این گونه شبکه‌ها تا حدودی الهام‌گرفته از شیوه کارکرد سیستم عصبی زیستی برای پردازش داده‌ها و اطلاعات به منظور یادگیری و ایجاد دانش میباشد. شبکه عصبی مصنوعی

روشی است که دانش ارتباط بین چند مجموعه داده را از طریق آموزش فراگرفته و برای استفاده در موارد مشابه ذخیره می‌کند. یک شبکه عصبی مصنوعی، از سه لایه ورودی، خروجی و پنهان تشکیل می‌شود. هر لایه شامل گروهی از سلول‌های عصبی (نورون) است که عموماً با کلیه نورون‌های لایه‌های دیگر در ارتباط هستند، مگر این که کاربر ارتباط بین نورون‌ها را محدود کند؛ ولی نورون‌های هر لایه با سایر نورون‌های همان لایه، ارتباطی ندارند. با استفاده از دانش برنامه‌نویسی رایانه می‌توان ساختار داده‌ای طراحی کرد که همانند یک نورون عمل نماید. سپس با ایجاد شبکه‌ای از این نورون‌های مصنوعی به هم پیوسته، ایجاد یک الگوریتم آموزشی برای شبکه و اعمال این الگوریتم به شبکه آن را آموزش داد. نورون کوچک‌ترین واحد پردازشگر اطلاعات است که اساس عملکرد شبکه‌های عصبی را تشکیل می‌دهد. یک شبکه عصبی مجموعه‌ای از نورون‌هاست که با قرار گرفتن در لایه‌های مختلف، معماری خاصی را بر مبنای ارتباطات بین نورون‌ها در لایه‌های مختلف تشکیل می‌دهند. نورون می‌تواند یک تابع ریاضی غیرخطی باشد، در نتیجه یک شبکه عصبی که از اجتماع این نورون‌ها تشکیل می‌شود، نیز می‌تواند یک سامانه کاملاً پیچیده و غیرخطی باشد. در شبکه عصبی هر نورون به‌طور مستقل عمل می‌کند و رفتار کلی شبکه، برآیند رفتار نورون‌های متعدد است. به عبارت دیگر، نورون‌ها در یک روند همکاری، یکدیگر را تصحیح می‌کنند[۲۰-۱۷]. شکل ۲ نمایی از یک شبکه عصبی مصنوعی را نشان می دهد



شکل ۲. نمایی از یک شبکه عصبی مصنوعی

۳. نتایج و بحث

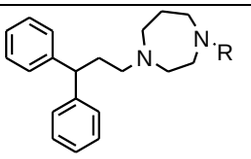
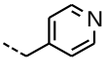
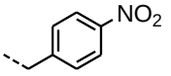
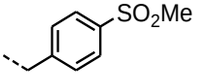
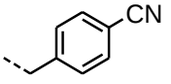
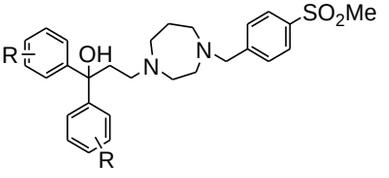
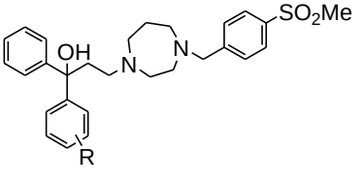
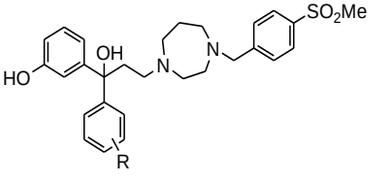
۳-۱. مدل‌سازی به روش رگرسیون خطی چندگانه(MLR)

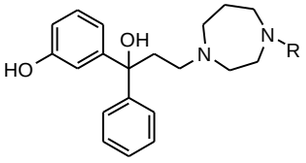
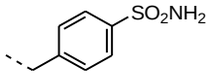
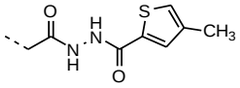
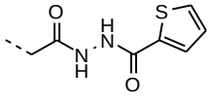
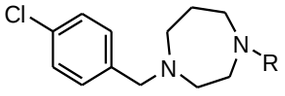
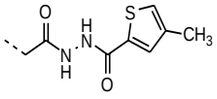
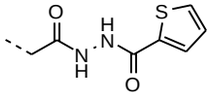
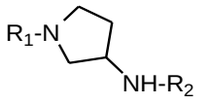
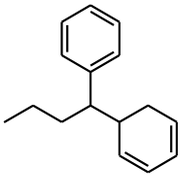
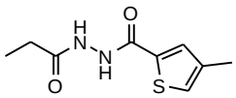
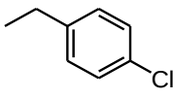
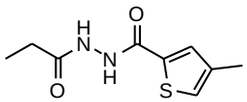
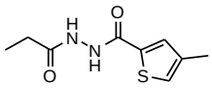
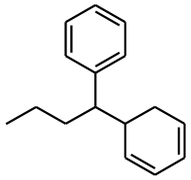
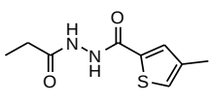
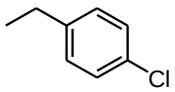
پس از انتخاب مناسب‌ترین توصیفگرها توسط روش مرحله‌ای، مرحله بعدی، ایجاد مدل میان توصیفگرهای انتخاب شده و pIC_{50} می‌باشد. بین توصیف کننده‌ها و فعالیت دارویی ترکیبات برای سری آموزش با استفاده از روش MLR رابطه زیر به عنوان مدل خطی بدست آمد

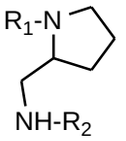
$$\mathbf{pIC_{50}} = 0.297 - 6.608\ (\mathbf{MATS2p}) + 0.183\ (\mathbf{PCWTe}) + 0.079\ (\mathbf{RDF045m}) + 0.046\ (\mathbf{RDF065m}) - 0.145\ (\mathbf{RDF115m}) - 0.886\ (\mathbf{C\text{-}003}) - 0.407\ (\mathbf{C\text{-}040})$$

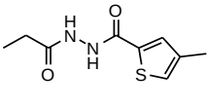
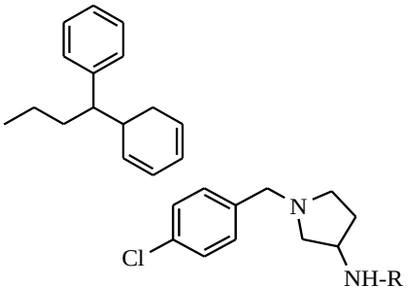
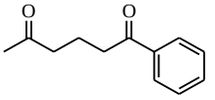
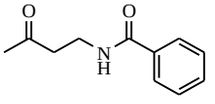
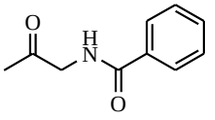
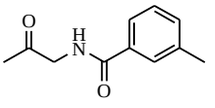
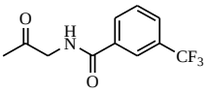
سپس از معادله بدست آمده برای پیش‌بینی فعالیت دارویی ترکیبات سری پیش بینی (تست) استفاده گردید. مقادیر تجربی و پیش‌بینی شده pIC_{50} برای کلیه ترکیبات مجموعه آموزش و تست در جدول (۲)آورده شده است. شکل(۳) نیز نمودار مقادیر پیش بینی شده بر حسب مقادیر تجربی را نشان می دهد.

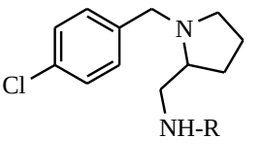
جدول۲. مقادیر تجربی و محاسبه شده pIC_{50} ترکیبات مختلف برای مجموعه‌های آموزشی و پیش‌بینی در مدل‌های SW-MLR, SW-ANN

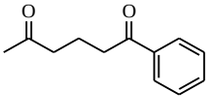
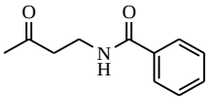
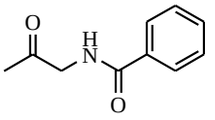
No.	R_1	R_2	Exp.	SW-MLR	SW-ANN
					
1		-	4.96	4.58	4.90
2		-	4.72	4.80	4.67
3		-	4.89	4.51	4.86
4[p]		-	4.37	4.64	4.63
					
5	H	-	5.39	5.16	5.65
6	4-NMe_2	-	4.35	4.58	4.31
7[p]	4-OH	-	4.68	5.62	5.17
8	3-OH	-	5.82	5.63	5.76
9	4-F	-	5.15	5.57	5.90
10	3-F	-	4.47	5.49	4.70
11	4-Cl	-	4.96	5.65	4.85
					
12	3-OH	-	6.15	5.46	6.33
13	3-CH_2OH	-	5.40	5.34	5.59
14[p]	3-NH_2	-	5.38	5.57	5.91
15	3-NHMe	-	5.30	5.46	4.98
16[p]	3-OMe	-	5.10	5.35	5.17
17	3-F	-	4.92	5.29	5.03
18[p]	3-Me	-	4.80	5.37	5.42
					
19	3-F	-	5.62	5.53	5.71
20	3-Cl	-	5.35	5.58	4.31
21	4-F	-	5.82	5.49	5.56

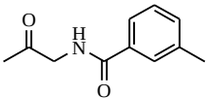
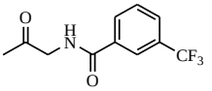
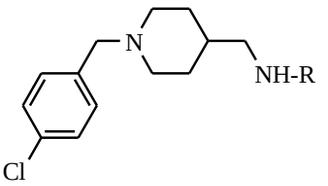
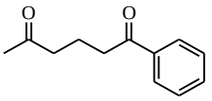
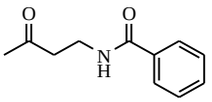
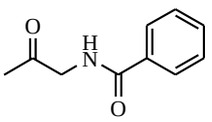
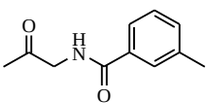
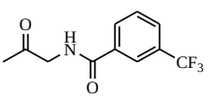
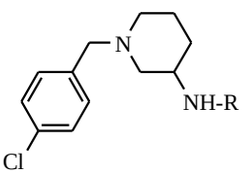
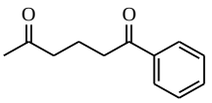
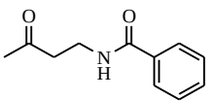
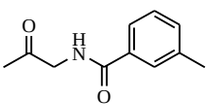
22	4-Cl	-	5.82	5.67	5.61
23	3,5-DiF	-	5.03	5.64	5.12
					
24		-	5.82	5.45	5.02
25[p]		-	5.19	4.05	4.69
26		-	4.55	3.99	4.67
					
27		-	5.13	4.94	4.85
28		-	4.52	4.09	4.85
					
29[p]				4.48 4.45	4.05
30				4.39 4.70	4.60
31			4.77	5.17	4.49
32			5.11	5.05	4.88

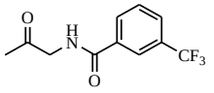
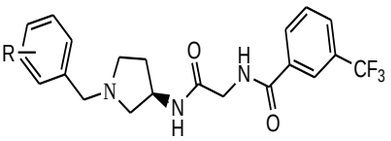
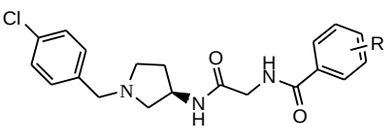


33			4.59	4.78	5.26
34[p]		-	4.96	5.11	5.09
35		-	4.66	4.85	4.88
36		-	4.38	4.87	4.43
37		-	5.16	5.30	4.94
38[p]		-	6.43	5.86	6.34

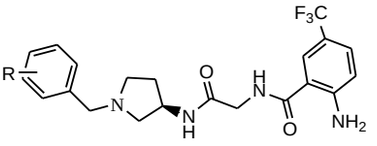


39		-	4.72	4.20	4.31
40		-	4.40	4.16	5.01
41		-	4.18	4.31	4.63

42		-	4.85	4.73	5.32
43		-	5.64	5.63	5.69
					
44[p]		-	4.30	4.73	4.29
45		-	4.19	4.76	4.53
46		-	4.96	4.85	4.50
47[p]		-	5.26	5.32	5.33
48		-	6.15	6.07	6.33
					
49		-	4.06	4.03	4.43
50		-	4.13	4.32	4.50
51		-	4.77	4.78	3.84

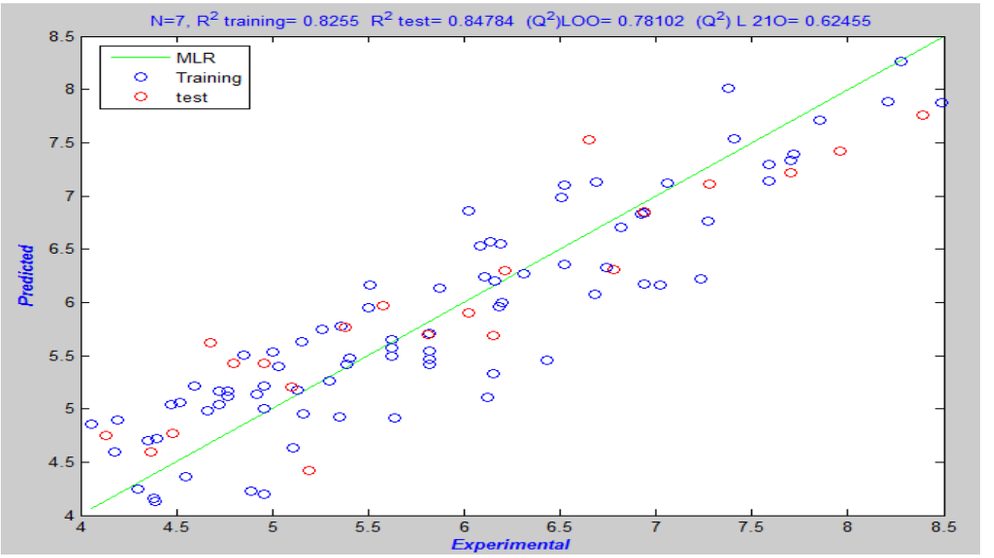
52		-	6.18	5.60	5.85
					
53[p]	H	-	6.16	6.01	6.26
54	2-Cl	-	6.20	6.22	6.00
55	2-CH_3	-	6.02	6.46	5.92
56	2-OCH_3	-	5.87	6.48	6.21
57	3-CH_3	-	5.51	6.33	6.24
58	3-OCH_3	-	5.50	6.35	6.01
59	4-Cl	-	6.74	6.13	6.27
60	4-CH_3	-	6.94	6.71	6.68
61[p]	4-OCH_3	-	6.94	6.69	6.81
62	4-Et	-	7.23	6.98	6.11
63[p]	4-Br	-	6.78	6.56	6.18
64	4-Vinyl	-	6.92	6.59	6.82
65	4-CH_3S	-	6.69	6.65	6.89
66[p]	4-OH	-	6.65	6.59	7.23
67	4-NHAc	-	6.52	6.96	7.32
68[p]	4-OCF_3	-	6.21	7.07	5.82
69	4-F	-	6.02	6.77	6.83
70[p]	4-NO_2	-	5.81	6.97	6.00
71	4-CN	-	5.58	6.29	6.01
72	2,4-$(CH_3)_2$	-	7.27	6.71	7.02
73	2,4-Cl_2	-	6.52	6.41	6.47
74	4-OH, 3-OCH_3	-	6.82	6.83	6.69
75	2-Naphthyl	-	6.12	5.91	5.43
					
76	3-CH_3	-	5.62	5.33	5.50
77	3-Cl	-	5.62	5.32	5.64
78	4-CH_3	-	5.00	5.64	5.02
79	3-F	-	5.36	5.31	5.49
80	3-Br	-	6.11	5.94	6.41
81	3-OCF_3	-	6.31	6.20	6.11
82	3-NO_2	-	6.08	5.98	6.71

83	2-NH_2, 5-NO_2	-	6.68	6.54	6.15
84	2-NH_2, 5-Cl	-	6.14	6.02	6.48
85	2-NH_2, 5-Br	-	6.19	6.69	6.48
86	2-NH_2, 5-I	-	6.51	6.82	7.21
87	2-NH_2, 5-OCF_3	-	7.06	7.17	7.23
88	2-NH_2, 5-CF	-	7.59	7.23	7.24



89	4-Cl	-	7.59	7.47	7.13
90	4-Br	-	6.94	7.58	6.27
91	4-CH_3	-	7.70	7.25	7.20
92	4-Et	-	7.96	7.44	7.28
93	4-Vinyl	-	7.72	7.54	7.40
94[p]	4-OCH_3	-	7.70	7.53	6.80
95	4-OH	-	7.38	7.39	7.77
96[p]	4-Cl, 3-NH_2	-	8.39	7.82	7.54
97	4-CH_3, 3-NH_2	-	8.21	8.05	7.79
98	4-OCH_3, 3-NH_2	-	8.28	8.24	7.78
99	4-OH, 3-NH_2	-	7.85	7.91	7.72
100[p]	4-OCH_3, 3-OH	-	7.28	7.85	7.13
101	4-OH, 3-OCH_3	-	7.41	7.57	7.56
102	2,4-$(CH_3)_2$	-	8.49	7.91	7.85
103	2,4-Cl_2	-	7.02	7.43	7.13

P: Used as test (prediction) set
v: Used as validation set



شکل۳. نمودار مقادیر پیش‌بینی شده pIC_{50} بر حسب مقادیر تجربی برای سری آموزش و تست به روش SW-MLR

۳-۲. مدل سازی و پیش بینی توسط شبکه های عصبی مصنوعی

در این روش توصیف کننده های انتخاب شده وارد شبکه عصبی مصنوعی می‌شوند. پردازش داده‌ها در محیط ویندوز۱۰و با استفاده از نرم افزار MATLAB انجام شد. یک شبکه سه لایه با تابع انتقال سیگموئیدی برای نرون ها طراحی شده است. مقادیر اولیه وزن ها بطور تصادفی از بازه [1 , 0] بوده و قبل از عمل آموزش مقادیر ورودی و خروجی در فاصله [0.9 , 0.1] نرمال شده است. بهینه‌سازی و بهنگام کردن وزنها و بایاس‌ها بوسیله الگوریتمBP [۱] انجام شده است. مجموعه داده‌ها به سه گروه تقسیم شده است: مجموعه آموزش، مجموعه ارزیابی و مجموعه تست. مجموعه آموزش(۵۰٪ داده ها) جهت آموزش دادن شبکه عصبی مصنوعی، مجموعه ارزیابی (۲۰٪ داده ها) برای ارزیابی مدل در طی آموزش دادن شبکه و ایجاد مدل مناسب و مجموعه پیش‌بینی (۳۰٪ داده ها) برای تست مدل ایجاد شده به کار رفت.

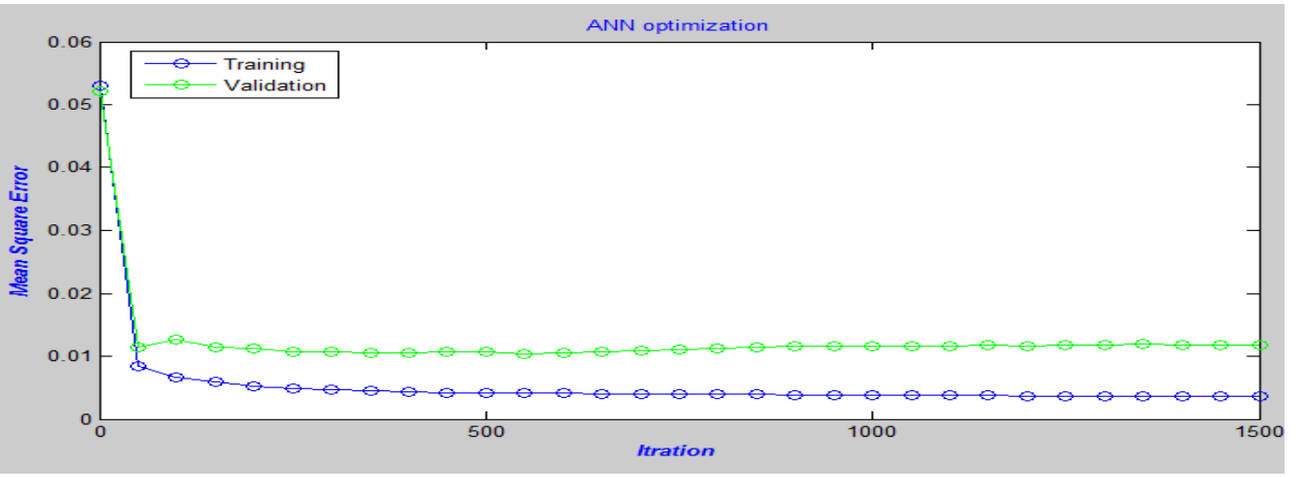
تعداد نرون‌ها در لایه ورودی با تعداد توصیف کننده های وارد شده به شبکه‌های عصبی مصنوعی برابر است. به ازای هر تعداد توصیف کننده وارد شده به شبکه عصبی، تعداد نرون‌ها در لایه مخفی بهینه می‌شود. بدین ترتیب که به ازای هر مدل ANN، تعداد نرون‌ها در لایه مخفی از ۱ تا ۱۰ تغییر داده شده و مقادیر RMSE برای مجموعه‌های آموزشی و پیش‌بینی محاسبه گردید. از رسم مقادیر RMSE برحسب تعداد نرون‌ها در لایه مخفی، تعداد نرون‌های لایه مخفی بهینه شد. سپس بقیه پارامترها از جمله وزنها و بایاس‌ها، سرعت یادگیری و مومنتوم نیز بهینه گردید. جدول ۳ مشخصات شبکه عصبی مصنوعی بهینه شده را نشان می‌دهد.

جدول ۳. ساختار و مشخصات ANN تولید شده

No. of nodes in the input layer	۷
No. of nodes in the hidden layer	۵
No. of nodes in the output layer	۱
Learning rate	۰/۴
Momentum	۰/۳
Number of epochs	۱۰۰۰
Transfer function	Sigmoid

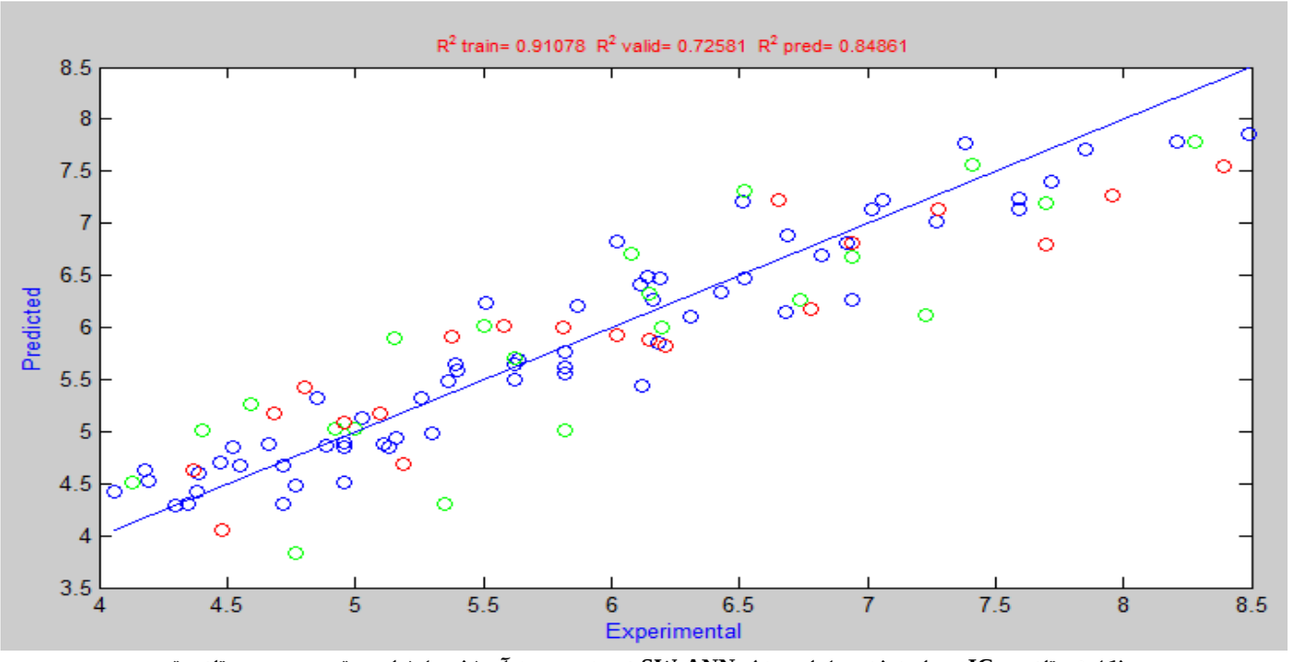
برای جلوگیری از Overfitting در طول آموزش مقادیر RMSE بعد از هر ۵۰ بار چرخه آموزشی، محاسبه و ثبت گردید. شکل ۴ نمودار میزان خطا بر حسب تعداد دورها برای این داده‌ها را نشان می‌دهد. همانطور که ملاحضه می‌شود میزان خطا برای سری آموزش همواره در حال کاهش است اما برای سری ارزیابی در ۵۵۰ چرخه آموزش، کمترین خطا مشاهده میشود و بعد از آن افزایش می‌یابد. بنابراین این مقدار به عنوان مقدار بهینه تعداد چرخه‌های آموزش انتخاب شد.

۱. Back-propagation



شکل ٤. مقادیر RMSE برای مجموعه‌های آموزشی و ارزیابی برحسب تعداد چرخه‌های آموزش

با استفاده از مدل ANN بهینه شده مقادیر فعالیتهای بازدارندگی(pIC_{50}) ترکیبات مورد نظر در مجموعه آموزشی، ارزیابی و پیش‌بینی (تست) مورد محاسبه قرار گرفت و درجدول(۲) نشان داده شده است. درشکل (۵) نیز مقادیر محاسبه شده pIC_{50} ترکیبات مورد نظر در مجموعه‌های مختلف برحسب مقادیر تجربی رسم شده‌اند.



شکل۵. مقادیر pIC_{50} محاسبه شده براساس مدل SW-ANN در سه مجموعه آموزشی، ارزیابی و تست برحسب مقادیر تجربی

۳-۳. ارزیابی مدل با استفاده از پارامترهای آماری

مطابق جدول ۴ دو پارامتر آماری، جهت ارزیابی توانایی پیش‌بینی مدل های ساخته شده به روش‌های MLR, ANN به کار گرفته شد. همانطور که در این جدول مشاهده می شود تمام پارامترهای آماری برای روش غیرخطی شبکه عصبی مصنوعی بهتر از روش خطی رگرسیون خطی چندگانه است.

جدول ٤. پارامترهای آماری برای مدل‌های انتخاب شده

		SW-MLR	SW-ANN
R^2	سری آموزش	۰/۸۲۵	۰/۹۱۰
	سری تست	۰/۸۴۷	۰/۸۴۸
RMSE	سری آموزش	۰/۴۲۵	۰/۳۳۰
	سری تست	۰/۲۳۸	۰/۲۱۲

٤. نتیجه گیری

امروزه با توجه به این موضوع که تحقیقات آزمایشگاهی نیازمند صرف هزینه و وقت بسیار بوده و در مواقعی با خطرات و عوارض زیست محیطی همراه است، استفاده از مدل‌های پیش‌بینی‌کننده جهت انجام تحقیقاتِ با هدف و نتیجه بخش، بسیار گسترش یافته است. با استفاده از روشهای QSAR مدل‌های بسیاری توسعه یافته‌اند که قابلیت اطمینان و قدرت پیش‌بینی‌کنندگی بالایی دارند. سهولت، سادگی، دقت بالا، قابلیت اطمینان و قدرت پیش‌بینی‌کنندگی بالا از مزایای اغلب مدل‌های ارائه شده با این روش‌ها میباشد. در این تحقیق از دو روش رگرسیون خطی چندگانه و شبکه های عصبی مصنوعی جهت مدل سازی و پیش‌بینی فعالیت دارویی برخی مشتقات کموکین ها استفاده شد. ابتدا به کمک رگرسیون مرحله‌ای، ۷ توصیف‌کننده‌ایی که بیشترین ارتباط را با فعالیت دارویی داشتند شامل MATS2p، PCWTe، RDF045m، RDF065m، RDF115m، C-003 و C-040 انتخاب شدند. جهت مدلسازی از دو روش رگرسیون خطی چندمتغیره(MLR) و شبکه های عصبی مصنوعی(ANN) استفاده گردید. هر دو روش خطی و غیر خطی نتایج نسبتا قابل قبولی ارائه داده اند. بنابراین از هر دو این مدل ها می توان برای پیش بینی فعالیت دارویی ترکیبات مشابه استفاده کرد. یعنی با محاسبه توصیف کننده های یک ترکیب دارویی جدید که با نرم افزار دراگون قابل انجام است و وارد کردن آنها به شبکه عصبی بهینه شده یا مدل MLR بدست آمده می توان مقدار فعالیت دارویی آن را بدون انجام آزمایشات وقت گیر و هزینه بر پیش بینی نمود.

٥. مراجع

[1] Coperchini, F., Chiovato, L., Croce, L., Magri, F., & Rotondi, M. (2020). The cytokine storm in COVID-19: An overview of the involvement of the chemokine/chemokine-receptor system. *Cytokine & growth factor reviews*, *53*, 25-32.

[2] Mollica Poeta, V., Massara, M., Capucetti, A., & Bonecchi, R. (2019). Chemokines and chemokine receptors: new targets for cancer immunotherapy. *Frontiers in immunology*, *10*, 379.

[3] Zhang, H., Chen, K., Tan, Q., Shao, Q., Han, S., Zhang, C., ... & Wu, B. (2021). Structural basis for chemokine recognition and receptor activation of chemokine receptor CCR5. *Nature communications*, *12*(1), 4151.

[4] Liu, K., Wu, L., Yuan, S., Wu, M., Xu, Y., Sun, Q., ... & Liu, Z. J. (2020). Structural basis of CXC chemokine receptor 2 activation and signalling. *Nature*, *585*(7823), 135-140.
[5] Miao, M., De Clercq, E., & Li, G. (2020). Clinical significance of chemokine receptor antagonists. *Expert opinion on drug metabolism & toxicology*, *16*(1), 11-30.
[6] Zacarías, N. V. O., Bemelmans, M. P., Handel, T. M., de Visser, K. E., & Heitman, L. H. (2021). Anticancer opportunities at every stage of chemokine function. *Trends in pharmacological sciences*, *42*(11), 912-928.
[7] Romero, J. M., Grünwald, B., Jang, G. H., Bavi, P. P., Jhaveri, A., Masoomian, M., ... & Gallinger, S. (2020). A four-chemokine signature is associated with a T-cell–inflamed phenotype in primary and metastatic pancreatic cancer. *Clinical Cancer Research*, *26*(8), 1997-2010.
[8] Jaeger, K., Bruenle, S., Weinert, T., Guba, W., Muehle, J., Miyazaki, T., ... & Standfuss, J. (2019). Structural basis for allosteric ligand recognition in the human CC chemokine receptor 7. *Cell*, *178*(5), 1222-1230.
[9] Feigl, F. F., Stahringer, A., Peindl, M., Dandekar, G., Koehl, U., Fricke, S., & Schmiedel, D. (2023). Efficient Redirection of NK Cells by Genetic Modification with Chemokine Receptors CCR4 and CCR2B. *International Journal of Molecular Sciences*, *24*(4), 3129.
[10] Markx, D., Schuhholz, J., Abadier, M., Beier, S., Lang, M., & Moepps, B. (2019). Arginine 313 of the putative 8th helix mediates Gαq/14 coupling of human CC chemokine receptors CCR2a and CCR2b. *Cellular signalling*, *53*, 170-183.
[11] Chtita, S., Belhassan, A., Bakhouch, M., Taourati, A. I., Aouidate, A., Belaidi, S., ... & Lakhlifi, T. (2021). QSAR study of unsymmetrical aromatic disulfides as potent avian SARS-CoV main protease inhibitors using quantum chemical descriptors and statistical methods. *Chemometrics and Intelligent Laboratory Systems*, *210*, 104266.
[12] Ewees, A. A., Abualigah, L., Yousri, D., Algamal, Z. Y., Al-Qaness, M. A., Ibrahim, R. A., & Abd Elaziz, M. (2022). Improved Slime Mould Algorithm based on Firefly Algorithm for feature selection: A case study on QSAR model. *Engineering with Computers*, *38*(3), 2407-2421.
[13] Mozafari, Z., Chamjangali, M. A., & Arashi, M. (2020). Combination of least absolute shrinkage and selection operator with Bayesian Regularization artificial neural network (LASSO-BR-ANN) for QSAR studies using functional group and molecular docking mixed descriptors. *Chemometrics and Intelligent Laboratory Systems*, *200*, 103998.
[14] Mozafari, Z., Arab Chamjangali, M., Arashi, M., & Goudarzi, N. (2021). Performance of smoothly clipped absolute deviation as a variable selection method in the artificial neural network-based QSAR studies. *Journal of Chemometrics*, *35*(5), e3338.
[15] Vahedi, Nafiseh, Majid Mohammadhosseini, and Mehdi Nekoei. "QSAR Study of PARP Inhibitors by GA-MLR, GA-SVM and GA-ANN Approaches." *Current Analytical Chemistry* 16, no. 8 (2020): 1088-1105.
[16] Cong, H., Zhao, X., Castle, B. T., Pomeroy, E. J., Zhou, B., Lee, J., ... & Zhuang, C. (2018). An indole–chalcone inhibits multidrug-resistant cancer cell growth by targeting microtubules. *Molecular pharmaceutics*, *15*(9), 3892-3900.
[17] Triolascarya, K., Septiawan, R. R., & Kurniawan, I. (2022). QSAR Study of Larvicidal Phytocompounds as Anti-Aedes Aegypti by using GA-SVM Method. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *6*(4), 632-638.
[18] Pramana, I. K. A. P. P., Septiawan, R. R., & Kurniawan, I. (2022). QSAR Study on Diacylgycerol Acyltransferase-1 (DGAT-1) Inhibitor as Anti-diabetic using PSO-SVM Methods. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *6*(5), 735-741.

[19] Ai, H., Wu, X., Zhang, L., Qi, M., Zhao, Y., Zhao, Q., ... & Liu, H. (2019). QSAR modelling study of the bioconcentration factor and toxicity of organic compounds to aquatic organisms using machine learning and ensemble methods. *Ecotoxicology and Environmental Safety*, *179*, 71-78.
[20] Minoru I., Tatsuki S., Ken-ichiro K., (2004) Small molecule inhibitors of the CCR2b receptor. Part 1: Discovery and optimization of homopiperazine derivatives. *Bioorganic & Medicinal Chemistry Letters*, *14*, 5407–5411.
[21] Wilna J., Ken-ichiro K., Michele M. (2004) Small molecule antagonists of the CCR2b receptor. Part 2: Discovery process and initial structure–activity relationships of diamine derivatives. *Bioorganic & Medicinal Chemistry Letters*. *14*, 5413–5416.
[22] Wilna J., Ken-ichiro K., Michele M. (2008) Potent antagonists of the CCR2b receptor. Part 3: SAR of the (R)-3-aminopyrrolidine series. *Bioorganic & Medicinal Chemistry Letters*. *18*, 1869–1873.
[23] Billones, L. T., Gonzaga, A. C., & Billones, J. B. (2019). Molecular descriptors for drugs: A discriminant analysis. *Phil J Health Res Devel*, *23*(4), 11-16.

The application of multiple linear regression and artificial neural networks to study the quantitative structure-activity relationship of a group of chemokine derivatives

Mohammad Reza Kianasab[1], Mehdi Nekoei[1]*, Majid Mohammadhosseini[1], Behnam Mahdavi[2], Tahmineh Baheri[3]

[1]Department of Chemistry, Shahrood Branch, Islamic Azad University, Shahrood, Iran

[2]Department of Chemistry, Faculty of Science, Hakim Sabzevari University, Sabzevar, Iran

[3]Department of Anti-Narcotics, Amin University of Police Sciences, Tehran, Iran

Submited: 25 October 2022, Revised: 11 February 2023, Accepted: 04 March 2023

Abstract

A quantitative structure-activity relationship (QSAR) study was conducted to predict the pharmacological activity of some chemokine derivatives using multiple linear regression and artificial neural networks (ANN). At first, the structure of pharmaceutical compounds was drawn and optimized with the help of Hypercam software. Then, a wide range of molecular descriptors were calculated by Dragon software. After reducing the number of descriptors that had a correlation above 0.9 and the descriptors that were more than 90% similar, stepwise regression was used to obtain the best descriptors that were most related to the pharmacological activity of the target compounds. became 7 descriptors including MATS2p, PCWTe, RDF045m, RDF065m, RDF115m, C-003 and C-040 were selected. Then, multiple linear regression (MLR) and artificial neural networks (ANN) methods were used to model and predict the activity of test series compounds. The obtained results show that both methods provide acceptable results that can be used to predict new pharmaceutical compounds.

Keywords: *Quantitative structure-activity relationship, chemokine derivatives, multiple linear regression, artificial neural networks.*

***Corresponding author :** Mehdi Nekoei

Address: Department of Chemistry, Shahrood Branch, Islamic Azad University, Shahrood, Iran

Tel: 02332394289 **E-mail:** m_nekoei1356@yahoo.com