دانشگاه آزاد اسلامی، واحد اراک
شاپا ٤٦٦٨-٢٠٠٨
http://jer.iau-arak.ac.ir

فصلنامه تخصصی تحقیقات حشره‌شناسی
(علمی- پژوهشی)
جلد ۱۱، شماره۳، سال ۱۳۹۸، (۲۵۷-۲٤٥)

پیشگویی سمیت و ضریب توزیع اکتانول-آب آفت کش‌های آلی کلردار با استفاده ازتوصیف گرهای مولکولی و روش‌های الگوریتم ژنتیک و رگرسیون خطی چند متغیره

*زهرا زمانی[1] ، فاطمه شفیعی[2]**

۱- دانشجوی دکتری شیمی فیزیک،گروه شیمی،واحد اراک،دانشگاه آزاداسلامی، اراک، ایران
۲ - استادیار شیمی فیزیک،گروه شیمی، واحداراک، دانشگاه آزاداسلامی، اراک، ایران

چکیده

در این تحقیق ازمطالعه ارتباط کمی ساختار- فعالیت (QSAR) جهت پیشگویی فعالیت‌های آفت کش‌های آلی کلردار استفاده شد. ابتدا به کمک نرم‌افزارگوس ویو ۰۵ ساختار شیمیایی تمام مولکول‌های مورد مطالعه رسم گردید و سپس بهینه‌سازی با نرم‌افزار گوسین ۰۹ و روش هارتری-فاک و سری پایه *G ۳۱-٦ انجام پذیرفت. خواص فیزیکی- شیمیایی نظیر سمیت ($logLD_{50}$) وضریب توزیع اکتانول-آب (logP) از منابع معتبر علمی به‌دست آورده شده است. برای تمامی سموم مورد مطالعه توصیف‌گرهای مناسب با استفاده از نرم‌افزار دراگون و روش‌های الگوریتم ژنتیک و برگشتی تعیین شدند. سپس ارتباط بین توصیف گرهای مولکولی و فعالیت‌ها با استفاده از روش چند متغیره خطی مورد بررسی قرار گرفت. به منظور ساخت و آزمون مدل‌های QSAR داده‌ها به‌طور تصادفی به دو دسته :آموزشی (۸۰ درصد) و آزمون (۲۰ درصد) تقسیم گردیدند. مدل‌ها با استفاده از ضرایب آماری نظیر ضریب همبستگی(R)، مجذور آن (R^2) و ریشه میانگین مربع خطا (RMSE) انتخاب شدند. برای بررسی توانایی پیش بینی و اعتبار مدل‌ها از اعتبار سنجی تقاطعی "یکی بیرون" و اعتبار سنجی خارجی استفاده شد. اعتبار سنجی خارجی با استفاده از ضرایب رگرسیونی مجموعه آزمون انجام گردید. نتایج اعتبار سنجی و کیفیت بالای ضرایب مدل‌ها نشان می‌دهد که مدل‌های GA-MLR به‌دست آمده مدل‌های قابل قبول QSAR می‌باشند .این مدل‌ها می‌توانند جهت شناسایی توصیف گر های مناسب و پیشگویی فعالیت‌های آن‌ها به کار گرفته شوند.

واژه‌های کلیدی: "سمیت" ; "ارتباط کمی ساختار- فعالیت" ; "روش رگرسیون خطی چند متغیره "; الگوریتم ژنتیک" ; "آفت کش‌های آلی کلردار" ; "ضریب توزیع اکتانول-آب".

* نویسنده رابط، پست الکترونیکی: *shafa38@yahoo.com*

تاریخ دریافت مقاله:۹۸/٤/۱٦ - تاریخ پذیرش مقاله:۹۸/۹/٤

مقدمه

آفت‌کش‌های آلی کلر دار[16] (OCPs) به سه گروه اصلی ددتِ(DDT) و آنالوگ‌های وابسته آن، سیکلودین‌ها و ترکیبات وابسته آن و هگزا کلرو بنزن تقسیم می‌شوند. این سموم به‌دلیل آنکه به کندی متابولیزه شده، از نظر شیمیایی پایدارند و در محیط باقی می‌مانند باعث کاهش باروری پرندگان و سایر جانداران شده و بروز برخی سرطان‌ها به آن‌ها نسبت داده می‌شود دارای منع مصرف می‌باشند (Jayaraj, *et al*., 2016). ددت از راه گوارش و تنفس جذب می‌شود و در صورت استفاده از حلال‌های روغنی در تهیه محلول سمی، امکان جذب پوستی آن هم وجود دارد (Hagmar, *et al*., 2001). سیکلودین‌ها (آلدرین، اندرین، دی اندرین، هپتاکلرو و اندوسولفان) و ترکیبات وابسته آن (توگزافن و دی انوکلرو) خاصیت حشره‌کشی بیشتری نسبت به ددت دارند (Jayaraj, *et al*., 2016). هگزا کلرو بنزن به تنهایی یکی از گروه‌های ترکیبات سمی کلردار را تشکیل می‌دهد (Zaffar, *et al*.,2016) و ایزومر گامای آن بنام گاما هگزاکلرو بنزن (لیندان) هنوز در امور بهداشتی استفاده می‌شود.

مقالات بسیاری در مورد سموم کلردار منتشر شده است که در ادامه به تعدادی از آن‌ها اشاره می‌شود. سموم آلی کلردار در نمونه‌های آب رودخانه کارون به روش‌های کروماتوگرافی گازی یونش شعله و میکرو استخراج مایع- مایع مورد شناسایی قرار گرفته است (Behfar, *et al*., 2013). نتایج این تحقیق نشان داد ادغام دو روش فوق روشی با عملکرد ساده، سرعت زیاد و درصد بازیابی بالا را ایجاد می‌نماید. مدل‌های ارتباط کمی ساختار – فعالیت برای پیشگویی میزان سمیت آفت کش‌های جدید از آنالوگ‌های DDT طراحی گردیده‌اند (Saini & Kumar, 2014). با استفاده از روش رگرسیون چند متغیره خطی ارتباط سمیت آنالوگ‌های DDT با قطبیت، آبدوستی و ثابت‌های استخلاف و تفت[17] مورد بررسی قرار گرفته (Yehia & Abdel, 1983) و نتایج نشان دادند ثابت‌های استخلاف و تفت، ارتباط خطی مطلوبی با سمیت (logLD50) ترکیبات مورد مطالعه دارند.

مدل‌های خطی جهت پیش‌بینی تابع توزیع اکتانول- آب برای مشتقات DDT و دی کوفول با استفاده از کروماتوگرافی مایع با فشار بالا و مخلوطی از حلال‌های آب و اتانول به‌عنوان فاز متحرک ارائه شده است (Han, *et al*., 2011). همچنین با استفاده از کروماتوگرافی گازی مدل‌هایی جهت پیشگویی تابع توزیع اکتانول-هوا (KO/A) و فشار بخاربرای ۹ سم آلی کلر دار از ایزومرهای DDT وDDE ارائه شده است (Zhang, *et al*., 2009).

مقالات بسیاری در خصوص آلوده بودن محصولات عمدتا گوشت، لبنیات و ماهی با آفت کش‌های آلی کلر دار منتشر شده است (Fitzgerald, *et al*., 2001; Hagmar, *et al*., 2001; Mwevura, *et al*., 2002; Bradman, *et al*., 2007).

ضریب توزیع اکتانول-آب[18] (logP)

ضریب توزیع معیاری برای میزان انحلال ماده مورد نظر در دو فاز آلی و آبی است و به‌عنوان یک ثابت فیزیکی در نظر گرفته می‌شود. ضریب توزیع آب و اکتانول Po/w یکی از خواص شیمی فیزیکی قابل قبول، برای اندازه گیری آب گریزی ترکیبات شیمیایی است و همچنین گزینه مفیدی برای تخمین توزیع یک ترکیب (دارو، سم و...) در سطح بدن می‌باشد. ترکیبات آب‌گریز با نسبت اکتانول به آب بالا معمولا در نواحی آب گریز بدن مانند غشاء دولایه لیپیدی سلول توزیع

[1]- Organochlorine pesticides
[2]- Taft
[1]- Octanol/ water partition coefficient

می‌شوند. در مقابل، ترکیبات آب‌دوست (نسبت اکتانول به آب پایین) در درجه نخست در نواحی آبی بدن مانند پلاسمای خون دیده می‌شوند (Scheytt, *et al*., 2005, Shargel, et al., 2012).

مواد و روش‌های محاسباتی

در این مقاله به بررسی ارتباط کمی ساختار-فعالیت[19] (QSAR) پرداخته شده است و در این راستا ارتباط سمیت (LD_{50})[20] و ضریب توزیع اکتانول-آب با ساختار مولکولی سموم آلی کلر دار نظیر ددت، کلردان، لیندن،آلدرین، اندوسولفان و ایزودرین و.... با کمک توصیف گرهای مولکولی، روش های چند متغیره خطی و الگوریتم ژنتیک مورد بررسی قرار گرفته و مدل‌هایی جهت پیشگویی مقادیرسمیت و ضریب توزیع اکتانول-آب سموم آلی کلردار ارائه گردیده است. بدین

منظور ۱۶ سم مورد مطالعه را به صورت تصادفی به دو دسته آموزشی با ۱۲ سم و مجموعه پیش‌بینی یا آزمون با ٤ سم تقسیم نموده و مقدار LD_{50} آن ها با کمک مقاله علمی (Jayaraj, *et al*., 2016) و مقادیر ضریب توزیع اکتانول-آب (logP) با استفاده از منبع معتبر علمی (Web search engine developed by ChemAxon) جمع‌آوری شد. نام و فرمول شیمیایی سموم به همراه سمیت، میانگین سمیت، لگاریتم میانگین سمیت و تابع توزیع اکتانول-آب آن‌ها در جدول ۱ ثبت می باشد.

جدول ۱- نام و فرمول شیمیایی ۱۶ آفت‌کش آلی کلردار به همراه سمیت، میانگین سمیت، لگاریتم میانگین سمیت و ضریب توزیع اکتانول-آب آن‌ها (*ترکیبات انتخاب شده برای مجموعه آزمون و اعتبارسنجی خارجی)

Table 1. The chemical names and formulas of 16 organochlorine pesticides and their toxicity ($logLD_{50}$) and octanol / water distribution coefficient (logP)(* Compounds selected for test set in external validation procedure).

No.	Chemical name	Chemical formula	logP	Toxicity LD_{50}	Mean of toxicity	log LD_{50}
1	Aldrin	$C_{12}H_8Cl_6$	5.270	39-60	49.5	1.695
2	Endrin	$C_{12}H_8Cl_6O$	4.480	3	3	0.477
3	Heptachlor	$C_{10}H_5Cl_7$	5.240	40-220	130	2.114
4	Isobenzan (telodrin)	$C_9H_4Cl_8O$	5.224	4.8	4.8	0.681
5	Isodrin	$C_{12}H_8Cl_6$	5.270	8.8	8.8	0.944
6	Lindane	$C_6H_6Cl_6$	3.644	88-270-	179	2.253
7	Mirex	$C_{10}Cl_{12}$	6.223	600-740	670	2.826
8	Pentachlorophenol	C_6Cl_5OH	4.659	27-211	119	2.075
9	1,4- dichlorobenzene	$C_6H_4Cl_2$	2.993	1516-2138	1827	3.262
10	Benzene hexachloride (BHC)	$C_6H_6Cl_6$	5.607	10000	10000	4
11	Endosulfan	$C_9H_6Cl_6O_3S$	3.690	18-220	119	2.075
12	Chlordane	$C_{10}H_6Cl_8$	5.683	200-700	450	2.653
13	*Chloropropylate	$C_{17}H_{16}Cl_2O_3$	5.240	5000	5000	3.699
14	*Dichloro diphenyl dichloroethane (DDE)	$C_{14}H_{10}Cl_4$	6.190	800-1240	1020	3.008
15	*Dicofol	$C_{14}H_9Cl_5O$	5.600	684-1495	1089.5	3.037
16	Dieldrin	$C_{12}H_8Cl_6O$	4.481	46	46	1.663

[2]- Quantitative Structure–Activity Relationship (QSAR)
[3]- Lethal Dose, which causes the death of 50% (one half) of a group of test animals

منظور ۱۶ سم مورد مطالعه را به صورت تصادفی به دو دسته آموزشی با ۱۲ سم و مجموعه پیش بینی یا آزمون با ٤ سم تقسیم نموده و مقدار LD_{50} آن‌ها با کمک مقاله علمی (Jayaraj, *et al*., 2016) و مقادیر ضریب توزیع اکتانول-آب (logP) با استفاده از منبع معتبر علمی (Web search engine developed by ChemAxon) جمع‌آوری شد. نام و فرمول شیمیایی سموم به همراه سمیت، میانگین سمیت، لگاریتم میانگین سمیت و تابع توزیع اکتانول-آب آنها در جدول ۱ ثبت می‌باشد.

جهت انجام این تحقیق ابتدا ساختار اولیه تمامی سموم با استفاده از نرم‌افزار گوس ویو۰۵[۲۱] ترسیم شده و با کمک نرم افزار گوسین.۰۹[۲۲] به روش آغازین وخط دستور HF/6-31G* اقدام به بهینه سازی ترکیبات گردید. رفتار شیمیایی یک ترکیب به ساختار مولکول‌های آن وابسته است به منظور برقراری ارتباط ساختار–فعالیت از توصیف گر های مولکولی بهره گرفته شده است. به منظور محاسبه انواع توصیف گرهای مولکولی (Todeschini & Consonni, 2009 & Todeschini Consonni, 2000 &)، نرم افزاردراگون[۲۳] به کار گرفته شد این نرم افزار برای هر ترکیب مورد مطالعه تعداد زیادی توصیف گر محاسبه نمود لذا به منظور تعیین مناسب ترین توصیف گرها از روش الگوریتم ژنتیک[۲٤] (Saxena & Prathipati, 2003, Ghosh & Bagchi, 2009) برنامه متلب 2010a[۲۵] و روش برگشتی[۲٦] (Goodarzi, et al., 2012) استفاده گردید. پس از تقلیل توصیف‌کننده‌های نامناسب، آن‌هایی که بیشترین میزان همبستگی را بافعالیت‌های مورد مطالعه برقرار نموده بودند و کمترین همبستگی را با توصیفگر های دیگر داشتند انتخاب نموده و در مدل‌سازی به روش رگرسیون خطی چند متغیره[۲۷] (MLR)از آن‌ها استفاده گردید و بدین وسیله چندین مدل برای هر فعالیت مورد مطالعه به‌دست آمد که با توجه به ضرایب آماری (Chatterjee, *et al*., 2013) معادله که عبارتند از ضریب همبستگی(R)، مجذور آن(R^2)،ضریب فیشر(F)، ضریب تعیین تعدیل شده R^2_{adj}، ریشه میانگین مربع خطا (RMSE)، آماره دوربین–واتسون(DW) و تعداد حداقل توصیف گرهای موجود در هر مدل، مناسب ترین مدل برای هر فعالیت مورد مطالعه تعیین گردید. در نهایت مدل‌های انتخاب شده مورد اعتبار سنجی قرار گرفتند.

نتایج و بحث

مدل‌های اولیه به‌دست آمده برای سمیت ($logLD_{50}$)

با توجه به اعداد گزارش شده برای سمیت که دارای حداقل و حداکثر مقدار می‌باشند سه نوع محاسبه یکبار با مقدار حداقل، یکبار با مقدار حداکثر و یکبار با مقدار میانگین سموم انجام گردید اما رابطه خطی مناسبی بین سمیت (مقادیر حداکثر، حداقل و میانگین) و توصیفگر های به‌دست آمده از نرم افزار دراگون حاصل نشد. ارتباط لگاریتم میانگین سمیت با توصیفگرها نیز مورد بررسی قرار گرفت که مشخص گردید رابطه خطی مطلوبی بین آن‌ها برقرار است .در جدول ۲ توصیف گرها و ضرایب آماری به‌دست آمده از نرم افزار آماری SPSS و روش رگرسیون خطی چندگانه (MLR) و روش برگشتی نشان داده شده است. همانطوری که از جدول ۲ مشخص است ضرایب آماری به‌دست آمده برای مدل‌های مختلف نزدیک به یکدیگر می‌باشد. اما یکی از شرایط معادله مناسب وجود تعداد کم توصیف کننده در آن است لذا مدل

[21]- Gauss View 05
[22]- Gaussian 09
[23]- Dragon
[4]- Genetic Algorithm (GA)
[1]- MATLAB 2010a
[2]- Backward
[3]- Multiple Linear Regression(MLR)

۸ با سه توصیف کننده مناسب‌ترین مدل در نظر گرفته شد (معادله۱). این توصیف گرها در دسته توصیف گرهای اتصالی[۲۸] (X3A,X1Av) و مسیر[۲۹] (PiID) تقسیم‌بندی می‌شوند.

(1) $\log LD_{50}$= -22.184+0.30 (piID)+ 34.516 (X3A) + 17.654 (X1AV)
(N=12, R=0.884, R^2 =0.781, R^2adj =0.686, F=8.658, DW=2.082, Sig=0.007, RMSE= 0.718)

جدول ۲- ضرایب آماری به‌دست آمده برای لگاریتم سمیت با روش GA-MLR

Table 2. Statistical parameters of the models calculated with the GA-MLR method for toxicity (logLD50).

Model	Independent variables	R	R^2	R^2_{adj}	RMSE
1	X5sol, X4Av, IC2, Wap, CIC3, X3A, X1Av, Mv, piID, Har2	0.984	0.968	0.860	0.640
2	X5sol, X4Av, IC2, CIC3, X3A, X1Av, Mv, piID, Har2	0.984	0.968	0.828	0.645
3	X5sol, X4Av, CIC3, X3A, X1Av, Mv, piID, Har2	0.981	0.962	0.859	0.667
4	X5sol, X4Av, CIC3, X3A, X1Av, piID, Har2	0.970	0.941	0.835	0.683
5	X5sol, CIC3, X3A, X1Av, piID, Har2	0.938	0.880	0.736	0.698
6	X5sol, X3A, X1Av, piID, Har2	0.915	0.838	0.702	0.701
7	X5sol, X3A, X1Av, piID	0.903	0.816	0.711	0.709
8	X3A, X1Av, piID	0.884	0.781	0.686	0.718

مدل‌های اولیه به‌دست آمده برای لگاریتم تابع توزیع اکتانول-آب (logP)

برای لگاریتم ضریب توزیع اکتانول-آب، ۷ مدل با ۲ الی ۸ توصیفگر با استفاده از روش چند متغیره خطی به‌دست آمد که نتایج این بررسی در جدول ۳ آورده شده است. مدل ۷ با دو توصیفگر (IC1 ,TIC4) مناسبترین مدل انتخاب گردید. این دو توصیف گر در گروه توصیف گرهای اطلاعاتی[۳۰] می باشند. معادله۲ حاصل این بررسی می‌باشد.

logP=6.915 - 10.555 (IC1) +0.378 (TIC4) (2)
(N=12, R=0.841, R^2 =0.707, R^2adj =0.714, F=6.501, DW=2.012, Sig=0.018, RMSE=0.825)

جدول ۳- ضرایب آماری به‌دست آمده برای لگاریتم تابع توزیع اکتانول-آب با روش GA-MLR

Table 3. Statistical parameters of the models calculated with the GA-MLR method for logP.

Model	Independent variables	R	R^2	R^2_{adj}	RMSE
1	S3K, IC1, MAXDN, TIC4, TIC1, STN, SIC0, WW	0.899	0.809	0.799	0.783
2	S3K, IC1, MAXDN, TIC4, TIC1, STN, SIC0	0.899	0.809	0.791	0.790
3	S3K, IC1, TIC4, TIC1, STN, SIC0	0.899	0.809	0.783	0.797
4	S3K, IC1, TIC4, TIC1, STN	0.897	0.805	0.765	0.799
5	IC1, TIC4, TIC1, STN	0.878	0.771	0.740	0.817
6	IC1, TIC4, TIC1	0.854	0.729	0.731	0.820
7	IC1, TIC4	0.841	0.707	0.714	0.825

[4]- Connectivity indices
[5]- Walk and path counts
[1]- Information indices

بررسی همبستگی[۳۱] بین متغیرهای مستقل

برای تائید مدل‌های پیشنهادی دو مرحله دیگر باید پشت سر گذاشته شود در اولین مرحله همبستگی بین متغیر های مستقل و یا همان توصیف گر های باقی‌مانده در مدل‌ها باید مورد بررسی قرار گیرد و در دومین مرحله اعتبار سنجی مدل‌ها بررسی می‌گردد. همبستگی بین توصیف گرها هر چه نزدیک به صفر باشد گویای مستقل بودن آنها است در حالی‌که همبستگی نزدیک به یک مشخص می‌نماید بین توصیفگر ها همبستگی کاملی وجود دارد و یکی از دو توصیف گر باید حذف گردد. برای بررسی همبستگی (Craney & Surles, 2002 & Pourbasheer, et al., 2017) دو ضریب در نظر گرفته می‌شود که عبارتند ازضریب نفوذ پذیری یا تاثیر پذیری[۳۲] (VIF) و دیگری ضریب پیرسون[۳۳] (PCC). این دو ضریب از خروجی نرم‌افزار SPSS بدست می آیند. ضریب نفوذپذیری که هم خطی متغیر های مستقل را بررسی می‌کند در صورتی‌که عددی بین یک و ده را داشته باشد مشخص کننده اینست که همبستگی کمی بین توصیفگر ها وجود دارد اما ضریب پیرسون هر چه به صفر نزدیکتر باشد گویای عدم همبستگی بین دو متغیر مستقل می‌باشد. در جدول ٤ ضرایب فوق به همراه تولرنس[۳٤] (عکس ضریب نفوذ پذیری) برای لگاریتم سمیت آورده شده است. با توجه به داده‌های جدول همبستگی بین دو توصیف گر X3A و PilD نزدیک به یک می‌باشد و ضریب نفوذ پذیری X3A بزرگتر از PilD است لذا این توصیف گر از ورودی نرم‌افزار حذف گردید. خروجی جدید نشان داد بین توصیف گرهای باقی‌مانده همبستگی کمی وجود داشته و ضریب نفوذ پذیری نیز از ده کمتر می‌باشد. معادله۳ حاصل این بررسی است.

$$\log LD_{50}=133.104+ 0.30\ (pilD) + 23.465\ (X1AV) \qquad (3)$$

$(N=12, R=0.864, R^2=0.746, R_{adj}=0.656, F=7.459, DW=1.96, Sig=0.007, RMSE=0.620)$

جدول٤- ضرایب همبستگی بین توصیفگر های مولکولی برای سمیت(معادله۱).

Table 4. Correlation coefficients between the molecular descriptors for toxicity (Equation1).

	Pearson correlations			Collinearity statistical		
	X3A	X1Av	PilD	Tolerance	VIF(1)	VIF(2)
X3A	1	0.776	0.909	0.051	19.544	-------
X1Av		1	0.727	0.364	2.748	1.53
piID			1	0.160	6.250	2.57

در جدول ۵ ضرایب پیرسون، تولرنس و نفوذپذیری توصیفگر های باقی‌مانده در مدل۲ آورده شده است. با توجه به مقدار ضرایب ذکر شده بین دو توصیفگر IC1، TIC4 همبستگی قابل ملاحظه‌ای وجود ندارد بنابراین معادله ۳ برای مرحله بعد مورد استفاده قرار می‌گیرد.

جدول۵- ضرایب همبستگی به دست آمده برای لگاریتم ضریب توزیع اکتانول-آب(معادله۲).

Table 5. Correlation coefficients between the molecular descriptors for logP (Equation2).

Pearson correlations			Collinearity statistical	
	IC1	TIC4	Tolerance	VIF
IC1	1	-0.258	0.933	1.071
TIC4		1	0.933	1.071

[1]- Multicollinearity
[32]- Variance inflation factor
[33]- Pearson coefficient correlation
[34]- Tolerance

اعتبار سنجی مدل‌ها

برای ارزیابی دقت نتایج، تصدیق و معتبر نمودن مدل‌ها باید از یکی از روش‌های اعتبار سنجی (Cramer III, 1988 & Votano, 2004 & Gramatica, 2007, Roy & Mitra, 2011, Saghaie, et al., 2013) استفاده نمود. در این تحقیق از روش اعتبار سنجی متقابل یا ضربدری[۳۵] و روش اعتبار سنجی خارجی[۳۶] استفاده شده است. برای این منظور از همان ابتدا سموم آلی کلردار مورد مطالعه به‌طور تصادفی به دو دسته آموزشی و آزمون تقسیم شدند. برای اعتبار سنجی متقابل یا ضربدری هر بار به‌طور تصادفی یکی از ترکیبات مجموعه آموزشی حذف و مقدار مجذور همبستگی با استفاده از معادله ٤ محاسبه گردید این عمل برای ۲۵٪ مجموعه آموزشی تکرار گردید و سپس میانگین مجذور همبستگی به‌صورت Q^2LOO[۳۷] محاسبه شد. درمعادله ٤، Y_i بیانگر مقدار مشاهده شده یا مقدار تجربی، $\hat{Y}_{i|i}$بیانگر مقدار پیشگویی شده یا محاسبه شده و $\overline{Y}$ بیانگر مقدار میانگین می‌باشد. Q^2LOO برای لگاریتم سمیت و ضریب توزیع اکتانول-آب به‌ترتیب ۰/۸۹۷ و ۰/۸۱۵ **محاسبه گردید.**

$$Q^2 = 1 - \frac{\sum(Y_i - \hat{Y}_{i|i})^2}{\sum(Y_i - \bar{Y})^2} \qquad Q^2 \leq 1 \qquad (4)$$

از مجموعه آزمون جهت انجام اعتبار سنجی خارجی استفاده گردید. برای این منظور با استفاده از مدل‌های به‌دست آمده برای مجموعه آزمون ضرایب آماری محاسبه شد. جدول ٦ مقادیر ضرایب آماری برای مجموعه آزمون و آموزشی فعالیت‌های مورد مطالعه را نشان می‌دهد.

جدول ٦- ضرایب آماری به‌دست آمده برای مجموعه آزمون و آموزشی logP و $logLD_{50}$

Table 6. Statistical parameters obtained for training and test sets of logP and $logLD_{50}$.

Activity	N	R	R^2	R^2_{adj}	RMSE	DW	Sig	F
$log(LD_{50})$	12	0.864	0.746	0.656	0.620	1.960	0.007	7.459
$log(LD_{50})$	4	0.908	0.824	0.889	0.265	2.178	0.060	38.623
logP	12	0.831	0.690	0.560	0.825	2.012	0.018	6.501
logP	4	0.898	0.806	0.747	0.599	1.918	0.088	9.870

ضرایب آماری به‌دست آمده برای مجموعه آزمون گویای معتبر بودن مدل‌های به‌دست آمده بوده و تائید می‌نماید با استفاده از معادلات و توصیف گرهای به‌دست آمده برای هر یک از فعالیت‌های مورد نظر می‌توان اقدام به پیشگویی آن فعالیت‌ها برای سموم آلی کلردار مطالعه نشده نمود.

آماره دوربین-واتسون[۳۸]

برای بررسی استقلال خطاهای مدل رگرسیون از آماره دوربین-واتسون (DW)استفاده می‌شود (Chatterjee & Simonoff, 2013). مقدار این ضریب بین صفر و چهار می‌باشد. اگر مقدار ضریب دوربین-واتسون به‌دست آمده نزدیک به ۲ باشد مشخص می‌نماید خطا های رگرسیون مستقل از هم می‌باشند که برای تمامی مدل‌های به‌دست آمده این آماره نزدیک به ۲ است پس می‌توان نتیجه گرفت خطاها معنادار نیستند.

[1]- Cross-validation (CV)
[2]- External validation
[3]- Leave - one - out (Q^2LOO)
[1]- Durbin-Watson Statistic

باقی‌مانده[۳۹]

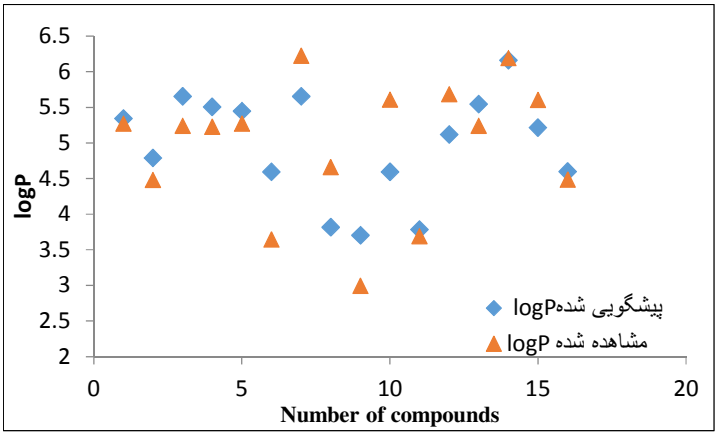
در آمار، باقی‌مانده تفاوت بین مقادیر مشاهده شده و پیشگویی شده است. یک نمودار باقی‌مانده نمایش مقادیر باقی‌مانده در روی محور عمودی و مقادیر مشاهده شده یا تجربی بر روی محور افقی می‌باشد. اگر نقاط به‌دست آمده از این نمودار به‌صورت تصادفی در اطراف محور افقی (خط صفر) پراکنده شده باشند و از الگویی خاصی تبعیت ننموده باشند گویای آن است که در مدل خطی به‌دست آمده هیچ خطای قاعده داری وجود ندارد. در جدول ۷ مقادیر مشاهده شده، پیشگویی شده و باقی‌مانده برای لگاریتم سمیت و لگاریتم ضریب توزیع سموم آلی کلردار آورده شده است. نمودارهای ۱و ۲ مقادیر مشاهده شده بر حسب مقادیر پیش‌بینی شده را برای هر دو مجموعه آموزشی و آزمون به نمایش می‌گذارد. نمودارهای ۳و ٤ مقادیر مشاهده شده بر حسب باقی‌مانده را برای لگاریتم سمیت و لگاریتم ضریب توزیع برای هر دو مجموعه آموزشی و آزمون نشان می‌دهد.

جدول ۷-مقادیر مشاهده شده، پیشگویی شده و باقیمانده برای لگاریتم سمیت و لگاریتم ضریب توزیع سموم آلی کلردار

Table 7. The experimental, predicted and residual values of logP and log(LD50) of organochlorine pesticides.

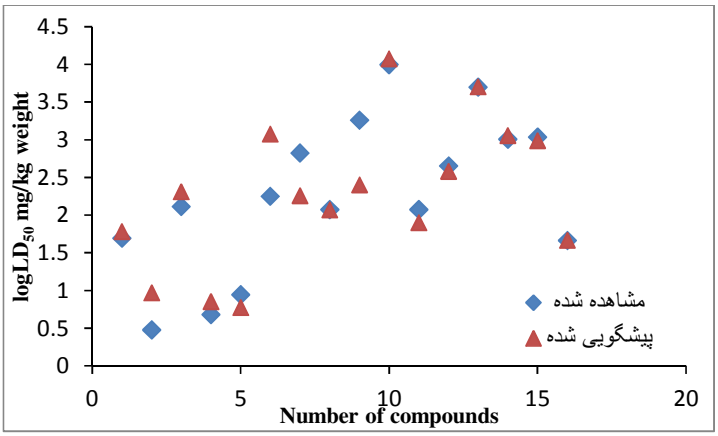
Compounds	Observed $\log(LD_{50})$	Predicted $\log(LD_{50})$	Residual	Observed logP	Predicted logP	Residual
Aldrin	1.695	1.784	-0.090	5.270	5.340	-0.070
Endrin	0.477	0.970	-0.493	4.480	4.788	-0.308
Heptachloro	2.114	2.312	-0.199	5.240	5.652	-0.412
Isobenzene	0.681	0.853	-0.172	5.224	5.506	-0.282
Isodrin	0.944	0.772	0.172	5.270	5.445	-0.175
lindane	2.253	3.075	0.823	3.644	4.591	-0.947
Mirex	2.826	2.257	0.569	6.223	5.653	0.571
Pentachlorophenol	2.075	2.069	0.006	4.659	3.818	0.841
1,4dichlorobenzene	3.262	2.402	0.859	2.993	3.703	-0.710
Benzene hexa chloro	4	4.075	-0.075	5.607	4.591	1.016
Endosulfan	2.075	1.900	0.175	3.690	3.782	-0.092
Chlordane	2.653	2.581	0.072	5.683	5.114	0.569
Chloropropylate*	3.699	3.703	0.004	5.240	5.544	-0.304
DDE*	3.008	3.057	-0.049	6.190	6.158	0.032
Dicofol*	3.037	2.987	0.050	5.600	5.211	0.389
Dieldrine*	1.663	1.667	0.005	4.481	4.597	-0.116

[2]- Residual



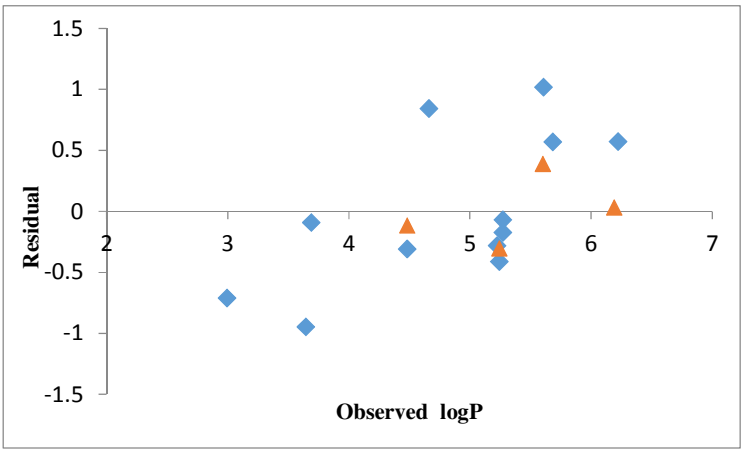
شکل -۱ مقادیر مشاهده شده بر حسب پیشگویی شده برای لگاریتم ضریب توزیع با روش GA-MLR.

Fig. (1). Comparison between predicted and observed values of logP by the GA-MLR method.



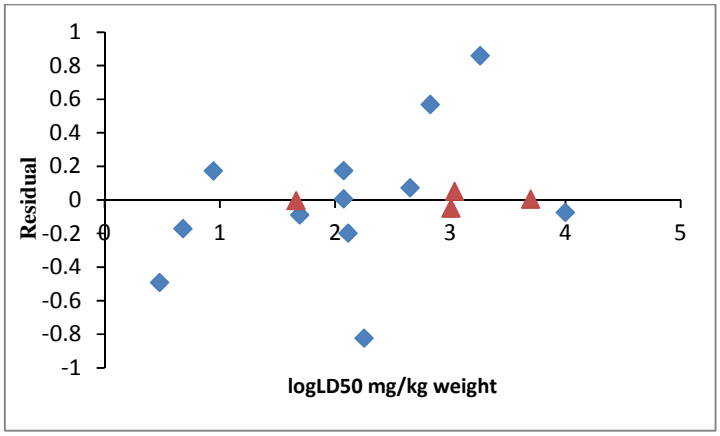
شکل ۲- مقادیر مشاهده شده بر حسب پیشگویی شده برای لگاریتم سمیت با روش GA-MLR.

Fig. (2). Comparison between predicted and observed values of logLD50 by the GA-MLR method.



شکل -۳ مقادیر مشاهده شده بر حسب باقی‌مانده برای لگاریتم ضریب توزیع با روش GA-MLR.

Fig. (3). Plot of the residual values for logP of compounds versus the experimental values by the GA-MLR method.



شکل ٤- مقادیر مشاهده شده بر حسب باقی‌مانده برای لگاریتم سمیت با روش GA-MLR.

Fig. (4). Plot of the residual values for logLD50 of compounds versus the experimental values by the GA-MLR method

توصیف گرهای منتخب

با توجه به نتایج جدول ٤ توصیف گرهای X1Av و PilD ازدو دسته توصیف گرهای اتصالی و مسیر از بقیه ۱۸ دسته توصیف گرهای محاسبه شده توسط نرم افزار دراگون جهت پیشگویی لگاریتم سمیت موثرتر می‌باشند. این دو توصیفگر جز توصیف گرهای توپولوژیکی محسوب می‌شوند. شاخص های اتصال مولکولی برای استخراج مدل ساده مربوط به غلظت بحرانی میسل های کاتیونی (کلرید) و همچنین جهت مدل‌سازی نقطه جوش هیدروکربن های بنزنوئید مورد استفاده قرار گرفته‌اند (Ramane & Yalnaik, 2017, Mozrzymas, 2017).

با توجه به نتایج جدول ٥ توصیف گرهای اطلاعاتی IC1 و TIC4 جهت پیشگویی لگاریتم ضریب توزیع اکتانول- آب انتخاب شدند. در سال‌های اخیر از توصیف گرهای اطلاعاتی به‌طور گسترده برای توصیف ساختارهای شیمیایی و همبستگی آن با خواص فیزیکی، شیمیایی و ساختاری استفاده شده است. این توصیف گرها در یک گراف براساس فاصله و با توجه به توانایی همبستگی و قدرت تفکیک پذیری آن‌ها در نظر گرفته می‌شوند (Konstantinova, 2006). درجدول ۸ توضیح مختصری در خصوص توصیف گرهای منتخب برای پیشگویی فعالیت‌های مورد نظر ارائه شده است.

جدول ۸-توصیفگرهای انتخاب شده در مدل های GA-MLR

Table 8. The Selected Set of descriptors presented in the GA-MLR Models.

Abbreviation	Description	Type of Descriptor
X1Av	average valence connectivity index of order 1	Connectivity indices
piID	conventional bond order ID number	Walk and path counts
TIC4	Total Information Content index (neighborhood symmetry of 4-order)	Information indices
IC1	Information Content index (neighborhood symmetry of 1-order)	Information indices

نتیجه گیری

مطالعات QSPR و QSAR ابزاری قوی جهت مطالعه و بررسی ارتباط ساختار- خاصیت و ساختار - فعالیت ترکیبات شیمیایی مهیا نموده است که به منظور پیش‌بینی خواص، فعالیت‌ها و طراحی ترکیبات جدید به کار می‌روند. در این مطالعه

با استفاده از روش رگرسیون خطی چند متغیره ارتباط کمی لگاریتم سمیت و ضریب توزیع اکتانول–آب با ساختار مولکولی ۱۶ سم از سموم آلی کلردار مورد ارزیابی قرارگرفت. تعداد زیادی توصیف کننده با استفاده از نرم افزار دراگون محاسبه گردید و با کمک روش‌های الگوریتم ژنتیک و برگشتی تعداد توصیف گرها تقلیل داده شد و با استفاده از نرم‌افزار SPSS و روش MLR مدل‌های اولیه برای فعالیت‌های مورد مطالعه به‌دست آمد. بعد از بررسی همبستگی بین توصیف گرها و بررسی اعتبار سنجی مدل‌های بدست آمده مشخص گردید ارتباط خطی مطلوبی بین ساختار سموم آلی کلردار و فعالیت‌های مورد مطالعه آن‌ها وجود دارد. برای پیشگویی لگاریتم سمیت، دو توصیفگر از گروه توصیف گرهای اتصالی و مسیر و برای پیش بینی لگاریتم ضریب توزیع اکتانول–آب دو توصیف گر از دسته توصیف گرهای اطلاعاتی از دیگر توصیف گرها از اهمیت زیادتری برخوردارند. با کمک این توصیف کننده‌های مولکولی و با استفاده از روش خطی می‌توان اقدام به مدل‌سازی و پیشگویی فعالیت‌های مورد نظر پرداخت و در وقت، هزینه صرفه جویی نموده و ضمن آسیب نرساندن به محیط زیست برای ترکیبات ناشناخته و یا شناخته شده فاقد اطلاعات مورد نیاز، به پیش بینی فعالیت‌های مورد نظر اقدام نمود.

Referance

Behfar, A., Nazari, Z., Rabiee, M. H., Raeesi, G., Oveisi, M. R., Sadeghi, N. and Jannat, B. 2013. The organochlorine pesticides residue levels in Karun river water. Jundishapur journal of natural pharmaceutical products, 8(1): 41-46.

Bradman, A. S. A., Schwartz, J. M., Fenster, L., Barr, D. B., Holland, N. T. and Eskenazi, B. 2007. Factors predicting organochlorine pesticide levels in pregnant Latina women living in a United States agricultural area. Journal of Exposure Science and Environmental Epidemiology, 17(4): 388-99.

Chatterjee, S. and Simonoff, J. 2013. Handbook of Regression Analysis. John Wiley & Sons: New York.

Cramer III, R. D., Bunce, J. D., Patterson, D. E. and Frank, I. E. 1988. Cross validation, Bootstrapping, and Partial Least Squares Comparedwith Multiple Regression in Conventional QSAR Studies. Molecular Informatics, 7: 18–25.

Craney, T. A. and Surles, J. G. 2002. Model-Dependent Variance Inflation Factor Cutoff Values. Quality Engineering, 3: 391-398.

Fitzgerald, E., Hwang, S. A., Deres, D. A., Bush, B., Cook, K. and Worswick, P. 2001. The association between local fish consumption and DDE, mirex, and HCB concentrations in the breast milk of Mohawk women at Akwesanse. Journal of exposure analysis and environmental epidemiology, 11: 381–388.

Ghosh, P. and Bagchi. M. C. 2009. QSAR modeling for quinoxaline derivatives using genetic algorithm and simulated annealing based feature selection. Current Medicinal Chemistrty, 16: 4032-4048.

Gramatica, P. 2007. Principles of QSAR models validation: Internal and external. QSAR & Combinatorial Science, 2007, 26(5), 694-701.

Goodarzi, M., Dejaegher, B. and Heyden, Y. V. 2012. Feature selection methods in QSAR studies. Journal of AOAC International, 95(3): 636-651.

Hagmar, L., Rylander, L., Dyremark, E., Klasson-Wehler, E. and Erfurth, E. M. 2001. Plasma concentration of persistent organochlorine in relation of thyrotropin and thyroid hormone levels in women. International Archives of Occupational and Environmental Health,74:184–188.

Han, S. Y., Qiao, J. Q., Zhang, Y. Y., Yang, L. L., Lian, H. Z., Ge, X. and Chen, H. Y. 2011. Determination of n-octanol/water partition coefficient for DDT-related compounds by RP-HPLC with a novel dual-point retention time correction. Chemosphere, 83(2): 131-136.

Jayaraj, R., Megha, P. and Sreedev, P. 2016. Organochlorine pesticides, their toxic effects on living organisms and their fate in the environment. Interdisciplinary toxicology, 9: 90-100.

Konstantinova, E. V. 2006. On some applications of information indices in chemical graph theory. In General Theory of Information Transfer and Combinatorics Springer, Berlin, Heidelberg. pp. 831-852.

Mozrzymas, A. 2017. Molecular connectivity indices for modeling the critical micelle concentration of cationic (chloride) Gemini surfactants. Colloid and polymer science, 295(1): 75-87.

Mwevura, H., Othman, O. C. and Mhehe, G. L. 2002. Organochlorine pesticide residues in sediments and biota from the coastal area of Dar es Salaam city, Tanzania. Marine Pollution Bulletin, 45(1-12): 262-267.

Pourbasheer, E., Ahmadpour, S., Zare-Dorabei, R. and Nekoei, M. 2017. Quantitative structure activity relationship study of p38α MAP kinase inhibitors. *Arabian* Journal of Chemistry,10: 33-43.

Ramane, H. S. and Yalnaik, A. S. 2017. Status connectivity indices of graphs and its applications to the boiling point of benzenoid hydrocarbons. Journal of Applied Mathematics and Computing, 55(1-2): 609-627.

Roy, K. and Mitra, I. 2011. On Various Metrics Used for Validation of Predictive QSAR Models with Applications in Virtual Screening and Focused Library Design. Combinatorial Chemistry & High Throughput Screen, 14: 450-474.

Saini, V. and Kumar, A. 2014. QSAR analyses of DDT analogues and their in silico validation using molecular docking study against voltage-gated sodium channel of Anopheles funestus, SAR and QSAR in Environmental Research, 25: 777-790,

Saghaie, L., Sakhi, H., Sabzyan, H., Shahlaei, M. and Shamshirian, D. 2013. Stepwise MLR and PCR QSAR study of the pharmaceutical activities of antimalarial3-hydroxypyridinone agents using B3LYP/6- 311++ G** descriptors. *Medicinal Chemistry Research*, 22: 1679-1688.

Saxena., A. K. and Prathipati. P. 2003. Comparison of MLR, PLS and GA-MLR in QSAR analysis. SAR and QSAR in Environmental Research., 14: 433-445.

Scheytt, T., Mersmann, P., Lindstädt, R. and Heberer, T. 2005. 1-Octanol/water partition coefficients of 5 pharmaceuticals from human medical care: carbamazepine, clofibric acid, diclofenac, ibuprofen, and propyphenazone. Water, air, and soil pollution, 165(1-4): 3-11.

Shargel, L., Susanna, W. and Yu, A. B. 2012. "Chapter 10: Physiological Drug Distribution and Protein Binding". Applied Biopharmaceutics & Pharmacokinetics. (secondary) (6th ed.). New York: McGraw-Hill Medical. p. 211.

Todeschini, R. and Consonni, V. 2000. Handbook of Molecular Descriptors. Weinheim: Wiley-VCH.

Todeschini, R. and Consonni, V. 2009. Molecular descriptors for chemoinformatics. Alphabetical listing (2nd ed., Vol. 1). Weinheim: Wiley-VCH.

Votano, J. R., Parham, M., Hall, L. H., Kier, L. B., Oloff, S., Torpsha, A., Xie, Q. and Tonga, W. 2004. Three new consensus QSAR models for the prediction of Ames genotoxicity. Mutagenesis, 19: 365- 377.

Web search engine developed by ChemAxon. 2019. software available at "http:// WWW. Chemicalize. Org".

Yehia, A. I. and Abdel-A, A. L. 1983. Quantitative Structure-Activity Relationship of DDT Analogs. Pesticide Biochemistry and Physiology, 20: 115-123.

Zaffar, H., Irshad, U., Pervez, A. and Alam Naqvi, T. 2016. Mode of Action, Toxicity and Biodegradation of Organchlorinated Pesticides: A Mini Overview. Journal of Applied Environmental and Biological Sciences, 6(8): 1-6.

Zhang, N., Yang, Y., Liu, Y. and Tao, S. 2009. Determination of octanol-air partition coefficients and supercooled liquid vapor pressures of organochlorine pesticides. Journal of Environmental Science and Health Part B, 44: 649-656.

Journal of Entomological Research

Volume 10, Issue 4, pages: 245-257

Islamic Azad University, Arak Branch
ISSN 2008-4668
http://jer.iau-arak.ac.ir

Prediction of toxicity and octanol – water partition coefficient of organochlorine pesticides using Molecular Descriptors and GA-MLR Method

Zahra Zamani[1], Fatemeh Shafiei[*2]*

1. PHD student, Department of Chemistry, Arak Branch, Islamic Azad University, P.O. BOX 38135-567, Arak, Iran
2. Assistant Professor, Department of Chemistry, Arak Branch, Islamic Azad University, P.O. BOX 38135-567, Arak, Iran

Abstract

In this research, Quantitative Structure–Activity Relationship (QSAR) studies have been used to predict activities of organochlorine pesticides. Firstly, the chemical structure of molecules was drawn with the Gauss view 05 program and optimized at Hartree–Fock level of theory and 6-31G* basis sets using Gaussian 09 software. The physiochemical properties namely octanol-water partition coefficient (logP) and toxicity (log LD_{50}) are taken from the scientific web book. The dragon software has been used for the calculation of molecular descriptors. The suitable descriptors were selected with the aid of the genetic algorithm (GA) and backward techniques. At the next step, the relationship between molecular descriptors and the activities was investigated by multiple linear regression (MLR) method. In order to build and test QSAR models, a data set of organochlorine pesticides was randomly separated into 2 groups: training (80%) and test (20%) sets.

The models were evaluated with regression parameters: correlation coefficient (R), squared regression coefficient (R^2), adjusted correlation coefficient (R^2_{adj}) and root mean squared error (RMSE).

For the predictive ability and verification of the models are discussed by using Leave-One-Out (LOO)

cross-validation and external test set. The external prediction accuracy of the obtained models was examined using the above regression parameters. Results of validations and high statistical quality of models indicate that generated GA-MLR models are reasonable QSAR models. These models help to delineate the important descriptors responsible for predicting their activities.

Keywords: "toxicity" **;** "QSAR" **;** "Multiple linear regression" **;** "genetic algorithm (GA)" **;** "organochlorine pesticides" **;** "octanol - water distribution coefficient"**.**

* Corresponding Author, E-mail*: shafa38@yahoo.com*
Received:7 Jul. 2019 – Accepted: 25 Nov.2019