
بسم الله الرحمن الرحيم

دانشگاه ولی عصر (عج) رفسنجان

دانشکده علوم پایه

گروه شیمی

پایان نامه کارشناسی ارشد رشته شیمی گرایش تجزیه

عنوان پایان نامه:

مطالعه ارتباط کمی بین ساختار و فعالیت ضد سالک یک سری از مشتقات هیدرازید با استفاده از روش‌های آنالیز چند متغیره

استاد راهنما :

دکتر زهرا گرکانی نژاد

دانشجو :

بهزاد احمدی رودی

مهر ماه ۸۸

Vali-e-Asr University of Rafsanjan

Faculty of Sciences

Department of Chemistry

M.Sc. Thesis

Title of the Thesis:

Quantitative structure –antileishmanial activity relationship study of some hydrazides by using multivariate analysis methods

Supervisor:
Dr. Zahra Garkani-Nejad

By:
Behzad Ahmadi-Roudi

September 2009

چکیده :

در این تحقیق روش QSAR برای مدل‌سازی و پیش‌بینی فعالیت ضد سالک یک سری ۴۹ تایی از مشتقات نیترو هیدرازید با استفاده از روش‌های مختلف کمومتریکس بکار برده شده است.

ابتدا تعداد زیادی از توصیف‌کننده‌های مولکولی با استفاده از نرم‌افزارهای هایپرکم، موپک و دراگون محاسبه شده است. سپس تعداد مناسبی از این توصیف‌کننده‌ها با استفاده از روش MLR انتخاب شده است. نتایج بدست آمده از روش MLR نشان می‌دهد که فعالیت ضد سالک این ترکیبات به پارامترهای مختلفی از قبیل: توصیف‌کننده‌های هندسی، خودهمبستگی دوبعدی، توصیف‌کننده‌های WHIMو GETAWAY بستگی دارد. سپس این توصیف‌کننده‌ها به عنوان ورودی برای شبکه‌های عصبی مختلفی از قبیل: ANN_LM، ANN_RP و ANN_GDX بکار برده شده است.

بهترین شبکه عصبی مصنوعی یک شبکه جلو رونده با انتشار به عقب خطا که با الگوریتم LM آموزش داده شده است می‌باشد. پس از بهینه‌سازی این شبکه دارای ساختار ۱–۵–۵ می‌باشد. کمترین خطای استاندارد و بهترین ضریب همبستگی برای مجموعه آموزشی به ترتیب ۰/۷۲۷ و ۰/۹۹۶ می‌باشد و برای مجموعه آزمایشی کمترین خطای استاندارد و ضریب همبستگی به ترتیب ۱/۵۸۹ و ۰/۹۵۱ می‌باشد. مقایسهٔ نتایج بدست آمده نشان می‌دهد که شبکه‌های عصبی مصنوعی قدرت پیش‌بینی بهتری نسبت به روش‌های MLR و PLS دارا می‌باشند.

عنوان **صفحه**

عنوان **صفحه**

عنوان **صفحه**

عنوان **صفحه**

فصل اول

۱- مقدمه

برای درک مکانیسم فرآیندهای مختلف شیمیایی، کشف و توسعه مواد جدید، حفظ محیط زیست و زمینه‌های دیگر شیمی هنوز توانایی حل مسائل به طور کامل وجـــود ندارد و برای عملی کردن بعضی از مسائل نیاز به سیستم‌های بسیار پیچیده‌ای می‌باشد که انجام آنها در گرو صرف هزینه‌های بسیار و مطالعات گسترده است. در جهت حل این مشکل، روش‌های محاسباتی کمومتریکس[1] می‌توانند مفید باشند. تجزیه و تحلیـــل آماری و ریاضی داده‌های شیمیایی معمولاً تحت عنوان کمومتریکس یاد می‌شود. به عبارتی کمومتریکس یک روش کارآمد برای خلاصه کردن اطلاعات مفید از یک سری داده مشخص و پیش‌بینی سری دیگر داده‌هاست. در حقیقت هدف کمومتریکس، بهبود بخشیدن فرآیندهای اندازه‌گیری و استخراج اطلاعات شیمیایی مفیدتر از داده‌های اندازه‌گیری شده فیزیکی و شیمیایی می‌باشد.

---

[1] chemometrics

کمومتریکس اولین بار در سال ۱۹۷۱ توسط سوانت ولد[1] شیمیدان سوئدی معرفی گردید. وی در سال ۱۹۷۴به اتفاق بروس کوالسکی[2] شیمی تجزیه‌دان آمریکائی جامعه بین‌المللی کمومتریکس را بنا نهاد [۳-۱].

چندین تعریف برای کمومتریکس بیان شده است که غالباً در متـــن‌های تجزیه‌ای به کار می‌روند. یکی از جامع ترین تعاریف به صورت زیر است:

کمومتریکس یک شاخه‌ای از شیمی است که از ریاضی و آمار و منطق استفاده می‌کند برای اینکه:

الف) فرآیندهای تجربی بهینه را طراحی و انتخاب کند،

ب) حداکثر اطلاعات شیمیایی قابل حصول را از تحلیل اطلاعات شیمیایی فراهم کند و

ج) بتوان اطلاعات بیشتری در مورد سیستم‌های شیمیایی بدست آورد[۴].

شیمیدانان با تعریف یک فرضیه، می‌توانند این فرضیه را آزمایش و اعتبار آن را ثابت نمایند. این کار نیاز به داده‌های تجربی دارد. بنابراین آنها ابتدا باید تصمیم بگیرند که چه آزمایش‌هایی را انجام دهند.

در کمومتریکس کارها به کمک ریاضی و تکنیک‌های آماری مانند استفاده از متدلوژی طراحی آزمایش انجام می‌شود. سپس با استفاده از داده‌های حاصل از آزمایش، اطلاعات را استخراج می‌کنند. برای مثال ساختن یک مدل بوسیله محاسبات رگرسیون که چگونگی ارتباط بین نتایج اندازه‌گیری و متغیرهای شیمیایی را توصیف می‌کند. یک شیمیدان می‌تواند با استفاده از این اطلاعات و دانش شیمیایی ، اطلاعات بیشتری در سیستم بدست آورد.

در واقع کمومتریکس مجموعه‌ای از یک سری قواعد شیمیایی است که با استفاده از ریاضی، آمار و کامپیوتر در عرصه‌های، ارزیابی و تفسیر اطلاعات، بهینه کردن و مدل‌سازی فرآیندها و آزمایشات و استخراج حداکثر اطلاعات شیمیایی از داده‌های تجربی به ما کمک می‌کند.

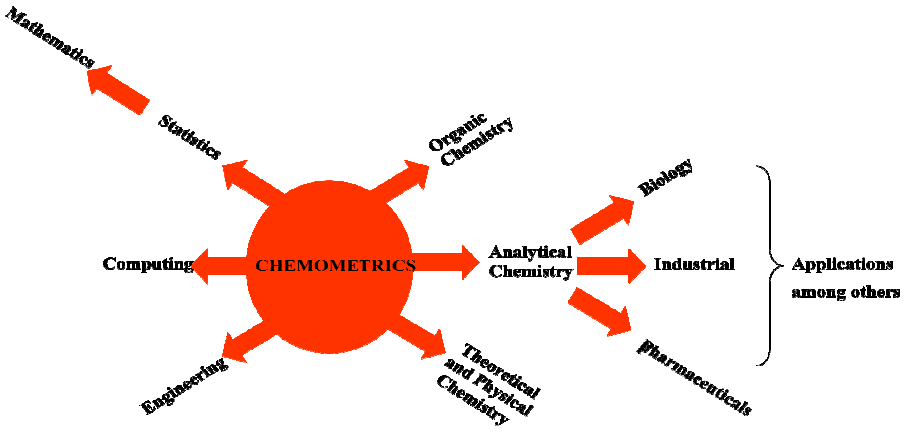
به طورکلی می‌توان گفت کمومتریکس، علم استفاده از کامپیوتر، ریاضی و آمار در شیمی می‌باشد. کمومتریکس به عنوان یک شاخه علمی جوان در دو دهه اخیر به سرعت توسعه پیدا کرده است. این رشد سریع مدیون پیشرفت دستگاه‌های هوشمند و خـــودکار آزمایشگاهـــــی و همچنین امکان استفاده از کامپیوترهای قدرتمند و نرم‌افزارهای ساده است. بنابراین کمومتریکس به عنوان یک وسیله در همه قسمت‌های شیمی و به طور زیادی در زمینه شیمی تجزیه مورد استفاده قرار گرفته است. امروزه یک شیمی‌دان تجزیه به صورت فزاینده‌ای به استفاده از روش‌های آماری و ریاضی در کارهای روزمره‌اش نیاز پیدا می‌کند[۵].

---

1 Svante Wold

2 Kowalski

امروزه کمومتریکس در شاخه‌های مختلف شیمی مورد استفاده قرار گرفته است که برخی از آنها عبارتند از: کنترل فرآیندها، تجزیه و تحلیل و شناخت الگوها، پردازش علائم، بهینه کردن شرایط. ارتباط کمومتریکس را با دیگر شاخه‌های علوم در شکل ۱-۱ نشان داده شده است.



شکل ۱-۱ ارتباط کمومتریکس با شاخه‌های علوم

یکی از زمینه‌های مهم کاربرد کمومتریکس در مطالعاتی است که خواص ملکول‌هـــــا را به ویژگی‌های ساختاری آنها نسبت می‌دهد. بررسی خواص بیولوژیکی فراورده‌های طبیعی و برخی از مشتقات آنها که خواص دارویی داشتند مقدمه‌ای جهت پیدا کردن وابستگی احتمالی ساختمان با فعالیت بیولوژیکی ارائه نمود. از نظر شیمی‌دانان فعالیت‌ها و خواص یک ترکیب ناشی از ویژگی‌های ساختاری آن است. بنا براین آگاهی از ساختار مولکولی کلید فهم عملکرد و خصوصیات مولکول‌ها است. این نوع از مطالعات به بررسی کمی ارتباط ساختمان با فعالیت، QSAR[1]، و همچنین بررسی کمی ارتباط ساختمان با ویژگی، QSPR[2]، معروف می‌باشد. هدف از مطالعات QSAR پیدا کردن رابطه‌ای است که بین رفتار فیزیک‌وشیمیایی یک ملکول با پارامترهای ساختاری آن وجود دارد[۶و۷].

نتایج این مطالعات علاوه بر شفاف سازی نحوه ارتباط بین خواص ملکول‌ها و ویژگی‌های ساختمانی آنها به پژوهشگران در پیش‌بینی رفتار ملکول‌های جدید براساس رفتار ملکول‌های مشابه کمک می‌کند. به مجموعه ابزارها و روش‌هایی که به این منظور مورد استفاده قرار می‌گیرند روش‌های پارامتری گویند. برخی روش‌های پارامتری در QSAR در شکل ۱-۲ نشان داده شده است

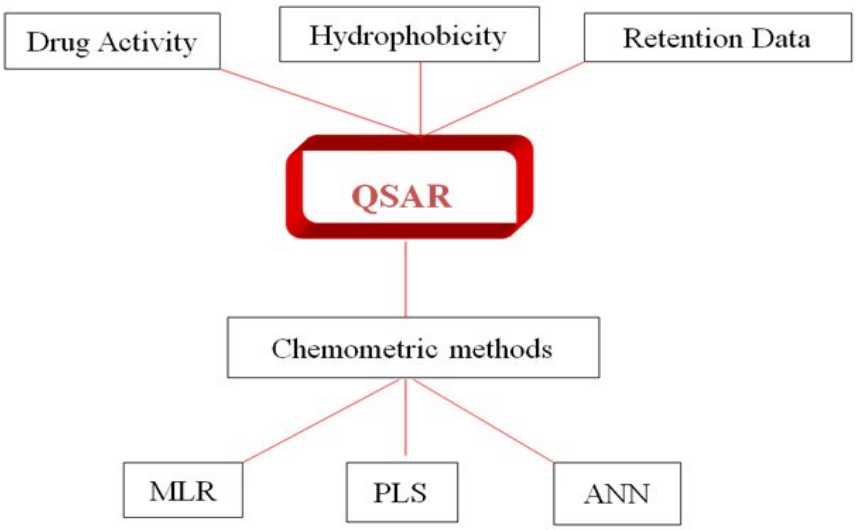
در روش‌های پارامتری سعی می‌شود بین یک سری توصیف‌کننده‌های ملکولی با فعالیت یا خاصیت مورد نظر ارتباط منطقی برقرار نمایند. توصیف‌کننده‌های ملکولی که به این منظور استفاده می‌شوند،

---

1 Quantitative Structure Activity Relationship

2 Quantitative Structure property Relationship

مقادیر عددی می‌باشند که جنبه‌های مختلف ساختاری ملکول را به طور کمی نشان می‌دهند. وقتی خصوصیات ساختاری گونه‌ها و فعالیت آنها توسط اعداد و ارقام بیان می‌شود می‌توان رابطه ریاضی یا کمی بین ساختار و فعالیت گونه ایجاد کرد. این رابطه می‌تواند برای پیش بینی پاسخ بیولوژیکی یا شیمیایی دیگر ساختارها مورد استفاده قرار گیرد.



شکل ۱-۲ روش‌های پارامتری در QSAR

برخی از روش‌های پارامتری کمومتری به اختصار به شرح زیر است:

۱- کالیبراسیون یک متغیره

۲- کالیبراسیون چند متغیره[۱]

۳- رگرسیون چند متغیره خطی[۲]

۴- حداقل مربعات کلاسیک[۳]

۵- حداقل مربعات معکوس[۴]

۶- رگرسیون اجزای اصلی[۵]

۷- حداقل مربعات جزئی[۶]

---

1 Multivariate Calibration
2 Multivariate Linear Regression
3 Classic Least Square
4 Inverse Least Square
5 Principal Component Regression
6 Partial_least_squares

۸- رگرسیون چند متغیره غیر خطی[۱]

۹- شبکه های عصبی مصنوعی[۲]

مراحل کالیبراسیون معمولا شامل طراحی آزمایش، انتخاب مدل، تخمین پارامترها و پیش‌بینی مجهولات می‌باشد. در کالیبراسیون یک متغیره، امکان تصحیح مزاحم‌ها بدون وارد کردن اطلاعات اضافی وجود ندارد، در حالیکه در روش‌های چند متغیره قادر به جدا کردن اطلاعات مفید از اطلاعات نامناسب مانند خطا می‌باشند، بدون این که اطلاعات اضافی دیگری برای این کار لازم می‌باشد.

مقاله‌های فراوانی در مورد استفاده از روش‌های پارامتری در مطالعات QSAR تا کنون منتشر شده اند. برخی از مقالات منتشر شده در این زمینه به قرار زیر است:

- پیش‌بینی اندیس بازداری یک سری از ترکیبات بنزن در کروماتوگرافی گازی توسط گرکانی نژاد و همکارانش [۸].

- مدل‌سازی تحرک الکتروفورتیکی[۳]، زمان بازداری و فاکتور پاسخ نسبی ترکیبات مختلف توسط گرکانی نژاد و همکارانش [۱۰و۹].

- پیش بینی ضرایب فعالیت ترکیبات در پارافین‌های با وزن مولکولی کم توسط اندراس[۴] و همکارانش [۱۱].

- مدل سازی فشار بخار تعداد بسیار زیادی از ترکیبات توسط لیانگ[۵] [۱۲].

استفاده از شبکه‌های عصبی مصنوعی به جای روش رگرسیون خطی چندتایی در مطالعات QSAR، به خصوص هنگامی که ارتباط بین توصیف کننده و فعالیت مورد نظر خطی نبوده و یا اینکه بین آنها بر هم کنش‌هایی وجود داشته باشد باعث بهبود مدل حاصله خواهد شد. در این پروژه این مطلب به وضوح نشان داده شده است و در آن سعی شده تا با استفاده ازسه روش شبکه‌های عصبی مصنوعی(ANN)، رگرسیون خطی چند تایی(MLR) و حداقل مربعات جزئی (PLS)، فعالیت یک‌سری ازمشتقات نیترو هیدرازید[۶] مدل‌سازی شود. در فصل‌های بعدی روش های QSAR و تکنیک شبکه‌های عصبی مصنوعی بطور کامل بحث شده است.

---

1 Multivariate Nonlinear Regression
2 Artificial Neural Network
3 Electrophoretic mobility
4 Andras
5 Liang
6 Nitro derivatives of Hydrazides

فصل دوم

۲ - مطالعه کمی ساختار ـ فعالیت

یکی از مهم‌ترین کاربردهای کمومتریکس ارتباط کمی ساختارـ فعالیت QSAR می‌باشد که به نحوه ارتباط بین فعالیت بیولوژیکی و ساختار شیمیایی ترکیبات می‌پردازد. هدف از QSAR، ایجاد رابطه ای منطقی بین کمیت‌ها و یا خواص ترکیبات (فعالیت) و ساختار شیمیایی آنها می‌باشد و این قانون برای مولکول‌های جدید مورد استفاده قرار می‌گیرد.

QSAR برای اولین بار در قرن نوزدهم مورد استفاده قرار گرفت.

مطالعات QSAR و QSPR به سه قسمت عمده تقسیم می‌شوند:

۱ ـ محاسبه و انتخاب توصیف‌کننده‌ها

۲ ـ مدل‌سازی و انتخاب بهترین مدل

۳ ـ ارزیابی اعتبار مدل‌های انتخاب شده

همان‌طور که می‌دانیم اساس QSAR مطالعه کمی بین ساختار و فعالیت است. برای رسیدن به این مهم باید فعالیت بیولوژیکی یا سایر کمیت‌هایی که بیانگر خاصیت ویژه‌ای از آن ترکیبات می‌باشند در شرایط آزمایشگاهی یکسان تعیین شده باشند تا بتوان آنها را از لحاظ کمی بررسی کرد. مجموعهٔ ترکیبات مورد مطالعه باید تا حدودی تشابهات ساختاری داشته باشند تا تعداد توصیف‌کننده‌های مورد نیاز برای ایجاد یک مدل مناسب، کم شود.

برای تحقق مرحلهٔ ۳، سری ترکیبات را به دو دسته سری مرجع (آموزشی) و سری پیش‌بینی یا به سه دسته سری مرجع (آموزشی)، سری پیش‌بینی و سری ارزیابی تقسیم می‌کنند. سری پیش‌بینی به گونه‌ای انتخاب می‌شود که نماینده کل ترکیبات موجود در سری آموزشی باشد. سری مرجع اکثریت

ترکیبات مورد نظر را در بر می‌گیرد و گروهی است که برای ایجاد مدل‌ها از آن استفاده می‌شود. سری پیش‌بینی شامل بقیه ترکیبات سری اولیه می‌باشد. از سری پیش‌بینی به منظور ارزیابی قدرت پیش‌بینی مدل‌های ایجاد شده استفاده می‌شود. در بعضی مواقع از روش‌هایی برای مدل‌سازی استفاده می‌شود که ممکن است سری پیش‌بینی را نیز به نحوی در مدل‌سازی درگیر کند. پس برای ارزیابی بهتر، از گروه سومی استفاده می‌شود که هیچ دخالتی در مدل‌سازی نداشته باشد.

برای محاسبه بعضی از توصیف‌کننده‌ها به مختصات سه بعدی اتم‌ها نیاز است. پس برای تحقق این مهم باید ساختمان ترکیبات بهینه شود. یعنی پایدارترین وضعیت ساختاری آنها با حداقل انرژی تعیین گردد.

۲ - ۱- محاسبه و انتخاب توصیف‌کننده‌ها

توصیف‌کننده‌های مولکولی:

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 مولکول به صورت یک عدد مفید می‌باشد[۱۳].

بکارگیری و تحلیل اطلاعات ساختار شیمیایی استفاده از توصیف‌کننده‌های مولکولی را ممکن ساخته است. توصیف‌کننده‌ها مقادیر عددی هستند که بیانگر ویژگی‌های مولکول می‌باشند. هر یک از این توصیف کننده‌ها اطلاعات خاصی از مولکول را در اختیار می‌گذارد.

برای مثال ممکن است توصیف‌کننده‌ها، ویژگی‌های فیزیکی و شیمیایی یک مولکول باشند و یا آنها ممکن است مقادیر ناشی از بکارگیری روش‌های الگوریتمی برای ساختارهای مولکولی باشند.

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. بطورکلی، افزایش محاسبات مورد نیاز در هر مرحله به واسطه اختلافاتی هست که ناشی می‌شود. برای مثال وزن مولکولی ویژگی‌های مولکولی زیادی را نشان نمی‌دهد ولی محاسبه آن خیلی سریع می‌باشد. در مقایسه توصیف‌کننده‌هایی که بر اساس مکانیک کوانتومی بدست می‌آیند خصوصیات دقیق زیادی ارائه می‌دهند، ولی زمان زیادی برای محاسبه مصرف می‌کنند.

۲-۱-۱- انواع توصیف‌کننده‌های مولکولی

توصیف‌کننده‌های مولکولی از نظر بیان چگونگی خصوصیات مولکول به دسته‌های متفاوتی تقسیم می‌شوند که برخی از آنها را به اختصار بیان می‌کنیم (توصیف‌کننده‌های مولکولی مربوط به نرم‌افزار دراگون) [۱۴].

۱- طبقه‌بندی بر اساس شکل مولکول‌ها

۲- طبقه‌بندی بر اساس نوع توصیف‌کننده‌ها

۲-۱-۱-۱- طبقه‌بندی بر اساس شکل مولکول‌ها

۱- توصیف‌کننده‌های مولکولی بدون بعد[1]
۲- توصیف‌کننده‌های مولکولی یک بعدی[2]
۳- توصیف‌کننده‌های مولکولی دو بعدی[3]
۴- توصیف‌کننده‌های مولکولی سه بعدی[4]

۲-۱-۱-۱-۱- توصیف‌کننده‌های مولکولی بدون بعد

توصیف‌کننده‌هایی هستند که مستقل از اتصال و صورتبندی مولکولی می‌باشند و می‌توان آنها را از روی فرمول مولکولی بدست آورد. تعداد اتم‌ها و تعداد پیوندهای موجود در یک مولکول از توصیف‌کننده‌های بدون بعد بشمار می‌روند [۱۳].

۲-۱-۱-۱-۲- توصیف‌کننده‌های مولکولی یک بعدی

توصیف‌کننده‌های مولکولی که بر اساس محاسبه تعداد گروه‌های عاملی شیمیایی بدست می‌آیند. از قبیل: گروه‌های عاملی[5]، توصیف‌کننده‌های تجربی[6] و خواص[7].

۲-۱-۱-۱-۳- توصیف‌کننده‌های مولکولی دو بعدی

این توصیف‌کننده‌ها از مهم‌ترین نوع توصیف‌کننده‌ها می‌باشند که شامل موارد زیر می‌باشد:
۱- توصیف‌کننده‌های توپولوژی[8]
۲- گام شمار مسیرهای مولکولی[9]
۳- شاخص‌های ارتباطی[10]
۴- شاخص‌های اطلاعاتی[11]
۵- خود همبستگی دو بعدی[12]
۶- شاخص‌های نزدیکی کناره[13]

---

[1] 0D-Descriptors
[2] 1D-Descriptors
[3] 2D-Descriptors
[4] 3D-Descriptors
[5] Functional Groups
[6] Empirical Descriptors
[7] properties
[8] Topological Descriptors
[9] Walk and Path Count
[10] Connectivity Indices
[11] Information Indices
[12] 2D-Autocorrelation
[13] Edge Adjacency Indices

۷- توصیف‌کننده‌های BCUT[1]
۸- شاخص‌های توپولوژی بار[2]
۹- شاخص‌های مبنی بر مقدار ویژه[3]

۲-۱-۱-۱-٤- توصیف‌کننده‌های مولکولی سه بعدی

شامل موارد زیر می باشد:
۱- توصیف‌کننده‌های هندسی[4]
۲- نیمرخ مولکولی رندیک[5]
۳- توصیف‌کننده‌های RDF[6]
۴- توصیف‌کننده‌های سه بعدی مورس[7]
۵- توصیف‌کننده‌های WHIM[8]
۶- توصیف‌کننده‌های GETAWAY[9]
۷- توصیف‌کننده‌های بار[10]
۸- شاخص‌های آروماتیکی[11]

۲-۱-۱-۲- انواع توصیف‌کننده مولکولی کلاسیک

۲-۱-۱-۲ -۱- توصیف‌کننده‌های توپولوژی

این توصیف‌کننده‌ها از روی گراف‌های مولکولی بدست می‌آیند. این توصیف‌کننده‌ها 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. این پارامتر را می‌توان بدون بهینه کردن ساختار مولکول محاسبه کرد.
از قبیل :
تعداد اتم‌ها[12]، شاخص‌های ارتباطی مولکولی[13] و وزن مولکولی[1].

---

[1] Edge Adjacency Indices
[2] Topological Charge Indices
[3] Eigen Value-Based Indices
[4] Geometrical Descriptors
[5] Randic Molecular Profile
[6] RDF Descriptors
[7] 3D MORSE Descriptors
[8] WHIM Descriptors
[9] GETAWAY Descriptors
[10] Charge Descriptors
[11] Aromaticity Indices
[12] Atom Counts
[13] Molecular Connectivity Indices

۲-۱-۱-۲-۲- توصیف‌کننده‌های هندسی

این توصیف‌کننده‌ها با ساختار سه بعدی مولکول‌ها در ارتباط می‌باشند. برای محاسبه این توصیف‌کننده‌ها ابتدا می‌بایست ساختار فضایی مولکول‌ها بهینه شود. ازقبیل:
حجم مولکولی، مساحت سطح[۲] و مساحت سطح در دسترس حلال[۳].

۲-۱-۱-۲-۳- توصیف‌کننده‌های شیمی کوانتومی[٤]

این توصیف‌کننده‌ها با استفاده از بهینه‌سازی نیمه تجربی ساختار مولکول‌ها در نرم‌افزارهای موپک[۵] و هایپرکم[۶] بدست می‌آیند.
از قبیل:
انرژی بالاترین تراز اشغال شد[۷]، انرژی پایین ترین تراز اشغال نشده[۸]، بار و الکترونگاتیویته اتم‌ها[۹].

۲-۱-۱-۲-٤- توصیف‌کننده‌های فیزیک‌وشیمیایی[۱۰]

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[۱۱]، نقطه ذوب و نقطه جوش.

۲-۱-۱-۲-۵- توصیف‌کننده‌های ارتباطی مولکولی

شاخص‌های ارتباطی مولکولی برخی اطلاعات را در موارد زیر فراهم می‌کند:
۱- اندازه و ساختار مولکول
۲- مرتبه شاخه‌دار شدن[۱۲]
۳- نحوه ارتباط اتم‌ها در مولکول
۴- نوع اتم‌ها در یک مولکول

---

[1] Molecular weight
[2] Surface Area
[3] Solvent Accessible Surface Area
[4] Quantum Chemical Descriptors
[5] Mopac software
[6] Hyperchem sofware
[7] Homo
[8] Lumo
[9] electronegativities
[10] Physico Chemical Descriptors
[11] Molecular Refractivity
[12] Degree of Branching

۲-۲- مدل‌سازی و انتخاب بهترین مدل

تا این مرحله انتخاب توصیف‌کننده‌ها براساس احتمالات و تشخیص نظری استوار است. بنابراین احتمال اینکه برخی از توصیف‌کننده‌ها برای مدل‌سازی مناسب نباشند وجود دارد و باید آنها را حذف کرد. توصیف‌کننده‌هایی که حذف می‌شوند معمولاً یک یا چند ویژگی زیر را دارند:

۱ ـ توصیف‌کننده‌هایی که کمتر از ۱۰ ٪ مقادیر غیر صفر دارند یا دارای بیش از ۹۰ ٪ مقادیر یکسان باشند.

۲ ـ توصیف‌کننده‌هایی که با توصیف‌کننده‌های دیگر همبستگی بالایی دارند ($r > ۰/۹$).

۳ ـ توصیف‌کننده‌هایی که با متغیر وابسته، همبستگی کمی دارند و محاسبه آنها مشکل است.

با توجه به موارد فوق، تعدادی از توصیف‌کننده‌ها حذف می‌شوند و از پیچیدگی محاسبات جلوگیری می‌شود. در این مرحله می‌توان با استفاده از روش‌های آماری مختلف به جستجوی مدل پرداخت. مدل، در واقع یک رابطه ریاضی است که بیان کننده ارتباط بین متغیر وابسته (فعالیت) و متغیرهای مستقل (ویژگی‌های مولکول یا توصیف‌کننده‌ها) می‌باشد. به کمک مدل می‌توان با داشتن مقادیر متغیرهای مستقل، متغیر وابسته را ارزیابی کرد. در این جا چون متغیر وابستهٔ ما با چندین متغیر مستقل تشکیل مدل می‌دهد، از رگرسیون خطی چند گانه (MLR) استفاده می‌کنیم. آنالیز رگرسیون خطی یا غیرخطی چندگانه بعنوان یک ابزار آماری جهت استخراج مدل‌های کمی، بررسی میزان اهمیت مدل‌های مذکور و اهمیت هر متغیر مستقل در معادله رگرسیون به کار می‌رود

۲-۲-۱- رگرسیون

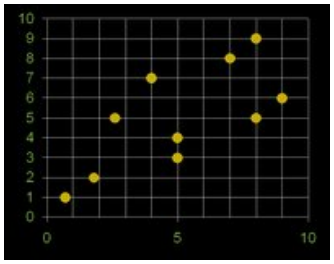
واژه رگرسیون در فرهنگ لغت به معنی بازگشت است و اغلب جهت رساندن مفهوم" بازگشت به یک مقدار متوسط یا میانگین " به کار می‌رود. بدین معنی که برخی پدیده‌ها به مرور زمان از نظر کمی به طرف یک مقدار متوسط میل می‌کنند.

بیش از ۱۰۰ سال پیش در سال ۱۸۷۷ فرانسیس گالتون[1] در مقاله‌ای که در همین زمینه منتشر کرد اظهار داشت که متوسط قد پسران دارای پدران قد بلند ، کمتر از قد پدرانشان می باشد . به نحو مشابه متوسط قد پسران دارای پدران کوتاه قد نیز بیشتر از قد پدرانشان گزارش شده است. به این ترتیب گالتون پدیده بازگشت به طرف میانگین را در داده‌هایش مورد تأکید قرار داد . برای گالتون رگرسیون مفهومی زیست شناختی داشت اما کارهای او توسط کارل پیرسون[2] برای مفاهیم آماری توسعه داده شده است. اگرچه گالتون برای تأکید بر پدیده "بازگشت به سمت مقدار متوسط" از تحلیل رگرسیون استفاده کرد، اما به هر حال امروزه واژه تحلیل رگرسیون جهت اشاره به مطالعات مربوط به روابط بین متغیرها به کار برده می‌شود [۱۵و۱۴].

[1] Francis Galton
[2] Karl Pearson

۲-۲-۱-۱- نمودار پراکندگی[1]

در حقیقت تحلیل رگرسیونی فن و تکنیکی آماری برای بررسی و مدل‌سازی ارتباط بین متغیرها است. رگرسیون تقریباً در هر زمینه‌ای از جمله مهندسی، فیزیک، اقتصاد، مدیریت، علوم زیستی، بیولوژی و علوم‌اجتماعی برای برآورد و پیش‌بینی مورد نیاز است. می‌توان گفت تحلیل رگرسیونی، پرکاربردترین روش در بین تکنیک‌های آماری است. شمایی کلی و خلاصه شده از یک تحلیل رگرسیونی ساده به صورت شکل۲-۱ می باشد [۱۵و۱۴].



شکل۲-۱. نمودار رگرسیونی

در ابتدا تحلیل‌گر حدس می‌زند که بین دو متغیر نوعی ارتباط وجود دارد ، در حقیقت حدس می‌زند که یک رابطه به شکل یک خط بین دو متغیر وجود دارد و سپس به جمع آوری اطلاعات کمی از دو متغیر می‌پردازد و این داده‌ها را به صورت نقاطی در یک نمودار دو بعدی رسم می‌کند. این نمودار که به آن نمودار پراکندگی گفته می‌شود نقش بسیار مهمی را در تحلیل‌های رگرسیونی و نمایش ارتباط بین متغیرها ایفا می کند(شکل شماره ۲-۱).

در صورتی که نمودار نشان دهنده این باشد که داده‌ها تقریباً (نه لزوماً دقیق) در امتداد یک خط مستقیم پراکنده شده‌اند، حدس تحلیل‌گر تأیید شده و این ارتباط خطی به صورت زیر نمایش داده می شود:

$$y = a\,x + b \qquad (۱-۲)$$

که در آن a عرض از مبدأ و b شیب این خط است.

۲-۲-۱-۲- متغیرها و خطا

بین برخی از نقاط و تصویر آنها بر روی خط رگرسیونی (خط y)کمی تفاوت به چشم می‌خورد که از آن به عنوان خطای برآورد یاد می‌کنیم. این خطا ممکن است از خطا در اندازه گیری، شرایط محیط، تفاوت‌های طبیعی و غیره ناشی شده باشد. بنابراین معادله اولیه را به صورت زیر اصلاح می‌کنیم :

$$y = ax + b + \epsilon \qquad (۲-۲)$$

[1] scatter plot