پردیس بین‌المللی ارس
گروه مهندسی کامپیوتر

پایان‌نامه
برای دریافت درجه کارشناسی‌ارشد در رشته مهندسی کامپیوتر گرایش نرم‌افزار

عنوان

بررسی مقایسه‌ای کاربرد روش‌های انتخاب پارامتر در مطالعات QSAR

استاد راهنما
دکتر محمدعلی بالافر

استاد مشاور
دکتر سمیه سلطانی

پژوهشگر
سیامک زنجانی

شهریور ۱۳۹۳


بسم الله الرحمن الرحيم

سپاسگزاری

در ابتدا از زحمات استاد راهنما آقای دکتر محمد علی بالافر و استاد مشاور خانم دکتر سمیه سلطانی تشکر می‌نمایم، همچنین از دوستان گرامی خانم‌ها محمدی‌وند، نورزاده و آقایان ذکریازاده و سعیدوند که مرا با راهنمایی‌های خودشان یاری کردند تشکر می‌کنم.

در پایان از زحمات پدر و مادر و خانواده خودم تشکر می‌کنم.

سیامک زنجانی

۱۳۹۳

نام خانوادگی دانشجو : زنجانی **نام** : سیامک

عنوان پایان‌نامه: بررسی مقایسه‌ای کاربرد روش‌های انتخاب پارامتر در مطالعات QSAR

استاد راهنما: آقای دکتر محمد علی بالافر
استاد مشاور: خانم دکتر سمیه سلطانی

مقطع تحصیلی: کارشناسی‌ارشد **رشته:** مهندسی کامپیوتر **گرایش:** نرم‌افزار **دانشگاه:** تبریز
دانشکده: پردیس بین‌المللی ارس **تاریخ فارغ‌التحصیلی:** ۱۳۹۳/۶/۲۰ **تعداد صفحه:** ۱۳۷

کلید واژه: انتخاب ویژگی، متلب، جعبه ابزار، استخراج داده، بررسی کمی رابطه‌ها ساختمان-فعالیت ترکیب‌ها

چکیده:

QSAR یا بررسی رابطه‌های کمی ساختمان و فعالیت ترکیب‌ها روشی است که مدل‌های ریاضی یا کامپیوتری را به منظور یافتن رابطه معنی‌دار آماری بین فعالیت و ساختار بکار می‌گیرد. یکی از مهمترین مرحله‌های انجام QSAR انتخاب ویژگی است. انتخاب ویژگی عبارت است از انتخاب یک زیر مجموعه از ویژگی‌های مجموعه داده بطوری‌که این ویژگی‌ها اطلاعات مفیدی در اختیار ما بگذارند. فایده‌های انتخاب ویژگی عبارت است از کاهش زمان محاسبه، حافظه مورد نیاز، کاهش زمان یادگیری، بهبود کارائی پیشگویی و حذف داده‌های مخدوش است. ما انواع روش‌های انتخاب ویژگی که شامل روش‌های خطی و غیرخطی است را بکار بردیم. برای انجام انتخاب ویژگی از جعبه‌ابزارهای مختلفی استفاده کرده‌ایم. برای مدل‌سازی از دو روش استفاده شده است: رگرسیون مرحله‌ای که یک روش خطی است و شبکه عصبی که روش غیرخطی محسوب می‌گردد.

به منظور انجام سریع و یکپارچه محاسبات 2D-QSAR، جعبه‌ابزاری را در متلب ایجاد کرده‌ایم. نام این جعبه ابزار FQSAR است. از مزیت‌های این جعبه‌ابزار افزایش سرعت و دقت در انجام محاسبه‌ها است. به راحتی می‌توان انواع روش‌های انتخاب ویژگی و مدل‌ها را به آن اضافه نمود. برای ارزیابی جعبه‌ابزار و روش‌های انتخاب ویژگی، نتیجه‌ی محاسبه‌ها بر روی سه تا مجموعه داده با استفاده از این جعبه ابزار مورد بررسی قرار گرفته است.

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

فصل اول - معرفی

QSAR یا بررسی رابطه‌های کمی ساختمان و فعالیت ترکیب‌ها روشی است که مدل‌های ریاضی یا کامپیوتری را به منظور یافتن رابطه معنی‌دار آماری بین فعالیت[1] و ساختار[2] بکار می‌گیرد. این روش برای پیش‌بینی فعالیت شیمیایی ترکیب‌های جدید بکار می رود.

۱-۱ انگیزه

امروزه بیش از ده‌ها میلیون ترکیب[3] در دنیا وجود دارد که هر کدام از آنها دارای خاصیت‌های مخصوص به خود است و این تعداد دائما در حال افزایش است. سه روش برای مطالعه فعالیت زیست‌شناسی ترکیب‌های شیمیایی وجود دارد : "in vivo" بکارگیری موجودهای زنده (مثل حیوان‌های آزمایشگاهی)،"in vitro" به کارگیری سلول‌ها، بیومولکول‌ها، ... و"in silico" که شامل به کارگیری روش‌های محاسبه‌ای برای پیش‌بینی فعالیت زیستی و یا خصوصیت‌های فیزیکی- شیمیایی ترکیب‌ها است.

روش‌های تجربی دارای عیب‌هایی هستند، از جمله اینکه هزینه و زمان زیادی برای اجرای آن‌ها نیاز است و همچنین غیرممکن بودن انجام برخی آزمایش‌ها بر روی انسان‌ها و یا حیوان‌ها محدودیت‌هایی را برای پژوهشگران ایجاد می‌کند. با توجه به مطلب‌های ذکر شده و نیز افزایش داده‌ها و اطلاعات در دسترس پژوهشگران، و در کنار همه این‌ها با افزایش قدرت پردازش کامپیوترها برای انجام محاسبه‌های پیچیده و حجیم، امروزه روش‌های "in silico" به عنوان ابزار کمکی و یا حتی جایگزین روش‌های تجربی برای پیش‌بینی فعالیت زیستی ترکیب‌ها مورد استقبال پژوهشگران قرار گرفته است. یکی از معمول‌ترین روش‌های مورد استفاده برای پیش‌بینی فعالیت ترکیب‌ها استفاده از رابطه‌های کمی ساختمان فعالیت (QSAR) است. QSAR به روش‌هایی اطلاق می‌شود که از یک رابطه کمی مبتنی بر مشخصه‌های ساختمانی ترکیب‌ها برای پیش‌بینی فعالیت آن‌ها استفاده می‌شود.

مزیت‌های روش QSAR به شرح زیر است:

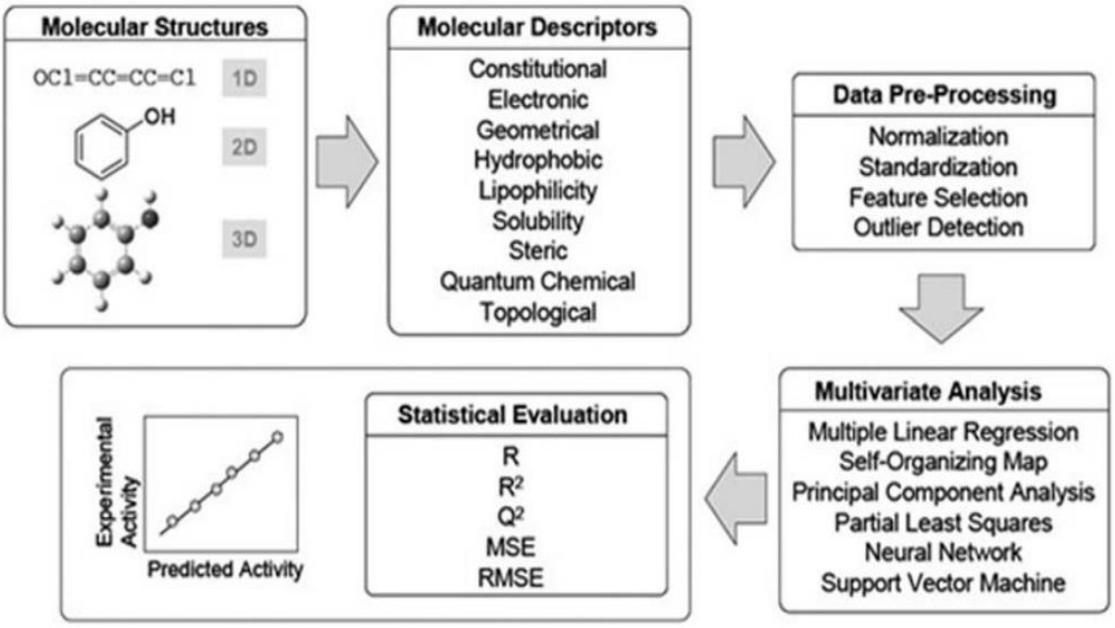
- امکان پیش‌بینی ویژگی‌های فیزیکی – شیمیایی و فعالیت زیستی مولکول‌ها.
- کاهش هزینه مراحل کشف و توسعه داروها.
- امکان پیش‌بینی اثرهای جانبی داروها قبل از تولید انبوه.

۱-۲ بیان مساله

1 Activity
2 Structure
3 Compounds

به طور کلی انجام QSAR را در شش مرحله کلی به شرح زیر می‌توان خلاصه کرد:

- تهیه پایگاه داده‌ها (گردآوری داده‌های مربوط به فعالیت زیستی یا خصوصیت‌های فیزیکی- شیمیایی)
- محاسبه و یا تعیین توصیفگرهای ساختمانی.
- آماده‌سازی داده‌ها[1] شامل : نرمال‌سازی[2] داده‌ها، استانداردسازی، انتخاب توصیفگرهای[3] مناسب و حذف داده‌های پرت
- ابداع رابطه کمی (مدل) مناسب بین توصیفگرها و فعالیت
- معتبرسازی مدل ارایه شده
- بازگشت به مرحله‌های قبلی و اعمال تغییرات مورد نیاز در صورت عدم موفقیت مدل



شکل ۱-۱: مرحله‌های کلی انجام QSAR

یکی از مهمترین مرحله‌های انجام QSAR مرحله انتخاب ویژگی است. این موفقیت منوط به انتخاب توصیفگرهای مناسب است که شامل حذف بعضی از توصیفگرهای نامرتبط (مخدوش و افزونگی) و یا انتخاب پارامترهای با قدرت تبیین بالاتر فعالیت مورد مطالعه است.

1Data Pre-Processing
2 Normalization
3 Descriptors

در این مطالعه روش‌های مختلف انتخاب توصیفگر از منابع استخراج شده و کاربرد آن‌ها در مطالعات QSAR مورد مطالعه قرار خواهد گرفت. بدین منظور جعبه‌ابزار مناسبی توسط نرم‌افزار متلب طراحی شده و مورد استفاده قرار خواهد گرفت. کارآیی روش‌های مختلف با استفاده از شاخص‌های معتبرسازی استخراج شده از دستورالعمل‌های بین‌المللی مورد بررسی قرار گرفته و مدل‌های مناسب ابداع خواهد شد.

۱-۳ اهداف

هدف اصلی این پایان‌نامه، مطالعه و مقایسه کاربرد انواع روش‌های انتخاب ویژگی در QSAR است. برای رسیدن به این هدف جعبه‌ابزاری[1] با ویژگی‌های ذیل طراحی خواهد شد.

- کلیه داده‌های خروجی را به صورت یک فایل اکسل در دسترس کاربر قرار دهد.
- یکپارچگی: نرم‌افزاری که با یک عملیات ساده بتوان کلیه مرحله‌های ابداع و معتبرسازی یک مدل QSAR را انجام داد.
- امکان افزودن الگوریتم‌های جدید به سادگی میسر است.

۱-۴ چهارچوب پایان‌نامه

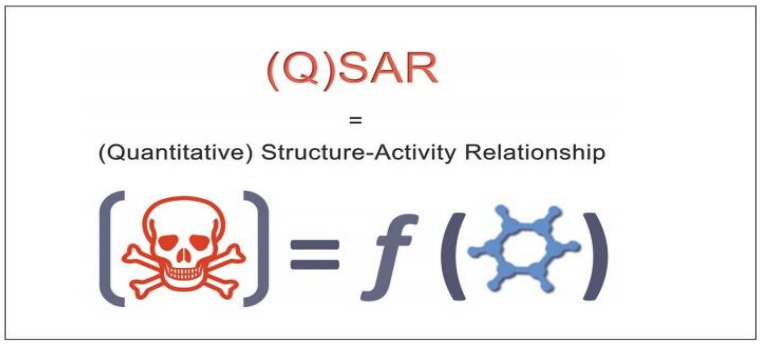
در فصل دوم ابتدا به معرفی انواع روش‌های QSAR پرداخته و سپس نرم‌افزارهای بکار رفته به طور مختصر توضیح داده می‌شود، در ادامه فصل دوم در مورد انتخاب ویژگی بحث و سپس روش‌های مختلف آن مورد بررسی قرار می‌گیرد، در فصل سوم روش بکار رفته در محاسبه‌ها توضیح داده خواهد شد، و سپس کدهای جعبه ابزار مورد تجزیه و تحلیل قرار می‌گیرد. در فصل چهارم نتیجه‌های بدست آمده برروی مجموعه داده‌ها بیان خواهد شد و در فصل پنجم پس از جمع‌بندی نتیجه‌ها، پیشنهادهایی برای مطالعه‌های آتی ارائه خواهد شد. در پیوست ب راهنمای اجرای برنامه آورده شده است.

1 Toolbox

فصل دوم - بررسی منابع

۲-۱ QSAR چیست؟

QSAR به رابطه‌های کمی بین فعالیت و یا خصوصیت‌های فیزیکی- شیمیایی یک ترکیب با خصوصیت‌های ساختمانی آن اطلاق می‌شود (شکل ۲-۱). روش‌های مدرن QSAR برای اولین بار در سال ۱۹۶۴ توسط فوجیتا[1]، هانس[2] [۱] ویلسون[3] ابداع گردید و پس از آن به سرعت گسترش یافت.



شکل ۲-۱: رابطه بین فعالیت و ساختار

۲-۱-۱ توصیفگرها

توصیفگرها در واقع مشخص کننده ویژگی‌های مولکول هستند، به دو دسته زیر تقسیم می‌گردند:

- توصیفگرهای دوبعدی (2D)

این نوع توصیفگرها در واقع توصیفگرهای توپولوژیکی گفته می‌شود، که نیاز به نمایش مسطح[4] مولکول دارند مانند تعداد اتم‌های مولکول.

- توصیفگرهای سه بعدی (3D)

گونه دیگری از توصیفگرها نیاز به نمایش سه بعدی[5] دارند که به آن‌ها توصیفگرهای سه بعدی گفته می‌شود، از قبیل توصیفگرهایی که مربوط به زاویه بین اتم‌های مولکول و غیره است. مقدار توصیفگرهای سه بعدی بسته به نوع نرم‌افزار، کاربر و تقریب‌هایی که بکار

1 Fujita
2 Hansch
3 Wilson
4 3D descriptors
5 Flat