Специализиран научен съвет
по електронна и компютърна техника при ВАК

инж. НИНА ГЕОРГИЕВА НИКОЛОВА

КОМПЮТЪРНИ МЕТОДИ ЗА МОДЕЛИРАНЕ НА ХИМИЧНИ СТРУКТУРИ

АВТОРЕФЕРАТ НА ДИСЕРТАЦИЯ
за присъждане на научната и образователна степен "Доктор"
научна специалност 02.21.04 "Компютърни системи, комплекси и мрежи"

Научни консултанти:

Чл.-кор. проф. д.т.н. инж. Кирил Боянов
проф. д.х.н. Ованес Мекенян

Научни рецензенти:

Ст.н.с. I ст. д.м.н. д.т.н. Красимир Атанасов
Ст.н.с. д-р Христо Турлаков

София, 2001

1. Обща характеристика на дисертационния труд

1.1. Актуалност

В последните години придобиват популярност методите за извличане на знания от големи масиви данни (*data mining*). Те успешно се използуват за компютърно откриване на общи характеристики и закономерности с цел получаване на информация от по-високо ниво (например правила, зависимости и др.), необходима за вземане на решения или при изследване, предсказване и моделиране на явленията, генерирали данните. Разработването на методи за анализ и извличане на зависимости от масиви данни с информация за химични съединения представлява интерес и от гледна точка на информатиката. Информацията за химични съединения се характеризира със следните особености: данните са с висока размерност (има изчислени и експериментално измерени стойности за множество параметри), данните са от различни типове, имат нестандартна структура (не всички вектори са с еднаква размерност) и са нехомогенни (съществуват различни зависимости между променливите в различни области на пространството). Поради тези особености директното прилагане на известните методи за извличане на знания обикновено не е подходящо.

Молекулният дизайн и молекулното моделиране са интердисциплинарни области, които имат за цел разработването на ефективни алгоритми за предсказване на биологичните свойства на химичните съединения, както и за създаването на нови съединения със зададени свойства. Използуват се методи и подходи на химията, информатиката, математическото моделиране, молекулярната биология и статистиката, както и на изкуствения интелект (разпознаване на образи, представяне на знания, лингвистични методи), бази данни, комбинаторика, стохастично моделиране, графи, паралелни изчисления.

Предсказването на физикохимични, биомедицински и токсикологични свойства на молекулите въз основа на параметри, които се изчисляват директно от структурата на химичното съединение е от голямо значение във фармацевтиката, химията и токсикологията (например при създаването на нови лекарства). При създаването на нови химични съединения, както и при изследването на на биологичните им свойства

е необходимо да се оценява терапевтичния ефект или токсичността на голям брой съединения, част от които все още не са синтезирани. Подобна е ситуацията и при оценката на риска от замърсяване на околната среда. Като пример може да се посочи, че повече от 15 милиона различни химични съединения са регистрирани в Chemical Abstract Service и този брой нараства с 775000 годишно. Всяка година започва използуването на около 1000 от тези съединения и само малка част от тях имат експериментално измерени свойства за оценка на риска от замърсяване на околната среда. В САЩ Toxic Substances Control Act Inventory съдържа около 74000 съединения и списъкът се увеличава с около 3000 годишно. Всяка година за предварителна оценка в Агенцията по опазване на околната среда на САЩ постъпват около 3000 съединения. За повече от 50% от тях няма експериментални данни, по-малко от 15% имат емпирични данни за мутагенност и само около 6% имат експериментални данни за екотоксикологичния им ефект и разграждането им в околната среда .

1.2. Цел и задачи на дисертационния труд

Основната цел на дисертационния труд е създаването на нови подходи към анализа и извличането на закономерности при обработката на масиви данни, и приложението им за предсказване на свойства на химични съединения и търсене в бази данни на химични съединения със зададени свойства, както и в молекулния дизайн.
Във връзка с тази основна цел, в дисертационната работа са поставени за решаване следните задачи:

- Разработване на метод за откриване и интерпретация на закономерности в масиви данни;
- Създаване на формален език за построяване на модели, описващи закономерностите в данните и прилагането му за предсказване и търсене в бази данни;
- Създаване на методи за генериране на нови обекти; при наложени моделни ограничения;
- Приложение на тези методи за молекулно моделиране.

1.3. Методи на изследването

Класификатор на Бейс
Оценка на плътността на вероятността с кернел функции
Построяване на дърво на решението
Генетични алгоритми
Формални езици
Методи за генериране на конформери
Методи за генериране на нови химични съединения

1.4. Кратка анотация на дисертационния труд

Глава 1 съдържа обзор на особеностите при съхраняване, обработка и анализ на информацията за химични съединения, както и на известни методи за разпознаване на образи, които представляват интерес при решаване на задачи за моделиране на свойствата на химичните съединения и молекулния дизайн.

Глава 2 съдържа описание на разработения метод за анализ и обобщение на масиви данни, съдържащи информация за химични съединения, чрез откриване на общи шаблони (метод *COREPA - Common Reactivity Pattern*). В резултат на анализа се построява дърво на решението, което представлява модел на разглеждано свойство. Дървото на решението се записва в синтаксиса на разработения език за описване на правила *RDL (Rule Description Language)*, което осигурява универсален метод за записване на правила, свързани със структурата на химичните съединения. Построеният модел може да се използува многократно за автоматично предсказване на моделираното свойство, търсене на молекули със зададени свойства или за генериране на нови структури със зададени свойства.

Глава 3 съдържа описание на разработения генетичен алгоритъм за намиране на множеството от максимално различни конформери (молекули с една и съща свързаност и състав, но различно разположение в пространството) - *Genetic Algorithm Search* , както и на разработения генетичен алгоритъм за компютърно конструиране на нови химични структури със предварително зададени свойства - *LeadGen*. Свойствата се задават чрез техния модел, записан във вид на правила на езика *RDL*. Генетичният алгоритъм на първо ниво конструира нови химични структури, които удовлетворяват зададените

условия. За всяка нова молекула се прилага втори генетичен алгоритъм за получаване на конформерите на молекулата.
Глава 4 съдържа кратко описание на програмните системи, в които са реализирани описаните алгоритми, както и резултати от приложението им при предсказването на естрогенна активност на химичните съединения в рамките на европейския проект *EDAEP (Endocrine Disrupting Ability of Environmental Pollutants*) за идентифициране на съединения с потенциален ефект върху ендокринната система.

2. Съдържание на дисертационния труд

2.1. Глава първа – методи за разпознаване на образи и използването им при моделиране на химични съединения

В глава първа са описани особеностите при моделиране на химични съединения и се анализират някои представителни методи за разпознаване на образи, които намират приложение при анализа на масиви данни с информация за химични съединения.

Представяне на химичните структури

Химичните съединения могат да бъдат зададени чрез структурната им формула, структурна диаграма или тримерен модел, както и чрез символен низ SMILES (*Simplified Molecular Input Line Specification*). Молекулата може да приема множество пространствени конфигурации ("*конформации*") под влияние на различни фактори, като промени в температурата, различни разтворители, ензими, взаимодействие с рецептор и други. Терминът "*конформер*" се използува за множеството конформации, съответстващи на локални енергийни минимуми. В литературата са известни детерминистични, стохастични и генетични алгоритми за генериране на конформери.
В голяма част от химичните бази данни съединенията се задават чрез информация за структурата им (*2D бази данни*). Появата на *3D бази данни* (съдържащи тримерни координати) е свързана с наличието на кристалографска информация за молекулите, както и с разработването на софтуер, позволяващ автоматично генериране на тримерни координати от таблицата на свързаност (CONCORD). Трябва да се отбележи, че понятието "бази данни", употребявано в литературата за моделиране на химични структури, не винаги съответствува на понятието "бази

данни", което се използува в компютърната литература. Информацията за химичните структури обикновено се съхранява в специфични файлови формати, които не са организирани чрез система за управление на бази данни (MOL, MOL2, SDF, PDB формати). Засега не съществува стандарт за файлове, съдържащи химични структури. Слабото използуване на стандартна технология като релационни БД може да се обясни със специфичния характер на софтуера за химично моделиране и факта, че средствата за SQL заявки не са достатъчни за формулиране на сложно търсене, свързано с намирането на фрагменти на химичните съединения (подграфи в молекулния граф) и взаимното им разположение. В изложението се използува понятието “бази данни”, както е прието в литературата за моделиране на химични структури.

Методи за моделиране

Биологичният ефект на химичните съединения е резултат от съвкупност от сложни събития, включващи движението на молекулата в организма, свързването и с рецептор в клетката и метаболитни процеси. Взаимодействието между молекулата и предполагаемия рецептор в общия случай не е известно, а експерименталните данни за биологична активност обикновено представляват измерената концентрация на съединението, необходима, за да се постигне съответния ефект. В този случай предсказването на биологичния ефект се свежда до намирането на зависимост между експериментално измерената активност и структура, свойства и параметри на молекулата, които могат да бъдат измерени и изчислени. Тези параметри се наричат молекулни дескриптори и могат да бъдат топологични, физикохимични, квантовохимични и други. Прилагат се ралични методи с цел да се намери множеството от подходящи молекулни характеристики, които имат връзка с биологичната активност и да се построи модел на най-вероятния механизъм на биологично действие. Механизмът обикновено се описва чрез *фармакофори* - функционални групи и фрагменти и/или техните електронни свойства и пространствени особености, отговорни за молекулните взаимодействия. След като моделът е построен, могат да се приложат *корелационни* методи за количествена оценка на активността в група съединения с общ механизъм на активност. Количественият анализ на връзката между химичната структура и активността (*Quantity Structure-Activity Relationship - QSAR*) се основава на хипотезата, че съединенията имат сходна биологична активност ако

притежават общи структурни свойства. Приложен за първи път от Hansh и сътрудници през 1960-те г., понастоящем *QSAR* е самостоятелна научна област, използуваща разнообразни статистически методи, дискриминантен анализ, невронни мрежи и други методи за разпознаване на образи, както и квантово химични методи за изчисляване на молекулните дескриптори.

Най-използуваните методи в QSAR анализа са многофакторна линейна регресия, редукция на информацията, метод на главните компоненти (Principal Component Analisys), Partial Latent Squares (PLS), клъстериране по молекулни дескриптори, невронни мрежи.

Широкото използуване на линейна регресия се основава на предположението на Hammet, че промените в свободната енергия на взаимодействието на даден клас структурно сходни съединения, с различни функционални групи като заместители, зависят линейно от промяната в свободната енергия на друг клас съединения, притежаващи същите заместители. Полученото уравнение може да се интерпретира като се обясни физическия смисъл на зависимостта. От друга страна, при получаване на регресионно уравнение с десетки и стотици параметри, физическият смисъл на зависимостта не е съвсем ясен.

Въпреки наличието на голям брой методи за клъстериране и дискриминантен анализ, прилагането им в молекулното моделиране среща редица проблеми, като:

- При отчитането на гъвкавостта на молекулите, всеки обект (молекула) се представя с множество подобекти (конформери), а не с единствен вектор, тъй като едно и също химично съединение (един и същи състав и свързаност) може да има представители с различно пространствено разположение. Това не позволява директното прилагане на известни алгоритми от изкуствения интелект. В литературата са известни детерминистични и стохастични методи за намирането на (под)множество от конформери. Откриването на всички възможни конформери е комбинаторен проблем. От друга страна, стохастичните методи създават проблеми при необходимостта за генериране на конформери при зададени ограничения.
- Необходимост от дефиниране на "подобие" между молекулите. Методите за разпознаване на образи и клъстеризация обикновено се основават на предварително дефинирано "разстояние" между обектите и класовете. При дефиниране на "разстояние" между

молекулите би трябвало да се отчитат освен голям брой скаларни и векторни параметри, така и топологичната и пространствена структура на химичните съединения. В литературата няма универсална дефиниция за сходството между молекулите. То се определя въз основа на експертна оценка и по различен начин за различни задачи. Задачата се усложнява още повече, ако се отчита гъвкавостта на химичните съединения.

- Клъстериране на набор от химични съединения според зададено сходство и извличане на критерии за предсказване на зададено свойство. В литературата са предложени голям брой методи, основани на пространствено разделяне или на идеи от теория на графите. Те се различават по алгоритмична сложност и приложимост в областта на молекулното моделиране. Не е очевидно какъв алгоритъм трябва да се избере за конкретна задача. Широко използваният алгоритъм на главните компоненти дава предимство на параметрите, притежаващи най-голяма дисперсия. За целите на клъстерирането е по-подходящо да се избират параметри, осигуряващи най-добро разделяне, а не тези с най-голяма дисперсия. Например, според публикувани изследвания, голяма част от данните, изследвани в химията имат вида на малък компактен клъстер, разположен във вътрешността на по-голям клъстер. Методи основаващи се на разстояния, както и на максимална дисперсия не могат да идентифицират клъстерите в данни с такава структура. Известни са също методи за класификация чрез нелинейно преобразуване на оригиналните данни в друго пространство с по-висока размерност, където обектите са линейно разделими (*Support Vector Machines, Self Organizing Maps, Radial Basis Networks, Statistical Learning Theory*), но те засега рядко се използват при анализа на химични съединения.

Вероятностният подход за класификация се основава на правилото на Бейс. То осигурява теоретичен оптимум за качеството на класификация, като практическите реализации могат само да се доближават до тази граница. За оценка на вероятностните плътности на класовете могат да се използват параметрични и непараметрични методи. Параметричните методи изискват предположение за вида на функцията на плътност на вероятността и оценка на параметрите на функцията. Поради това параметричните методи крият опасността от

допускане на грешка при предположението на вида на функцията и при оценка на параметрите. Ако за дадения проблем не съществуват основания за избор на конкретен вид на функцията, то се прилага непараметричен метод. Методите за непараметрична оценка на вероятностната плътност предоставят обосновани алгоритми за оценка на произволна плътност, като се избягва необходимостта от предположение за вида на функцията. Намерените оценки за плътност на вероятността могат да се използуват за класификация, непараметричен дискриминантен анализ, клъстериране и редукция на информацията.

Методите за построяване дърво на решението се ползуват с по-широка популярност, отколкото други непараметрични методи, поради своята нагледност (възможност за интерпретиране на класификацията) и ефективност на обучаващите алгоритми. Известни са алгоритмите *CART*, *ID3*, *C4.5.* Пример за реализация е системата Oncologic, която дава оценка за потенциалната канцерогенност на няколко групи химични съединения. Недостатъци: при работа с непрекъснати данни някои методи изискват предварително разделяне на параметрите на интервали; построеното дърво може да не е добро обобщение на данните.

Генетичните алгоритми са разработени по аналогия с естествения еволюционен процес. Генетичният алгоритъм е итеративна процедура, която се прилага върху популация от постоянен брой индивиди. Новите индивиди се генерират чрез *кръстосване* и *мутация*. Тези от тях, които ще образуват следващата генерация, се *избират* според стойността на зададена оценъчна функция. Генетичните алгоритми предоставят недетерминистично решение на комбинаторни оптимизационни проблеми. Приложенията им в конформационния анализ имат за цел да оптимизират генерирането на конформери според зададени геометрични/стерични условия.

Под названието *3D QSAR* са известни методи, които имат за цел да намерят количествената връзка между пространствената структура на химичното съединение и неговата активност. Обикновено се изчисляват стойностите на електростатични и други полета на молекулата и се търси корелация с биологичната активност. Като пример може да се посочи методът *CoMFA (Comparative Molecular Field Analisys)*. Тези методи обикновено работят с огромен брой

параметри, описващи повърхностите на молекулите и изискват предварителното им пространствено изравняване (наслагване).

Молекулен дизайн

Молекулният дизайн има за цел намирането на съществуващи или нови съединения с предварително зададени свойства. За идентифицирането на съединения със зададени свойства в съществуващи бази данни се използува получената в резултат на молекулно моделиране зависимост на свойството от известни молекулни дескриптори или се търси структура, която да съответствува най-добре на рецептора.. Голяма част от методите за молекулен дизайн се основават на концепцията на Емил Фишер от 1894 г. за взаимодействието на химичните съединения и рецептора като "ключ" и "ключалка" (пространствено съответствие между молекулата и специфичната част на рецептора, което позволява да се осъществи свързването им и да се прояви биологичното свойство). Генерирането на нови структури е известно под името *de novo* методи. Методите за генериране, които изискват наличието на информация за пространствената структура на съединението и рецептора се наричат *директни* и конструират структури, така че да съответствуват най-добре на рецептора. Тримерна информация за рецептора обаче често не е известна. От друга страна, оптималното съответствие не гарантира че съответната структура ще има максимална биологична активност и минимални странични ефекти. *Индиректните* методи са подходящи, когато структурата на рецептора е неизвестна, но са налице експериментални данни за набор от съединения, които проявяват разглежданата биологична активност. От тези данни чрез методи за молекулно моделиране се построява модел на биологичното свойство, който се използува за търсене или генериране на нови структури. Всяка генерирана структура подлежи на *оценка* доколко е подходяща. *Енергетичните функции* оценяват приноса на различните типове взаимодействие между структурата и рецептора, като кулонови, хидрофобни, стерични и други. Недостатък на тази оценка е, че се извършва бавно и зависи от качеството на избраната енергетична функция. *Правилата* обикновено определят вероятността за взаимодействие между новата структура и рецептора, като се извеждат въз основа на анализ на бази данни със структурна информация и честотата на срещане на различните типове контакти между фрагменти. Правилата обаче често са опростени и могат да

бъдат изведени само при наличието на достатъчно информация, а освен това този процес също е бавен.
Генерацията може да се извършва чрез систематичен или случаен метод. При *систематичния* подход се генерират всички възможни структури и след това се избира от тях по даден критерий. При *случайните* методи се извършват случайни модификации на структурата, оценяват се новите структури и се взима решение дали да бъдат приети или отхвърлени. Не всички методи генерират цели структури - някои методи имат за цел само намирането на подходящи позиции на атоми или фрагменти, които по-нататък могат да се използуват за създаване на цяла структура. Отчитането на конформационната гъвкавост на молекулата и рецептора е важен критерий за адекватността и полезността на метода.
Надеждността на метода е от голямо значение и е трудна за оценка. Новостта на методите за de novo дизайн често е свързана с нежелание да бъдат синтезирани новите структури само защото дадена структура е била генерирана от компютърна програма. Съществуват малък брой публикувани резултати на синтезирани химични съединения, предложени от de novo дизайн, което отчасти се дължи на факта, че изследванията се правят от фармацевтични компании и изследователите не публикуват най-интересните резултати и рецептори.

2.2. Глава втора – методи за моделиране и анализ на данни

Предложеният в дисертационната работа метод *Common Reactivity PAttern* за извличане на общи шаблони в набор данни от химични съединения, оценява вероятностното разпределение на данните и използува класификатор на Бейс за определяне критериите за принадлежност към даден клъстер. Това позволява извличане на значимите параметри и техните интервали, отчитане на гъвкавостта на съединенията, дава възможност за дефиниране на сходство между молекулите, без да е необходимо пространственото им наслагване.
Множеството от химични съединения се разделя на групи със сходно биологично действие. Намират се най-значимите параметри, т.е. тези, които осигуряват максимално различие между вероятностните плътности на групите. Използува се непараметричен метод за оценка на вероятностните плътности спрямо съответните параметри (*kernel*

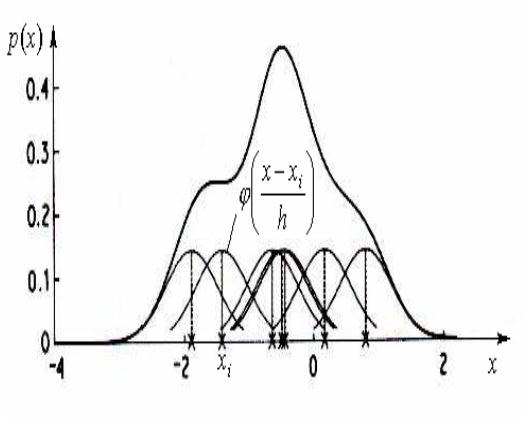
density estimation). Методът е модифициран, за да се вземе предвид конформационното разнообразие на химичните съединения (наличие на повече от един конформер за всяко химично съединение). За целта конформерите се отегловяват според статистиката на Болцман.

Вероятностната плътност на извадка (x_1, x_2,…, x_n) се оценява като $p(x) = \frac{1}{nh}\sum_{i=1}^{n}\varphi\left(\frac{x-x_i}{h}\right)$, където φ(x) е функция, удовлетворяваща условията:

$\int \varphi(x)dx = 1$

$\int x\varphi(x)dx = 0$

а h е изглаждащ параметър



Фигура 2.1

Оценката на вероятностната плътност има за цел намирането на функция, близка до вероятностната плътност, с която са били генерирани данните. Прилага се оценка с кернел функции, която за разлика от параметричния подход за оценка (намиране на параметрите на стандартно вероятностно разпределение, например нормално), не изисква избор на вида на разпределението. Изглаждащият параметър може да се определи чрез използуване на кръстосана проверка на достоверност или други известни алгоритми. Използуването на бързо преобразуване на Фурие позволява бързо изчисляване на оценката на плътността с кернел функции.

Алгоритъмът за оценка на плътността е модифициран поради необходимостта да се отчита повече от един конформер за съединение:

$$\alpha_{ij} = \frac{p(C_{ij})}{N_{ij}} = \frac{e^{-\Delta E_{ij}/k_B T}}{N_{ij}\sum_{m=1}^{R_i} e^{-\Delta E_{im}/k_B T}}$$

$$p(x|group_m) = \frac{1}{M_m}\sum_{i=1}^{M_m}\sum_{j=1}^{R_i}\sum_{k=1}^{N_{ij}}\frac{\alpha_{ij}}{h}\varphi\left(\frac{x-x_{ijk}}{h}\right)$$

където M_m е броя на съединенията в *група* m, R_i е броя на конформерите на *i*-то съединение, N_{ij} е броя на параметрите за *j-я* конформер на *i-то* съединение. C_{ij} обозначава *j-я* конформер на *i-то*

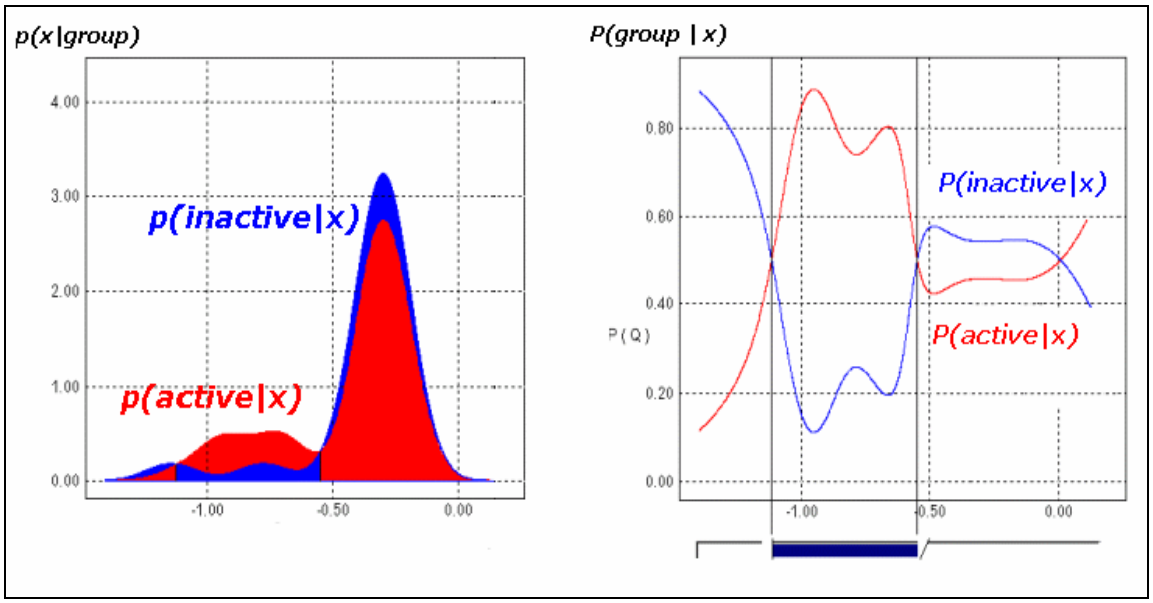
съединение S_i, а α_{ij} е отегловяващ коефициент за конформера, изчислен според статистиката на Болцман, k_B е константата на Болцман.

След като е получена оценката на вероятностната плътност, може да се приложи известната формула на Бейс:

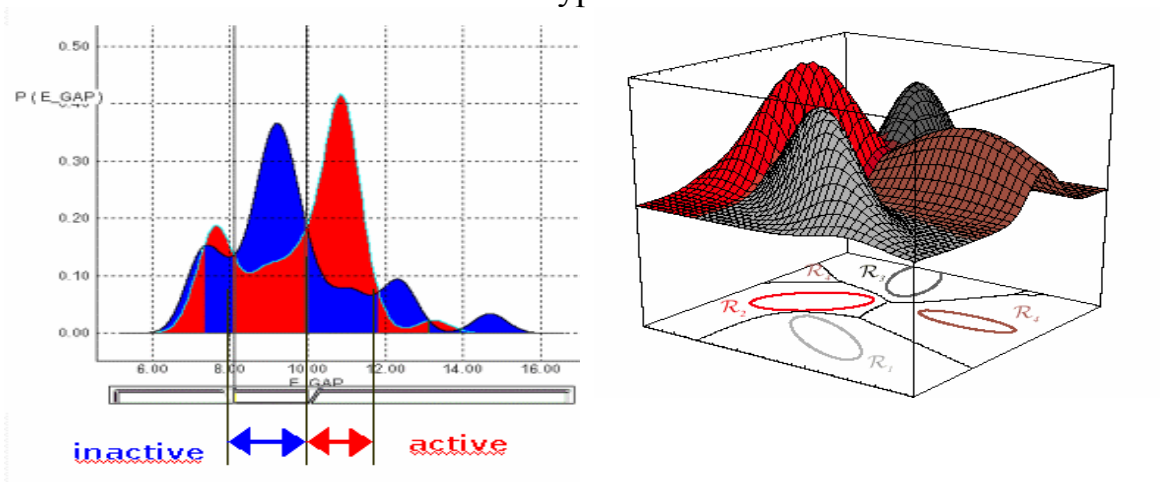
$$p(group_m|x) = \frac{p(group_m)\,p(x|group_m)}{\sum_{j=m}^{M_m} p(\omega_j)\,p(x|group_m)}$$

което позволява:

1. Да се вземе решение за принадлежността на дадено химично съединение към група. Съединенията се класифицират към *група m*, ако вероятността за принадлежност към групата е по-голяма от зададен праг *(0<threshold<1)*. Фигура 2.2 е илюстрация на случая с две групи (активни и неактивни съединения)
2. Да се извличат областите, за които може да се взема решение (където $p(group_m|x) > threshold$). В едномерния случай областите се свеждат до интервали (Фигура 2.3).
3. Да се оцени значимостта на параметрите чрез способността им да разделят зададените групи. В случай на две групи (активни и неактивни съединения), най-значимият параметър е този, за когото разстоянието между $p(group_{active}|x)$ и $p(groupin_{active}|x)$ е максимално. Известни са множество дефиниции за разстояния между вероятностни разпределения (*Kullback-Leibler divergence*, *Chernov, Bhattacharyya, Matusita, Hellinger, Mahalanobis, Patrick-Fisher distances*). В програмната реализацията COREPA се използува разстояние на Hellinger , което е дефинирано като: $D^2(p_1, p_2) = \int (\sqrt{p_{group1}(x)} + \sqrt{p_{group2}(x)})^2\,dx$
4. Да се изчисли степен на сходство между съединение и група, въз основа на оценените вероятностни плътности.
5. Да се изчисли степента на сходство между химични съединения, въз основа на оценените вероятностни плътности. По този начин се избягва необходимостта да се съпоставят тримерните структури на химичните съединения, като се дава и възможност за сравнение на гъвкави (с много конформери) съединения.



Фигура 2.2



Едномерен случай Двумерен случай

Фигура 2.3

Възможността да се намери най-значимия параметър и да се определят критериите за вземане на решение позволява построяването на **дърво на решението**, където на всяка стъпка като условие във възела на дървото се поставя намерения интервал за най-значимия параметър. Множеството данни се разделя според това условие и се търси следващият най-значим параметър за съответното подмножество данни.

Разработеният формален език *Rule Description Language* позволява задаване на множество различни правила за структурата на химичното съединение (свързаност, пространствено разположение, електронни условия). Функционалните групи се задават чрез стандартно SMILES означение. Разширение на SMILES позволява задаване на стереоинформация за фрагментите и молекулите. Могат да се дефинират и да се задават ограничения за разстояния между произволни фрагменти. Ограничения могат да се задават и на произволен параметър на структурата. Езикът позволява комбиниране на дефинираните условия в правила чрез използуване на логически оператори *and, or, not,* което осигурява гъвкавост и приложимост на подхода в множество различни ситуации. Дървото на решението се записва като съвкупност от правила. Правилата позволяват компютърна реализация и могат да бъдат достатъчно подробни, за да бъдат използувани за предсказване и търсене. Предложеният подход дава общо решение на проблема за задаване на сложни параметри, свързани с молекулната структура, тъй като може да се комбинира подфрагментно търсене с молекулни дескриптори, стереохимични конфигурации и взаимно 3D разположение на функционалните групи.

2.3. Глава трета – Методи за генериране на нови обекти

Обектите, които представляват интерес в дисертационната работа са химични съединения, зададени с тяхната структура и физикохимичните им свойства. Биологичната активност е свойство, дефинирано в зависимост от изследвания проблем, например токсичност, мутагенност или лекарствено действие на съединенията. Задачата се състои в това да се "разпознаят" биологично активните съединения и да се предложат методи, които позволяват целенасочено "конструиране" на нови биологично активни съединения. Тъй като биологичният ефект може да се предизвиква от един или повече конформери на молекулата, които не за задължително най-ниско енергетичните, то е важно да се намери множеството от най-различаващите се конформери.

В дисертационната работа е предложен и реализиран генетичен алгоритъм за намиране на множество конформери най-добре покриващо конформационното пространство. Алгоритъмът дава

възможност за генериране на конформери при наложени ограничения върху структурата на химичните съединения. Тези ограничения се описват чрез създадения формален език *RDL*.

Генетичен алгоритъм за намиране на множество конформери, които най-добре покриват конформационното пространство (Genetic Algorithm Search)

Предложеният генетичен алгоритъм намира малък брой конформери с ниска енергия, максимално покриващи конформационното пространство. Вместо изчерпателно търсене на всички конформери се прилага генетичен алгоритъм, който максимизира структурните разлики между генерираните конформери. По този начин става възможно да се изследва конформационното пространство дори и за големи гъвкави молекули. Приликата между конформерите се оценява като реципрочна на стойността на *RMSE (Root Mean Square Error*) разстоянието между идентични атоми при наслагване на структурите, което обезпечава неговия минимум. RMSE за два обекта X и Y в тримерно евклидово пространство се дефинира като:

$$R(X,Y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{3}(x_{ij} - y_{ij})^2}$$

,

където N е броят на точките на X и Y.

За разлика от класическия генетичен алгоритъм, тук не се оценява годността на индивидуалния конформер, а годността на популацията, към която той принадлежи.

Методът използува следните преобразувания за генериране на конформери: ротация около ациклична връзка и двойна връзка, инверсия на стереоцентровете, обръщане на свободен ъгъл в наситени цикли, отражение на пирамиди в 2 или 3 наситени цикъла. Последните две са въведени с цел да се обхване структурното разнообразие при полициклични структури. Стереохимичните изменения могат да бъдат забранявани, ако трябва да се запази стереохимията на изходната структура. Изследвани са три качествени критерия за оценка на изпълнението на алгоритъма: стабилност, възпроизводимост и покритие на конформационното пространство.

Генетичен алгоритъм за генериране на нови химични съединения при зададени ограничения (Lead Generator - LEADGEN)

Компютърният дизайн на нови лекарства и други химични съединения с предварително зададени свойства включва следните задачи: оптимизация и генериране на лидери. Оптимизацията на съединението се състои в стъпково (постепенно) модифициране на известни съединения, което ограничава намерените структури до близки на изходната структура. Създаването на лидери (*lead compounds*) може да бъде извършено чрез търсене в бази данни на структури, които удовлетворяват зададени структурни изисквания за активност, или чрез генериране нови структури при тези условия. В молекулния дизайн критериите за активност могат да се основават на ограничения за разстояния между атоми и фрагменти, разпределение на заряди и др.
За създаване на нови структури е разработен генетичен алгоритъм на две нива, който използува сглобяване на фрагменти и оценка на новите структури чрез правила и енергетични критерии. В класическия генетичен алгоритъм обектите се кодират, като се получават гени, върху които се прилагат генетичните операции. В настоящия алгоритъм обектите се представят с естествения тримерен модел на молекула (свързаност на атомите и 3D координати). Фрагментите представляват тримерни структури и/или части от структури. Първоначалната популация от структури се генерира чрез случайно свързване на фрагменти от зададения речник. Създаването на нова структура се извършва чрез последователно свързване на произволни фрагменти, до достигане на предварително зададена стойност за минималния брой на атоми в молекулата или минимална молекулна маса. Следващите популации се получават чрез прилагане на генетичните операцията *кръстосване* и *мутация*. *Кръстосване* се извършва на два етапа: първо, разкъсване по ациклична връзка на всяка от двете родителски структури и второ, съединяване на получените фрагменти (по един фрагмент от всяка родителска структура формира новата структура). Фрагментите се свързват по ациклични връзки с еднакъв тип, което позволява конструирането на структури без явното въвеждане на правила за валентност. След свързването на два фрагмента се извършва локална геометрична оптимизация като се максимизират междумолекулните разстояния между двата фрагмента, чрез въртене около простата или тройна връзка. Ако връзката е двойна,

то фрагментите се разполагат в планарна конфигурация. В допълнение е въведено и свързване чрез слепване (кондензиране) на цикли по обща връзка, така че да се получат полициклични съединения. В този случай са въведени правила за спазване на валентностите. Операцията *мутация* представлява замяна на един атом с друг, като предварително е зададена таблица с възможните замени. Операцията *селекция(избор)* се извършва въз основа на оценката за всяка нова структура. Структурите с по-добра оценка заместват тези с по-лоша оценка от родителската популация. Оценката е комплексна и съдържа множество критерии – молекулна маса, енергия, ранг (който се получава чрез прилагане на съвкупност от *RDL* правила). Правилата позволяват задаване на критерии, специфични за дадения проблем – търсене на съединения с определени свойства.

За всяка новосъздадена структура се прилага алгоритъма за генериране множеството от максимално различни конформери.

2.4. Глава четвърта - Приложение на разработените методи за моделиране и изследване на химични структури

Програмни системи

Описаните в глави 2 и 3 методи са реализирани в съвкупност от приложни програми за MSWindows 95, 98, NT и по-нови версии, с използуване на обектен Pascal и средата за разработка на Borland International - Delphi 2,3,4,5. Настоящите версии се компилират с Borland Delphi 5. Всички програмни системи използуват общи модули за визуализация на химичните бази данни, химичните структури (двумерни и тримерни), изчисляване на параметри, въвеждане на нови структури. За визуализация на тримерните модели се използува стандартната OpenGL библиотека. Химичните съединения в разработените програмни системи се съхраняват в специфичен файлов формат (*.cmp*). В настоящите версии не се използува система за управление на бази данни. Този формат е избран поради необходимостта от съвместимост с по-стари версии на софтуера, разработван в Лабораторията по Математична Химия при Университет "проф. А.Златаров" - Бургас. *CMP* файловете се създават от приложния софтуер, като химичната информация може да се въведе чрез SMILES, текстови файлове в SDF/MOL/MOL2/XYZ формати, интерактивно

задаване на структурите чрез двумерен или тримерен редактор, реализирани в програмната система. Необходимите молекулни дескриптори (геометрични, електронни, физикохимични и други) се изчисляват автоматично.
Програмните системи позволяват:
COREPA - реализация на *Common Reactivity PAttern* метода за построяване на модел на биологичната активност чрез дърво на решението:

- Избор на параметри, както и дефиниране на сложни и обобщени параметри, зависещи от молекулната структура чрез интерактивен двумерен редактор на химични структури;
- Визуализиране на молекулните и атомни дескриптори, на които се дължи реакционния образ;
- Оценка на изведения модел според различни критерии;
- Определяне на общите интервали за конформерите на всички съединения;
- Визуализиране полученото дърво на решението и записването му като текстов файл с правила на езика RDL;
- Работа върху подмножество от базата данни от химични съединения, избрано от експерт-химик по различни условия, включително подфрагментно търсене.

FORECAST - програмна система, позволяваща количествено и качествено прогнозиране на биологични свойства при готови файлове с правила на езика RDL:

- Дефиниране на свойство за предсказване;
- Прилагане на предварително дефинирани правила с цел разделяне на базата данни от съединения на групи с общ механизъм на действие;
- Извеждане на регресионни модели за всяка група, въз основа на обучаваща база данни с химични съединения;
- Изчисляване на стойностите на разглежданото свойство въз основа на изведените модели;
- Визуализация на причините, поради което едно химично съединение е класифицирано към дадена група (атоми, атомни характеристики, фрагменти, разстояния между фрагменти, молекулни дескриптори);
- Конформационно размножаване на съединенията при желание от потребителя, чрез разработения генетичен алгоритъм;

- Визуализация на разпределението на конформерите по енергия;
- Изследване на конформационната гъвкавост на съединенията в процеса на скрининг, чрез прилагане на алгоритъма tweak към единични конформери или чрез тяхното конформационно размножаване чрез алгоритъма GAS;
- Прогнозиране на свойствата на химичните съединения:
  1. Класификация на химичните съединения, чрез прилагане на предварително построено дърво на решение.
  2. Количествено прогнозиране. След класификация за всяка група се извежда регресионен модел, който дава количествена оценка на активността на съединенията с общ механизъм на действие.

LEADGEN - програмна система, реализираща описания алгоритъм за генериране на нови съединения с предварително зададени свойства. Системата позволява:

- Задаване на условията, които трябва да се удовлетворяват от новите химични съединения чрез RDL файл;
- Използуване на съединенията от *.cmp* файл като множество от фрагменти, от които се конструират новите съединения;
- Създаване на библиотека от прости фрагменти.

GAS32 - програмна система, реализираща описания алгоритъм за намиране на множеството от максимално различни конформери. Този алгоритъм е интегриран и във всички останали приложни програми, с цел прозрачно генериране на конформери при решение на съответната задача. GAS32 дава възможност и за изследване на възпроизводимостта и стабилността на алгоритъма при многократно изпълнение върху едни и същи структури и при различни настройки.

Скрининг на бази данни от химични съединения за оценка на биологичната им активност

Разработените в дисертацията методи са използувани в няколко проекта на Лабораторията по математична химия при Университет "А.Златаров" - Бургас, част от които са:

- Скрининг на базите данни с химични съединения на Европейската общност (*EINECS - European Inventory of Existing Commercial*

Substances) и Американската Агенция за опазване на околната среда (*TSCAI - Toxic Substances Control Act Inventory*) за идентифициране на потенциално токсични съединения;

- Прогнозиране на биоразградивността на индустриални замърсители;
- Скрининг на базите на Европейската общност за идентифициране на съединения с потенциален ефект върху ендокринната система (проект EDAEP). Резултатите от този проект са представени по-долу като илюстрация за използуването на разработените методи.

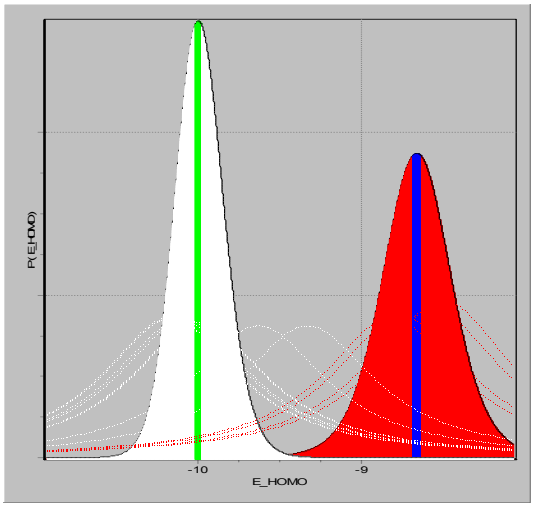
Проектът EDAEP има за задача прогнозиране на способността на съществуващи съединения да предизвикват разпадане на ендокринната система. Чрез метода COREPA е построено дърво на решението, което позволява предсказването на естрогенната активност на съединенията. Експерименталните данни за естрогенната активност на дадено съединение се получават като радиоактивно белязан естествен хормон *естрадиол* се свързва с естрогенния рецептор в препарат, получен например от черен дроб на риби. Добавя се изследваното химично съединение в определена концентрация, и се измерва каква част от естествения хормон е изместена от рецептора. Предполага се, че изследваното химично съединения се свързва с естрогенния рецептор, като измества естествения хормон. Количеството на изместения (свободен, несвързан с рецептора) *естрадиол* се използува като числена оценка на относителния афинитет на свързването (*relative binding affinity - RBA*) на изследваното съединение към естрогенния рецептор по отношение на естествения хормон.

За извеждане на модела се използува множество от съединения с известна експериментална оценка на естрогенната активност.

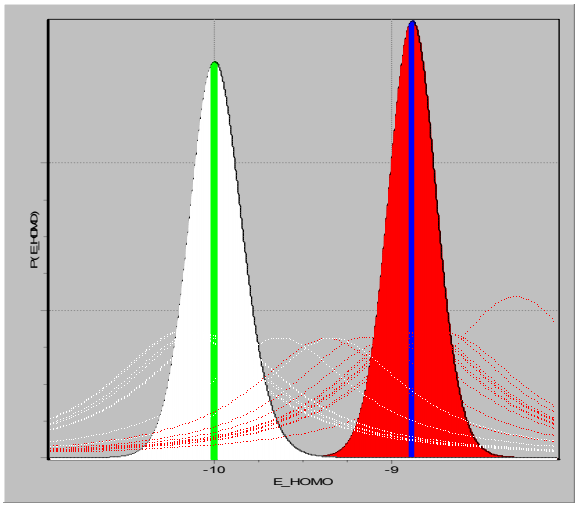
Базите данни с химични съединения, които подлежат на прогнозиране са **HVPC** - *High Production Volume Chemicals* - съединения с висок обем на производтсво и **LVPC** - *Low Production Volume Chemicals* - съединения с нисък обем на производство, **TSCAI (***Toxic Substances Control Act Inventory*), **EINECS (***European Inventory of Existing Commercial Substances).* Оригиналните бази данни съдържат само структурна информация за съединенията (SMILES означение или химично наименование), което изисква генериране на тримерна информация за съединенията, както и изчисляването и въвеждането на множество физикохимични параметри, които ще се използуват за

моделиране на биологичната активност. Тези дейности са извършени с помощта на разработените програмни системи.

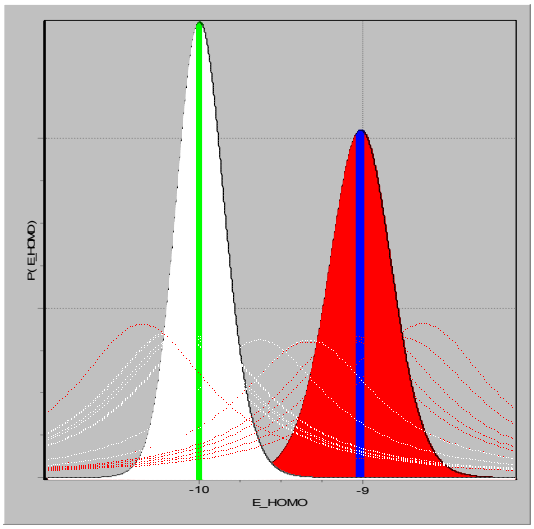
Чрез прилагането на COREPA метода са определени молекулните дескриптори, осигуряващи най-голямо различие между оценените плътности на вероятността за активните и неактивни съединения. Това са: глобалната нуклеофилност, E_{HOMO} ; разстоянието между нуклеофилни центрове и заряда (локалната нуклеофилност) на тези центрове. На Фигура 2.4 е показана илюстрация на оценените плътности на вероятността спрямо параметъра E_{HOMO}., съответно за активните (с червено) и неактивни (с бяло) съединения. Образът на съединенията с активност *RBA>100% ; 10% < RBA<100%; 1% < RBA<10% и 0.1% < RBA<1%* е сравнен с този на неактивните съединения (*RBA<0.1%*). От фигурите се вижда, че с нарастване на биологичното сходство между тестовите подмножества от Фигура 2.4а до Фигура 2.4.г, нараства и сходството между вероятностните плътности за активните и неактивни съединения.



а)Активни - RBA>100%
Неактивни - RBA<0.1%

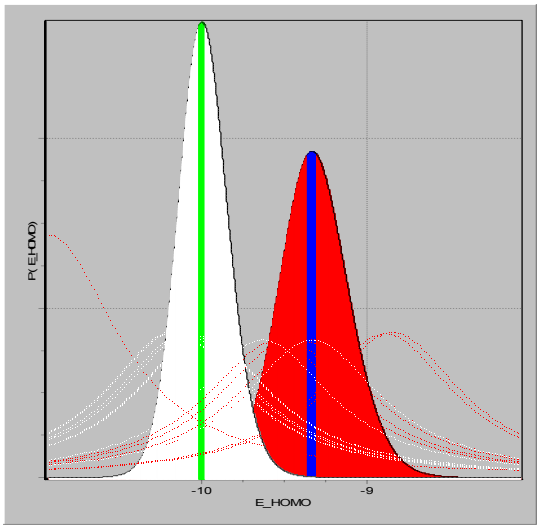


Б)Активни – 10%<RBA<100%
Неактивни - RBA<0.1%



в)Активни – 1%<RBA<10%

Неактивни - RBA<0.1%



г)Активни – 0.1%<RBA<1%

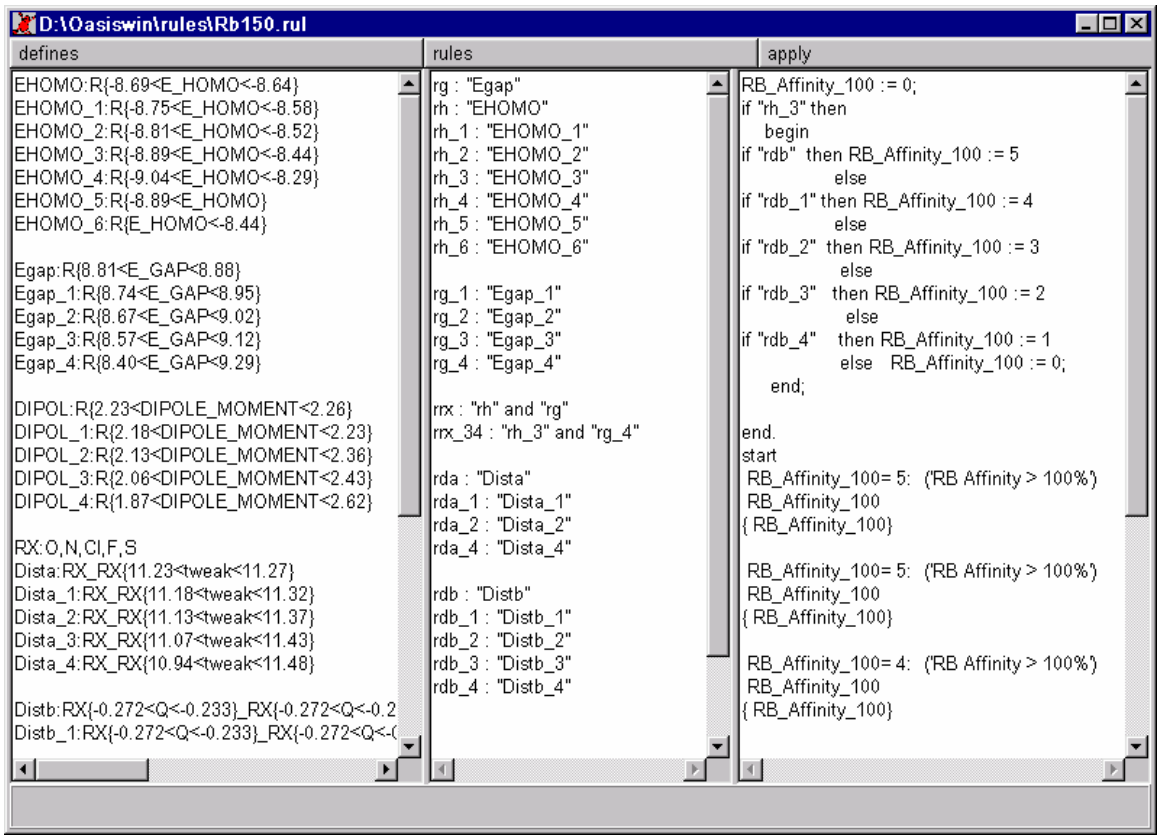
Неактивни - RBA<0.1%

Фигура 2.4

Съгласно *COREPA* алгоритъма е изведено по едно дърво на решението за всяка от разглежданите 5 групи съединения с близка естрогенна активност.

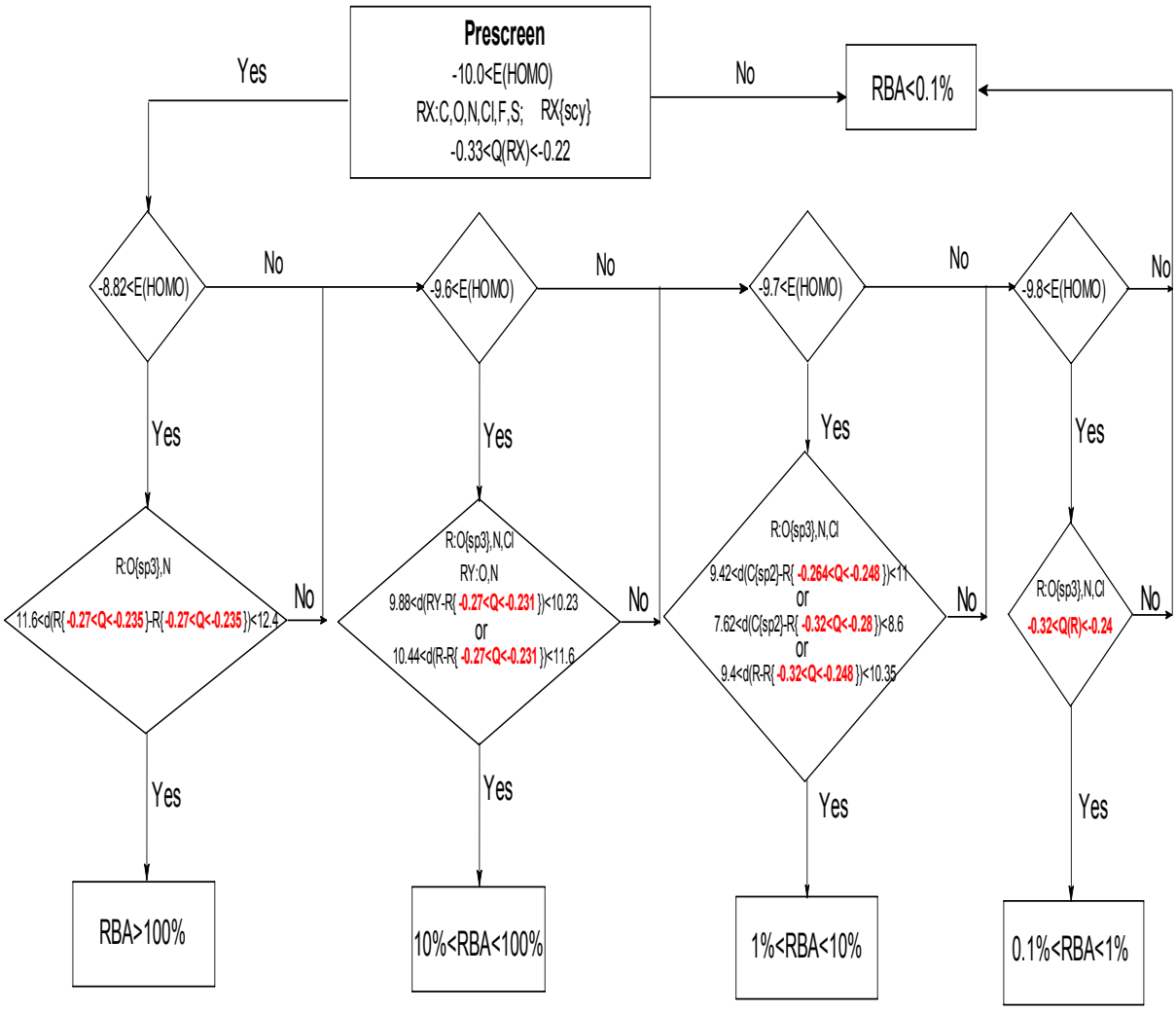
Група 1 (висока активност)	RBA>100%
Група 2	10% < RBA<100%
Група 3	1% < RBA<10%
Група 4 (ниска активност)	0.1% < RBA<1%
Група 5 (неактивни)	RBA<0.1%

На Фигура 2.5 е показано дървото на решението за групата на съединения с висока активност (RBA > 100%) , записано като RDL правила.



Фигура 2.5

На Фигура 2.6 е показано опростен вариант на дърво на решението за всички групи на активност.



Фигура 2.6

Моделът за биологичната активност може да бъде изведен, като се използуват данни за структурно сходни съединения, получени от експерименти с минимално разнообразие на тестваните биологични видове, или да се разшири множеството данни с такива за структурно различни съединения, за които данните за биологична активност са получени при експерименти с различни организми. В първия случай съществува вероятност изведеният модел да не е валиден за съединения с различна структура от тази на използуваните. Във втория случай проблемът се състои в евентуалните различия в механизма на рецепторно взаимодействие при различните организми. Построяването на QSAR модела е извършено на три етапа - модел I, модел II, модел III.

За построяването на модел I е използувана база данни, съдържаща 58 стероидни и нестероидни съединения. Модел II е построен въз основа на данни за афинитета на свързване към човешкия естрогенен рецептор (hERα). Модел III е построен въз основа на данни за афинитета на свързване към естрогенен рецептор при човека, мишки, плъхове и ракови клетки (MCF7 cells). Модел III трябва да се разглежда като усреднен модел за афинитета на свързване към естрогенния рецептор за млекопитаещи. Най-добри резултати са постигнати с последния модел III, който се основава на база данни със структурно различни съединения (използуването на база данни със структурно различни съединения е за сметка на комбинирането на данни, получени при експерименти с различни организми). Резултатите от прогнозите, направени с Модел III) показват, че от 1000 съединения от базата данни на съединения с висок обем на производство (HPVC) (това е целият списък, предоставен от Европейското Химично бюро), само за седем съединения е предсказан нисък афинитет на свързване с естрогенния рецептор. Четири от тях са тествани с ERE-CALUX и hERα-binding анализи, според които за три от съединенията резултатите са потвърдени. За четвъртото съединения (riboflavin), потвърждение на прогнозираните резултати е получено при in vitro и in vivo анализ. За останалите 13 съединения, които са прогнозирани като неактивни според Модел III, също е получено че са неактивни според ERE-CALUX и hERα-binding анализите.

В Таблица 2.1 са показани резултати от прогнозиране и експериментални резултати от ERE-CALUX и hERalpha-binding анализи за част от съединенията от базата данни от химикали с висок обем на производство, получени в рамките на проекта EDAEP.

Таблица 2.1

	CAS	Химично наименование	Прогнозирани стойности			Експериментални стойности	
			Модел I pKi rank	Модел II RBA	Модел III RBA	ERE-CALUX анализ	HERalpha -binding
1	60004	edetic_acid	+' a	n b	n b	negd	neg
2	79947	4,4-isopropylidenebis(2,6-dibromphenol)	+	1-10%	n	-' e	-' e
3	81118	4,4'diaminostilbene-2,2'-disulphonic acid	+	1-10%	n	neg	neg
4	83885	riboflavin	+	1-10%	1-10%	neg	neg
5	85563	2-(4-Chlorobenzoyl)benzoic acid	+	1-10%	1-10%	-	-

6	106752	oxydiethylene bis (chloroformate)	+	n	n	-	-
7	112243	trientine	+	n	n	neg	neg
8	112572	3,6,9-triazaundecamethylenediamine	+	n	n	neg	neg
9	112607	3,6,9-trioxaundecane-1,11-diol	+	n	n	neg	neg
10	118821	4,4'-Methylenebis (2,6-di-trt-butylphenol)	+	1-10%	1-10%	-	-
11	119802	2,2'-dithiodi(benzoic acid)	+	10-100%	1-10%	1.15E-05	neg
12	5102830	2,2'-[(3,3'-dichloro[1,1'-biphenyl]-4,4-diyl)bis(azo)]bis[N-(2,4-dimethylphenyl)-3-oxobutyramide]	+	10-100%	n	-	-
13	1559348	3,6,9,12-tetraoxahexadecan-1-ol	+	n	n	-	-
14	1675543	bisphenol A diglycidyl ether	+	10-100%	10-100%	-	0.002
15	1761713	4,4'-methylenebis(cyclohexylamine)	+	n	n	neg	neg
16	2310170	phosalone	+	1-10%	n	neg	neg
17	2494895	2-[(p-aminophenyl)sulphonyl]ethyl hydrogensulphate	+	1-10%	n	-	neg
18	4035896	1,3,5-tris(6-isocyanatohexyl)biuret	+	n	n	-	-
19	4098719	3-isocyanatomethyl-3,5,5-trimethylcyclohexyl_isocyanate	+	n	n	neg	neg
20	5567157	2,2'-[(3,3'-dichloro[1,1'-biphenyl]-4,4-diyl)bis(azo)]bis(azo)bis[N-(4-chloro-2,5-diethoxyphenyl)-3-oxobutyramide]	+	1-10%	n	-	-
21	6358856	2,2'-[3,3'-dichloro[1,1'-biphenyl]-4,4'-diyl)bis(azo)]bis[3-oxo-N-phenylbutyramide]	+	1-10%	n	-	-
22	7434404	ethane-1,2-diylbis(oxyethane-2,1-diyl)bisheptonate	+	n	n	-	-
23	13674855	tris(2-chloro-1-methylethyl)phosphate	+	n	n	-	-
24	13684634	Phenmedipham	+	1-10%	n	-	-
25	14861177	4-(2,4-diphenoxy)aniline	+	1-10%	n	-	-
26	15894709	N,N'"-1,6-hexanediylbis[N'-cyanoguanidine]	+	n	n	-	-
27	22839470	I-Aspartyl-L-phenylalanine methyl ester	+	1-10%	n	-	-
28	24800440	[(methylethylene)bis(oxy)]dipropanol	+	n	n	neg	neg
29	26919506		+	1-10%	n	-	-
30	27375526		+	1-10%	1-10%	-	-
31	36734197	Iprodion	+	1-10%	n	-	-
32	38051104	2,2-bis(chloromethyl)timethylene bis(bis(2-chloroethyl)phosphate	+	n	N	-	-
33	40843730		+	1-10%	N	-	-
34	42576023	Biphenox	+	1-10%	N	-	-
35	56966520	5-chloro-2-(2,4-dichlorophenoxy)aniline	+	n	n	-	-
36	80057	bisphenol A	+	1-10%	1-10%	0.007	0.005
37	95487	o-cresol	nrf	n	n	neg	neg

38	96764	2,4-di-tert-butylphenol	nr	n	n	neg	neg
39	50293	Clofenotane	nr	n	n	neg	neg
40	106445	p-cresol	nr	n	n	neg	neg

За 9 съединения с висок афинитет на свързване към естрогенния рецептор, избрани от QSAR списъка на съединения с нисък обем на производство (LPVC) е получено 100% съвпадение между QSAR прогнозите и резултатите от ERE-CALUX и hERα-binding анализ. Прогнозите за висока естрогенна активност са потвърдени и от други *in vitro* и *in vivo* анализи. В Таблица 2.2 са показани прогнозни и експериментални резултати от моделиране и *in vitro* и *in vivo* анализ за част от базата данни съединения с нисък обем на производство, получени в рамките на проекта EDAEP.

Таблица 2.2

CAS	NAME	Прогнозирани стойности		Експериментални стойности				
				in vitro				in vivo
		Model II	Model III	ERE-CALUX	hERa Binding	CARP-HEP	E-scree assay	Uterotropic
		RBA	RBA	assay a	(EC50)	Vtg assay c	on MCF-7	assay
56951	Chlorhexidine diacetate salt	>100%	>100%	1,6.10-2	-' b	-' b	-' b	-' b
79970	4,4'-isopropylidenedi-o-cresol	>100%	>100%	1,2.10-2	10mM	100mM	positive	positive
84162	meso-3,4-bis(4-hydroxyphenyl)hexane	>100%	>100%	30	1nM	10-25 nM	positive	positive
84195	3,4-bis-(4-acetoxyphenyl)-2,4-hexadiene	>100%	100%	1.2	25nM	-	positive	positive
446720	4',5,7-Trihydroxyisoflavone	>100%	>100%	-' b	-	-	positive	negative
458377	Curcumin	>100%	100%	1,2.10-2	10mM	-	positive	positive
500389	Nordihydroguaiaretic acid	>100%	>100%	1,3.10-3	>100mM	+' d	positive	positive
1944123	Fenoterol hydrobromide	>100%	100%	10	-	-	-	-
2411894	o-Cresolphthalein Complexone	>100%	100%	-	-	-	positive	negative
15875135	1,3,5-tris[3-(dimethylamino)propyl]hexahydro-1,3,5-triazine	>100%	100%	-	-	-	positive	negative

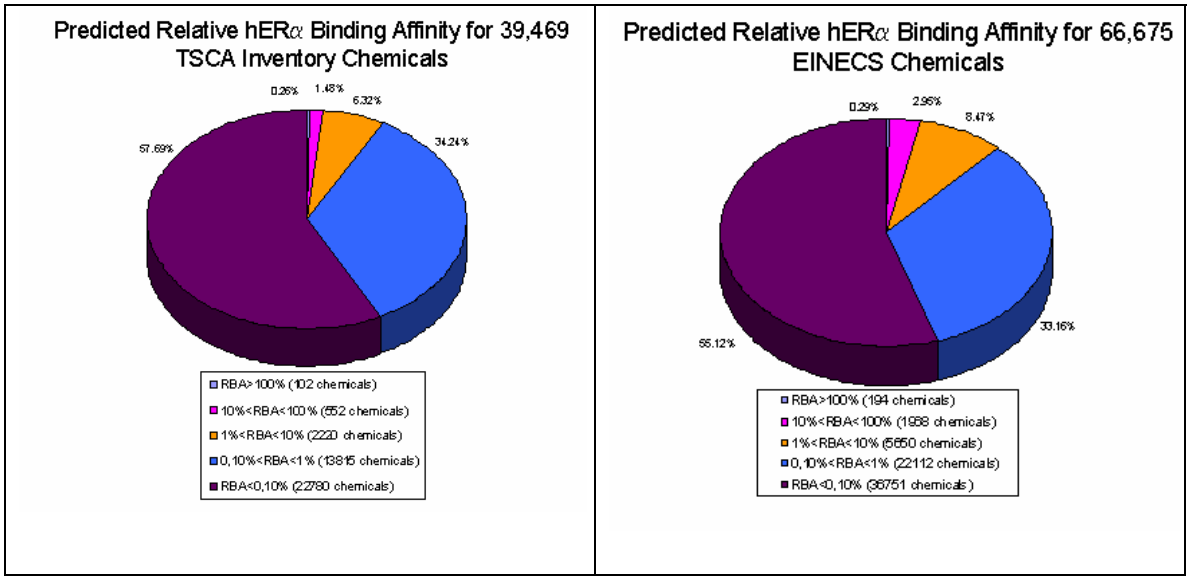
a Relative potency to Estrodiol

b not tested

c lowest effect levels

d potent antiestrogen

Въз основа на добрата предсказваща способност на QSAR модела за съединения с високи афинитет, и това, че такива съединения не са намерени в списъка от съединенията с висок обем на производство, може да бъде направено заключението, че сред тези съединения, няма високо активни. Получените резултати са обещаващи по отношение на възможностите за скрининг на големи бази данни с химични съединения. На Фигура 2.7 са показани резултатите от предсказване на афинитета за свързване с естрогенния рецептор за базите данни *TSCA* и *EINECS*.



Фигура 2.7

3. Заключение

Разработените методи в дисертационния труд позволяват да се открие връзка между свойствата на молекулите и биологичната им активност, която впоследствие да се използува за обяснение на биологичния ефект, предсказване на биологични свойства и генериране на нови съединения със желани свойства.

- Разработен е вероятностен метод за откриване и интерпретация на закономерности в масиви данни (*COREPA*). Методът е използуван за извличане на правила за оценка на биологичната активност на различни по структура химични съединения и построяване на модел на биологичното свойство във вид на дърво на решението. Интерпретацията на условията във възлите на дървото на решението дава възможност на химика да създаде хипотеза за обяснение на биологичния ефект. Полученото дърво на решението може да се

използува за предсказване на активността на молекули с неизвестен биологичен ефект, както и за класификацията им. Като предимство може да се посочи, че не се изисква наличието на детайлно описание на рецептора, нито тримерно наслагване на структурите, а се използуват изчислени или експериментално измерени параметри.

- Създаден е формален език *Rule Description Language*, който позволява записването на дървото на решението като съвкупност от правила. Правилата се прилагат многократно върху различни бази данни. Правилата могат да включват сложни обобщени фрагменти от молекулната структура. Тези фрагменти могат да се задават чрез вградения в софтуера редактор на молекулни структури.
- Поради необходимостта от отчитане гъвкавостта на молекулите (всяко съединение може да има повече от един конформер) е разработен генетичен алгоритъм за генериране на множеството от максимално различните конформери на зададено химично съединение. За разлика от останалите генетични алгоритми, които оценяват всеки индивид, предложеният метод оценява цялата популация и има за цел намирането на оптимално множество от индивиди. Алгоритъмът се използува както за намиране на набор от конформери и записването им в база данни с цел по-нататъшен анализ, така и за целенасочено генериране на конформери, удовлетворяващи зададени критерии.
- Разработен е генетичен алгоритъм на две нива, с цел генерирането на нови съединения със желани биологични свойства. Свойствата се задават във вид на правила, записани на езика *RDL*. На първо ниво генетичен алгоритъм генерира новите съединения, а за намиране подходящите конформери се прилага алгоритъмът за генериране на максимално различни конформери. Предимство на алгоритъма е че дава възможност и за отчитане гъвкавостта на съединенията в хода на генерацията, което го отличава от известните алгоритми.
- Предложените алгоритми и методи са реализирани в съвкупност от приложни програми *COREPA, FORECAST, GAS32, LEADGEN* за MSWindows 95, 98, NT, с използуване на обектен Pascal и средата за разработка на Borland International - Delphi. Всички изброени приложни програми позволяват визуализиране (двумерно и тримерно) на химичните съединения, въвеждането на нови химични съединения чрез двумерен и тримерен редактор, запис и четене на

химичните съединения във/от файлове, изчисляване на голям брой параметри на съединенията и други.

- Разработените алгоритми и програми са използувани за предсказване на естрогенна активност на химичните съединения, както и на фототоксичност, остра токсичност и други биологични свойства. Химичното съединение се класифицира в една от няколко групи (например на активни и неактивни съединения), въз основа на изведения модел за механизма на действие на биологичното свойство (построеното дърво на решението). След като моделът е построен, и съединенията класифицирани, то може да се приложи метод за количествена оценка на токсичността в групата съединения с общ механизъм на действие. За количествена оценка също се използуват различни типове молекулни параметри, като физикохимични свойства, биологични свойства, квантово- химични електронни или стереохимични параметри. Експерименталните тестове показват добро съвпадение с прогнозните резултати.

4. Списък на публикациите, свързани с дисертационния труд

1. Mekenyan O.G. Dimitrov D. ***Nikolova N***., Karabunarliev S.H. Conformational Coverage by a Genetic Algorithm. Chem. Inf. Comput. Sci. . 1999, 39/6, 997-1016.
2. Mekenyan O.G., ***Nikolova N.***, Karabunarliev, S.H., Bradbury, S.P., Ankley G.D., Hansen, B., New Developments in a Hazard Identification Algorithm For Hormone Receptor Ligands. Quant Struct. – Act. Relat. ,1999, 18, 139 –153
3. Karabunarliev, S.H., ***Nikolova N.,*** Nikolov N. and O.G. Mekenyan. . Rule interpreter: A chemical language that implements decision rules based on molecular structure. THEOCHEM (in press)
4. Mekenyan O., Karabunarliev S., ***Nikolova N.***, "The New OASIS tools for Fuzzy Modelling of Chemical Structures" , Годишник на Софийския Университет, т.91, Годишник на Софийския Университет “св. Климент Охридски”, Химически Факултет, т.91, 2001г , стр. 145-157
5. Mekenyan O., Karabunarliev S., ***Nikolova N.***, Dimitrov D., Ivanov J., Grancharov I., Nikolov N., "The OASIS tools for Fuzzy Modelling of Chemical structures: Applications for Toxicity Screening of the EU inventories", 8-th International Workshop on Quantitative Structure-

Activity Relationship (QSAR) in the Environmental Sciences, May, 16-20, 1998, Baltimore, Marylend, USA.

6. Mekenyan O.G. ***Nikolova N***., and Schmieder P. Dynamic 3D QSAR Techniques: Applications in Toxicology, THEOCHEM (in press).