Централна Лаборатория по паралелна обработка на информацията

Българска Академия на Науките

инж. НИНА ГЕОРГИЕВА НИКОЛОВА

КОМПЮТЪРНИ МЕТОДИ ЗА МОДЕЛИРАНЕ НА ХИМИЧНИ СТРУКТУРИ

Дисертация за присъждане на образователната и научна степен "Доктор"

Научни ръководители:

чл.-кор. проф. д.т.н. инж. Кирил Любенов Боянов

проф. д.х.н. Ованес Гаро Мекенян

София, 2001

Въведение

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 за създаването на нови лекарства във фармацевтиката, химията и токсикологията. При създаването на нови химични съединения, както и при оценката на биологичните им свойства, трябва да се оценява терапевтичния ефект или токсичността на голям брой съединения, част от които дори 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 се увеличава с 775000 годишно. Всяка година започва използуването на около 1000 от тези съединения. Само малка част от тези съединения имат експериментално измерени свойства за оценка на риска от замърсяване на околната среда. В Съединените щати базата данни на Агенцията за опазване на околната среда (*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 USA 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 .

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 за създаването на нови химични съединения (молекулен дизайн).

Във връзка с тази основна цел, в дисертационната работа са поставени за решаване следните задачи:

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

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*– т.2.1). 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 правила (*RDL* – т.2.2), което осигурява универсален метод за записване на правила, зависещи от структурата на химичните съединения. Построеният модел може да се използува многократно за автоматично предсказване на моделираното свойство, търсене на молекули с зададени свойства(т. 4.1.3) или за генериране на нови структури със зададени свойства (т.3.2).

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* – т.3.1, както и на разработения генетичен алгоритъм за компютърно конструиране на нови химични структури със предварително зададени свойства - *LeadGen* - т.3.2 . Свойствата се задават чрез техния модел, записан като правила в езика *RDL*. Генетичният алгоритъм на първо ниво конструира нови химични структури, удовлетворяващи зададените условия, а на второ ниво за всяка нова молекула се прилага генетичният алгоритъм от т. 3.1 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 върху предсказването на естрогенна активност на химичните съединения в рамките международен проект на Европейската общност.

1. Методи за разпознаване на образи и използването им при моделиране на химични съединения

Разпознаване на образите е научна област, изучаваща функционирането на системи, които извличат общите характеристики на съвкупност от обекти, както и методите за създаване на такива системи. За решението на проблема за разпознаване на образи (*pattern recognition*) се използват широк кръг от методи, които включват статистическо разпознаване на образи, дискриминантен анализ, извличане на свойства, клъстерен анализ, синтактично разпознаване на образи, невронни мрежи, нелинейна оптимизация, методи за извличане на знания (*knowledge discovery*) в големи бази данни (*data mining*). Понастоящем методите за разпознаване на образи намират приложение в огромен брой най-разнообразни области, като разпознаване на ръкописен текст и разпознаване на мимики и жестове, анализ на геоложки данни, търсене в документи, разпознаване на следите на елементарните частици и много други[23] [24].

Методите за извличане на знания се прилагат за да се открият нови зависимости в дадена съвкупност от данни, които по-нататък да бъдат анализирани от специалистите в областта. За тази цел се построява модел на изследваната област въз основа на предварително събрани данни. Модел на обект A е друг нов обект A^*, на когото наблюдателят B може да задава въпроси, с цел получаване на отговори за обекта A. Моделите могат да бъдат класификационни, регресионни, времеви серии, клъстериращи, откриване на последователности и други. Класификационните модели, регресионните модели и моделите на времеви серии се използват главно за *предсказване*, докато останалите могат да бъдат полезни за *описание* на поведението на обектите, чиито събрани данни се анализират. Най-често използваните модели са за класификация и регресия.

Процесът на разпознаване включва измерване и/или получаване на данните за формиране на вектор на образа; извличане на отличителни признаци (намиране на

характеристики) и определяне принадлежността на непознатия обект към клас (класификация) чрез прилагането на решаващи правила. Процесът на разпознаване на образи може да се разглежда и като процес на преобразуване на информацията от подсимволно ниво (сигнали) до символно ниво (значения). В класическата книга на Дуда и Харт [102] е отбелязано, че класифицирането на обектите не е достатъчно за да се извлече тяхното описание. Важна задача е да се намерят дескриптори от по-високо ниво, които описват съответната реална ситуация.

Изследването на химичните съединения се състои от няколко ключови етапа: въвеждане и съхраняване на молекулните структури, получаване на молекулните дескриптори и последващо прилагане на различни методи за разпознаване на образи, статистически анализ и други, с цел откриване на общите характеристики обуславящи дадено свойство (обикновено биологична активност). Молекулните дескриптори представляват множество количествени параметри, получени чрез емпирични или числени методи, химични измервания и тестове. Резултатите от изследванията дават възможност за:

- Изясняване на механизмите, водещи до конкретното свойство (биологична активност);
- Предсказване на токсичност и други биологични свойства и оценка на замърсяването на околната среда;
- Търсене на структури с предварително зададени свойства (лекарства, агрохимикали, други) сред известните химични съединения;
- Компютърно генериране на нови структури със зададени свойства и последващото им експериментално синтезиране (de novo design).

1.1. Представяне на химичните структури

Обектите, които представляват интерес в дисертационната работа са химични съединения, зададени с тяхната структура и физикохимичните им свойства. Задачата се състои в това да се "разпознаят" биологично активните съединения и да се предложат методи, позволяващи целенасочено "конструиране" на нови

биологично активни съединения. Биологичната активност е свойство, дефинирано в зависимост от изследвания проблем, например токсичност, мутагенност или лекарствено действие на съединенията. Обикновено тя се измерва с концентрацията на съединението, необходима, за да се постигне съответния ефект.
Биологичният ефект на химичните съединения е резултат от съвкупност от сложни събития, включващи движението на молекулата в организма, свързването и с рецептор и метаболитни процеси. Тъй като в общия случай поведението на молекулата в организма не е добре известно, предсказването на биологичния ефект се свежда до намирането на зависимости между свойства и параметри на молекулата, които могат да бъдат измерени и изчислени. След като такава зависимост е намерена, тя може да се използува за предсказване на свойстватана химичните съединения.

1.1.1. Структура на химичните съединения

Структурата на химичните съединения се задава по следните начини [35]:

- едномерно представяне чрез символен низ, който може да бъде получен от структурната формула
- двумерна структурна диаграма, въз основа на структурната формула
- тримерен модел, където атомите са представени със сфери, а връзките със пръчки, разположени под съответните ъгли (*ball-and-stick model, 3D модел*)

Химичните съединения се определят от техния състав (от кои атоми се състоят) и свързаност (как и с какви връзки са свързани атомите). Добре познатата на химиците структурна диаграма всъщност не е изображение на молекулата, а граф, описващ нейната топологична структура. Атомите се представят с възлите на графа, а връзките - с ребра на графа. Топологичната структура обаче не определя как атомите са разположени в пространството. Координатите на атомите в пространството могат да бъдат експериментално получени чрез рентгено-кристалографски анализ или теоретично изчислени чрез квантово-химични методи.

1.1.2. Конформационен анализ

Тримерната конфигурация на дадена молекула обикновено не е единствена. Съвременните теории не ограничават пространственото разположение на химичното съединение до това с най-ниска енергия. Възможно е също да се окаже, че търсената активна структура не е най-ниско енергетичната. Във всички известни примери, за които има експериментални данни, се появяват промени в пространственото разположение и в рецептора при свързване с химичното съединение.

Молекулата може да приема множество пространствени конфигурации под влияние на различни фактори, като промени в температурата, различни разтворители, ензими, взаимодействие с рецептор и други. Тези конфигурации се наричат "*конформации*". Изследването на множеството конформации, възможни за дадена структура, се нарича конформационен анализ. Обикновено с термина "*конформация*" се наричат неидентичните положения на атомите в молекулата, които се получават при въртене на една или повече връзки без те да се прекъсват. В по-широк смисъл, терминът се употребява за да обозначи структури, получени след промяна на валентни ъгли. Терминът "*конформер*" се използува за множеството конформации, съответстващи на локални енергийни минимуми.

Молекули, които могат да приемат множество конфигурации, се наричат *гъвкави* за разлика от *твърди* молекули, които могат да приемат само единствена или много малък брой конфигурации. Броят на възможните конформации зависи от гъвкавостта на изследваните молекули. Гъвкавите модели са по-реалистични, но по-трудни за описание и изследване и изискващи повече време за изчисление. При гъвкави структури намирането на всички възможни конформации е комбинаторен проблем и не е възможно да се приложи изчерпателно търсене. Предложени са различни детерминистични, стохастични и генетични алгоритми за генериране на множество конформации [39, 75].

Методите за генериране на конформери обикновено използуват изчерпателно търсене в конформационното пространство, след което множеството конформери

се филтрира въз основа на дадени ограничения. Така се получава редуцирано множество от конформери, удовлетворяващи зададени критерии. Разпространен детерминистичен и систематичен подход е използуваният в пакета за молекулно моделиране SYBYL [89], където всички циклични части на молекулата се считат за твърди и не се деформират. По-гъвкава процедура за изчерпателно търсене е процедурата на Lipton и Still [53], използуваща алгоритъм за търсене в дърво (*internal coordinate tree-search*). Въз основа на нискоенергетичен конформер се генерират всички комбинации от ротамери (конформери, получени чрез ротация около връзка) за зададена стъпка на завъртане на торзионните ъгли. Приемат се само конформерите, които удовлетворяват дадени геометрични тестове за отстраняване на напрегнатите структури. Подобен подход се използува и при намирането на конформер, подходящ за даден рецептор. Методът позволява да се изследва конформационната гъвкавост на наситени циклични структури. Недостатъкът на всички систематични алгоритми е, че броят на възможните решения (конформери) зависи експоненциално от степените на свобода на структурата. Един от начините за преодоляването на този проблем е използуването на стохастичен алгоритъм. Chang [11] описва Монте Карло алгоритъм за конформационно търсене и показват неговата ефективност за намиране на нискоенергетични конформери. Недостатък на стохастичния подход е ниската му ефективност, ако са зададени голям брой ограничителни условия за генерираните конформери.

Разглеждането на химичното съединение като съвкупност от възможни разположения в пространството поражда следните проблеми, както при извеждането на модели за активността, така и при търсенето в бази данни от химични съединения:

- Дали да се записва само един или повече конформери при съхраняването на химичните съединения и кои от тях. При представяне и търсене на съединението по единствен конформер може да се пропусне съответствие със заявката за търсене поради липсата на подходящия активен конформер, както и

да се изведе непълен модел, поради липса на данни за останалите конформери. В случая на съхраняване на няколко конформера в БД е от значение въпроса кои конформери да бъдат записани. Генерирането на всички възможни конформери е комбинаторен проблем и не е реално осъществимо, особено при големи и гъвкави молекули. Генерирането на конформери, съответстващи на енергетичен минимум също изисква значителни изчислителни ресурси и има следния недостатък - активните конформации могат да не са тези в енергетичен минимум.

- Дали конформерите да се генерират при необходимост или да се записват в базата данни предварително. Недостатък на генериране на конформерите в процеса на търсене е увеличаване на времето за търсене с това, необходимо за генерацията.
- Оценката на сходството между химичните съединения, представени с набор от конформери, се усложнява в сравнение с оценката за сходството между съединения, представени с единствено пространствено разположение. Необходимо е освен това да се оценява приликата между самите конформери.

1.1.3. Компютърно представяне и съхраняване на химични съединения

В голяма част от химичните бази данни се съхранява и обработва именно топологичната информация за съединенията. Такива бази данни са известно под името 2D бази данни. Структурите се описват с таблици на свързаност, в които са изброени атомите и връзките. Топологичната информация се използува за разпознаване на структури или търсене на структурни фрагменти в бази данни. През 80-те г. повечето фармацевтични компании създават 2D бази данни от съединенията, които произвеждат или продават [20]. 2D базите данни продължават да се използуват основно за поддържане на архиви от химични съединения.

Необходимостта да се отчита не само топологията, но и формата на химичното съединение е довело да разработване на методи, използуващи и анализиращи информацията за неговата тримерна структура. Базите данни, съдържащи тримерни координати на химични съединения са известни под името *3D бази данни*. Първият пример за 3D база данни е системата на Cambridge Crystallographic data Centre[20],

която съдържа тримерна информация за малки органични съединения, получени от кристалографски данни. Тази система естествено претърпява развитие през годините, както и се създават множество други 3D бази данни [14,54,89], особено с появата на софтуер, позволяващ автоматично генериране на тримерни координати от таблицата на свързаност (CONCORD [18]).

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. Така например, огромната част от известния софтуер записва и обработва информацията за химичните структури в специфични файлови формати, които не се управляват от система за управление на бази данни. Пример за това са разпространения MOL file формат, разработен и публикуван от MDL Information Systems, MOL2 (разработен от Tripos Associates), SDF (space delimited) формат [19], PDB (Protein Database Format), както и десетки други формати, като засега не съществува стандарт за файлове, съдържащи химични структури. В последните години има примери за системи, използуващи СУБД - MACCS-3D [54] позволява част от информацията да бъде поддържана в ORACLE база данни; Catalyst е разработен в тясна интеграция с релационна СУБД, но предоставя на потребителя информация само чрез файлове. Програмните системи на MDL Information Systems са построени въз основа на йерархичната база данни ISIS. Слабото използуване на стандартна технология като релационните БД може да се обясни със специфичния характер на софтуера за химично моделиране и фактът, 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.

Преди широкото разпространение на графичните дисплеи и средства за въвеждане на информация, основният начин за въвеждане на данни чрез клавиатура е наложил разработването на представяне на химичните съединения като символни низове. С въвеждането на графичния интерфейс обаче става възможно да се въвеждат

директно структурните диаграми. Системите за молекулно моделиране обикновено предоставят тримерна картина на съединението.

Някои от методите за вход от клавиатура са се оказали достатъчно бързи и удобни. Означението SMILES [97] (описание на свързаноста на химичното съединение чрез символен низ) се е превърнало в стандарт за линейно представяне и широко се използува като алтернатива на графичния вход. От символния низ с подходящи алгоритми може да се генерира 2D и 3D изображение на химичното съединение [17].

1.1.4. Молекулни дескриптори

Освен свързаността и пространственото разположение на молекулата, интерес представляват и множество числени параметри, отразяващи различни свойства на съединението. Моделирането на химични съединения има за цел да изведе модел на дадено биологично свойство, което от своя страна зависи от реакционната способност на химичното съединение.

Активността (или реакционната способност) на химичното съединение се влияе от множество фактори.. Активността може да е резултат от размера и формата на молекулата, типове на връзките между атомите, както и други свойства. Например възможността за реакция между съединението и вътрешно клетъчните структури зависи от неговата електрофилност. Възможно е също така, проявата на биологична активност да се дължи не на самото съединение, а на продуктите, получавани от него при метаболитния процес в организма [17, 50].

За да се отчетат тези фактори се използуват множество молекулни дескриптори – топологични, физикохимични, квантовохимични, взаимно разположение на атоми и атомни групи, свойства на метаболитите на разглежданото химично съединение и други.

Топологични индекси

Молекулната топология се характеризира чрез т.нар. топологични индекси. Това са числени характеристики на молекулата, основани на определени топологични инварианти на съответен молекулен граф. Топологичните индекси се делят на две

групи: индекси, основани на атомната свързаност и индекси, при които се отчита и типа на атомите и връзките. Установена е корелация между някои химични свойства и топологичните индекси (температура на кипене, токсичност [6]). Пример за топологичен индекс е числото на Wiener. Ако молекулната структура е представена като химичен граф, където атомите са възлите на графа и връзките са ребрата на графа, то числото на Винер е сумата от всички най-къси разстояния между атомите, като разстоянието се дефинира като сума от броя на връзките между атомите.

Физикохимични свойства

При анализа на активността на химичните съединения се използуват и експериментални физикохимични свойства, като Ван дер Ваалсова повърхност и обем, диполен момент, молекулна форма, стерични свойства, температура на кипене, налягане на парите, електронегативност и някои квантови или електронни и други свойства [17,35,100,106]. Някои физикохимични свойства се изчисляват, така например, обемът, изчислен по уравнението на Ван дер Ваалс за реалния газ се нарича Ван дер Ваалсов обем. Ван дер Ваалсов обем на молекулата се изчислява като сума от Ван дер Ваалсовите обеми на отделните нейни структурни елементи. Ван дер Ваалсова повърхност е сума от Ван дер Ваалсовите повърхности на отделните атоми на молекулата. Може да се изчисли и повърхността на молекулата, която е достъпна за атакуване отвън.

Квантовохимични дескриптори

Методите на квантовата механика позволяват изчисляване на електронната плътност, енергетичните нива и молекулната енергия, с цел да се предскаже реакционната способност на молекулите. Квантовохимичните индекси, наречени още стереоелектронни, са въведени въз основа на методи за изчисляване на молекулната електронна структура за основно (а някои и за възбудено) състояние на молекулите. Като глобални електронни параметри могат да се посочат енергетичните нива на граничните молекулни орбитали, като най-ниската заета молекулна орбитала (*LUMO*), най-високата заета молекулна орбитала (*HOMO*), тяхната разлика E_{gap}, изчислен диполен момент и др. Като локални електронни индекси се из-

ползуват атомните заряди q_i, порядъците на връзките p_{ij}, донорните S_i^E и акцепторни S_i^N свръхделокализационни индекси и атомните поляризуемости π_{ij} [100,106].

Необходимостта от изчисляване на огромен брой интеграли за молекулите с размер на съвременните лекарства е довела до използуването на апроксимации на силово поле [38] с цел по–бързо пресмятане енергията на съединенията. Молекулата се моделира като състояща се от сфери (атоми) свързани с еластични връзки. Вътрешната енергия се представя като сума от еластичността на връзките, огъването на ъглите и несъседните взаимодействия като Ван дер Ваалсови и Кулонови сили на привличане и отблъскване.

Изчисляването на молекулната динамика в даден разтвор моделира важни явления като загубата на степени на свобода при свързването на лекарството с рецептора. Приложението на методи от изкуствения интелект, като симулирано закаляване [81] или генетични алгоритми [75], може да допринесе за по-бързото обхождане на конформационното пространство на молекулата. Съгласно статистиката на Болцман [101] преобладават структурите с минимална енергия.

Съвременните компютри с висока производителност позволяват квантово механично изчисляване на молекули от порядъка на ДНК и молекулна динамика за системи от порядъка на 50000 атома. С комбиниране на двете техники в последните години се появява квантовата динамика, която дава възможност да се симулира динамиката на реакциите между молекулите. Въпреки че изследваните системи са груби приближения на реалните живи организми или клетки, възможността да се изучават молекулярните взаимодействия и реакции с помощта на компютър вместо микроскоп се оказва не само привлекателен, но и ефективен метод за изучаване и моделиране на молекулния строеж и предсказването на свойства на химичните съединения.

Изброените молекулни дескриптори предоставят огромно количество информация за всяко химично съединение. Следвайки класическия подход от системите за разпознаване на образи, можем да представим числовите молекулни характеристики като n-мерен вектор от свойства

$$\boxtimes = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \quad u \quad \boxtimes = (x_1 \quad x_2 \quad \dots \quad x_n)$$

В този случай могат да се използуват методи на математическата статистика за анализ на многомерни данни, стохастични и вероятностни методи. Тук съществуват обаче следните проблеми:

- Непълнота на данните - Не всички разглеждани съединения имат един и същ набор от свойства или не се разполага с данни за всички свойства;
- Методите за клъстериране и класификация изискват оценка за геометричната близост на векторите. При използуване само на числовите характеристики, не се взима предвид структурата на химичното съединение. Ако структурата на химичното съединение трябва да се вземе предвид, то трябва да се дефинира понятие за сходство между химичните съединения.
- Разглеждането на химичното съединение не като уникална структура, а като съвкупност от конформери изисква дефиниция на понятието "сходство", такава, че съвкупността от конформери да се разглежда като цялостен обект.

Фармакофори

Редица свойства на молекулите, като някои типове биологична активност, произтичат от взаимодействието на тяхна специфична част с рецептори, клетъчна мембрана, специфични части от ензими и макромолекули. Тази характерна подструктура на молекулата, която е съществена за свързването с рецептора се нарича *фармакофор* (в случай на полезно, лекарствено действие) или *токсикофор* (в случай на вредно, токсично действие). Използува се още и терминът *биофор*, за да се обозначи структурната особеност на молекулата, която води до биологична активност. Тези особености обикновено не се свеждат до единичен фрагмент от няколко свързани атома в молекулата, а са съвкупност от взаимно влияещи си фрагменти (всеки от тези фрагменти е отговорен за взаимодействия, които като цяло определят биологичната активност). Следователно, биофорът се определя като съвкупност от всички особености на молекулата, действуващи като "ключ" за разпознаване на специфичната част на рецептора ("ключалката"), осъществяване на свързването им

и проявата на биологичното свойство. Под токсикофор и фармакофор трябва да се разбира тримерната структура на фрагмент, състоящ се от атомни центрове, разстоянията между тях, и електронните им характеристики, които определят биологичните му свойства.

За компютърното моделиране на химични съединения начинът на представяне на фармакофорите е от съществетно значение. В Chem-X [14,20] фармакофорите се представят като съвкупност от центрове (заряди, донори и акцептори на водородна връзка, центрове на ароматни пръстени, и други) и разстоянията между тях. От тези разстояния се конструират 15 32-битови ключове за всяка молекула, като във всяка 32-битова дума са записани разстоянията между двойки центрове. Отчитането на функционални групи и други структурни особености на съединенията се задава чрез набор от ключови думи като CHAIN, RING, HDON, HACC, AROMATIC, n-CYCLIC, MEDIUM, LARGE. В ALADDIN [91] се отчитат всички типове геометрични отношения, като могат да се дефинират точки (център на масите на група атоми), линии (връзки между атоми), равнини (равнина на ароматен пръстен) и разстояния между тях, както и ъгли между прави и равнини. В CAVEAT [91] не се използуват функционални групи, а само векторни съотношения.

Идентифицирането на фармакофор е сложна задача. Откриването на тези в общия случай тримерни характеристики изисква прилагане на различни методи, включително методи за разпознаване на образи, върху бази данни от съединения с известно биологично действие. Поради липса или непълнота на данните, част от характеристиките могат да се окажат неуместни и да бъдат отхвърлени въз основа на експериментални наблюдения или компютърно моделиране. След успешното им намиране обаче, те могат да се използуват за оценяване на сходството между структурите, като се отчита наличието на общи характеристики в различните съединения.

Компютърното идентифициране на фармакофор е все още нерешен проблем. Съществуват множество изследвания и софтуер, но нито един не дава изчерпателно

решение. В идеалния случай, системата за компютърно конструиране на фармакофори трябва да дава поне толкова добри решения, колкото би предложил химик; решенията да бъдат устойчиви на малки изменения и грешки във обучаващата база данни и статистически значими [91].

1.2. Сходство на химичните съединения

По-голямата част от методите за разпознаване на образи използуват под една или друга форма сравнение между обектите или между обект и множество обекти. Ако изследваните обекти се описват с многомерен вектор от своиства $\mathbf{x}= (x_1,\ldots,x_n)^t$, сравнението означава оценка на разстоянието $\delta(\mathbf{x}_i,\mathbf{x}_j)$ между два обекта $\mathbf{x}_i$ и $\mathbf{x}_j$. Най-използуваното разстояние е евклидово разстояние. Други разстояния са разстояние на Минковски от ред *s*, City block разстояние, разстояние на Чебишов. Разстоянието на Танимото [23, 105] определя разстоянието между две множества, дефинирано като:

$$D_{Tanimoto}(S_1,S_2)=\frac{n_1+n_2-2n_{12}}{n_1+n_2-n_{12}},$$

където n_1 и n_2 са броя на елементите съответно на множествата S_1 и S_2 и n_{12} е броя на общите за двете множества елементи.

Вероятностното разстояние е оценка на сходството между условната плътност на вероятностите $p(x|\omega_i)$ и $p(x|\omega_j)$ за два класа ω_i и ω_j. По- висока стойност на вероятностното разстояние означава по-голямо различие между класовете. Важни вероятностни разстояния са *Kullback-Leibler divergence*, разстояния на *Чернов, Bhattacharyya, Matusita, Hellinger, Mahalanobis, Patrick-Fisher* [9,22,31, 55,85,57].

Особен интерес представлява дефинирането на сходството между химичните съединения. Цитираните по-горе дефиниции за разстояния между вектори могат да се използуват за определяне разстоянието между вектори от молекулни дескриптори. Тези разстояния обаче не отчитат структурата на химичните съединения и съответно не отразяват структурните различия и прилики между молекулите. От своя страна структурните характеристики имат огромно значение за свойствата на молекулите.

При един от подходите, за критерий за сходство се приема наличието на общи структурни фрагменти (MACCS [54]). В Chem-X [14] сходството се оценява чрез съпоставяне на атомите, които могат да бъдат донори или акцептори на водородна връзка, както и чрез съпоставяне на разстоянията между тях.

При отчитане на пространствената структура и свързаността на химичните съединения, сравнението между тях се извършва чрез съпоставянето на тримерните им структури. Броят на възможните начини, по който две молекули могат да бъдат наложени една върху друга, обаче е огромен. Задачата се усложнява още повече, ако се разглеждат гъвкави молекули.

За оценка на приликата между структурите се използува разстоянието между идентични атоми при наслагване на структурите [86], което се получава при минимизиране на средната квадратична грешка RMSE:

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

където N е броят на точките на два вектора **x** и **y** в тримерно евклидово пространство.

В методите CoMFA, CoMSIA[48] (т.1.4.1) се използува метод за сравнение на молекули чрез налагането им в пространството и съпоставянето на техните полета в 3D решетка.

За автоматичното съпоставяне (наслагване) на структури са разработени различни методи:

При **твърдото съпоставяне** се избира множество от торзионни ъгли, които ще бъдат променяни през зададен интервал. За всеки набор от стойности на ъглите се изчислява енергията на молекулата. **Адиабатното съпоставяне** е аналогично на твърдото съпоставяне, но за всеки набор стойности на ъглите структурата се подлага на минимизация на енергията. При използуването на **Monte Carlo методи** за съпоставяне, ъгълът се избира по случаен начин и се променя със случайна стойност. Изчислява се енергията на молекулата. Ако молекула с такава геометрия не е била получена, то тя се запазва и се продължава с по-нататъшното и модифициране. Минимизация на енергията на молекулата може да бъде извършвана за всяка получена структура. При използуването на **симулирано зака-**

ляване се започва със система с висока температура, където почти всички конформации могат да съществуват. Постепенно температурата се понижава, за да се намери структурата с минимална енергия. Процедурата се повтаря многократно при различни начални условия, с цел намирането на различни локални минимуми. За съпоставяне на химични структури се използуват също и **генетични алгоритми**[75].

Обикновено експертите правят предположения за биологичната активност въз основа на собствената си концепция за подобие, която обаче не е строго дефинирана и е субективна. Всяка система има собствен критерий за сходство, като различните критерии се оказват подходящи при различни задачи. Следователно сходството на химичните съединения може да се разглежда като свойство, зависещо от контекста на задачата (например разглежданото химично взаимодействие).

1.3. Методи за моделиране

Моделирането на химичните съединения има за цел установяването на връзка между биологичната активност на молекула (или серия от молекули) и други свойства на молекулата (структура, физикохимични свойства и др.). Тъй като в повечето случаи механизмът на биологично действие не е известен, връзката се търси чрез анализ на структурните прилики и разлики за активните и неактивни съединения. Основната хипотеза при изследванията на активността на химичните съединения е, че съединенията имат сходна биологична активност ако притежават общи структурни свойства. Количественият анализ на връзката между химичната структура и активността (*Quantity Structure-Activity Relationship - QSAR*), приложен за първи път от Hansh и сътрудници [17,48,70,79] през 1960-те г., понастоящем 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.

Методите за разпознаване на образи се прилагат с цел да се идентифицира общия образ на реактивност на набор от структурно различни химични съединения, които имат сходна биологична активност. Общият образ представлява множество от подходящи молекулни характеристики. Този подход води до построяване на модел на преобладаващия механизъм на биологично действие, или на вероятността на даден механизъм на действие, поради което е известен като *механистичен* подход. Механизмът обикновено се описва чрез фармакофори - функционални групи и фрагменти и/или техните електронни и стерични свойства, отговорни за молекулните взаимодействия. След като моделът е построен, се прилагат *корелационни* методи [46,78] за количествена оценка на активността в група съединения с общ механизъм на активност (например токсично действие). За количествена оценка обикновено се използуват различни типове молекулни параметри, като физикохимични свойства, биологични свойства, квантово химични електронни или стереохимични параметри, при което се търси тяхната корелация с биологичната активност.

Механистичният и корелационният подход взаимно се допълват. От една страна методите за разпознаване на образи успешно се използуват за полуколичествена оценка на токсичността [61], а от друга страна, корелационният подход разкрива връзката между молекулните дескриптори и начина на токсично действие [44,92]. При предсказването на токсичност на химичните съединения в големи бази данни от разнородни химични съединения, двата подхода взаимно се допълват [46]. Например, метод за разпознаване е необходим, за да се предскаже преобладаващият механизъм на биологично действие и да се разграничат несходните химични съединения, с последващо количествено оценяване на всяка група химични съединения с общ механизъм на действие. При предсказване на токсичност въз основа на сравнение на молекулите с рецептор може да се използуват само методи за разпознаване без да е необходим по-нататъшен корелационен анализ, тъй като самото съответствие на молекулата с рецептора е достатъчно, за да идентифицира молекулата като токсична. Известните методи

COMFA [48], GRID [29], Active Analogue Approach [56] изискват детайлно описание на тримерния модел на рецептора, което не винаги е достъпно.

Широко използувани методи в QSAR анализа са многофакторна линейна регресия, метод на главните компоненти (*Principal Component Analisys*), *Partial Latent Squares* (PLS) , клъстериране по молекулни дескриптори, невронни мрежи.

1.3.1. Линейна регресия

Преобладаващото използуването на многофакторна линейна регресия има своите основания в предположението на Hammet [70,79], 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. Този принцип е известен като *LFER - Linear Free Energy Relationship*. Класическото уравнение на Hansh, което установява връзка между биологичната активност и химичната структура има вида:

$$y = a\pi + b\pi^2 + c\sigma + dE_S + e$$

където y е биологичната активност, $y = log\ (1/C)$, а C е концентрацията на веществото, необходима да се достигне биологичният ефект. Параметрите π, σ, E_S са широко използувани константи, които отразяват хидрофобността, електронни и стерични свойства на съединението. Свободните коефициенти на регресията, които подлежат на оценка, са означени с a,b,c,d,e.

Методът на Hansh е ползотворен подход при търсенето на най-активната молекула в серия от структурно сходни съединения. Счита се, че съществува линейна корелация между промените в молекулната структура, със съответните физикохимични свойства.

Линейната регресия е популярен метод и защото полученото уравнение може да се интерпретира като се обясни физическият смисъл на зависимостта. От друга страна, при получаване на регресионно уравнение с десетки и стотици параметри, физическият смисъл на зависимостта не е съвсем ясен.

1.3.2. Редукция на информацията

Наличието на голям брой параметри, не всички от които оказват влияние върху активността на молекулата, изискват метод за избиране на подходящите параметри. Такива методи са известни в литературата за разпознаване на образи като предварителна обработка (preprocessing) или редукция на информацията. Като пример може да се даде изборът или извличането на свойства (параметри), При избора на параметри, се избират някои от параметрите и те се използуват непроменени. Останалите параметри не се използуват. При извличането на параметри, новите параметри са комбинация от всички съществуващи параметри, като оригиналните параметри не се използуват.

Избор на параметри

Известните алгоритми за избор на параметри се различават според използуваните стратегии за търсене, критерий за избор и условие за край. **Критерият за избор** има за цел да се намери най-доброто подмножество n от възможни N параметъра, където $n<N$. Най-доброто подмножество максимизира критерия по всички възможни комбинации от d на брой свойства. Най-важните критерии за селекция са следните:

Минимална грешка: избира се тази съвкупност от свойства, при която грешката при класификация е минимална. Обикновено пресмятането на аналитичния израз на грешката не е възможно, тъй като трябва да е известна условната плътност на вероятността за всеки клас $p(\mathbf{x}|\omega_i)$. Вместо това се използува нейната оценка.

Разстояние между класовете: Критерият за избор е средното разстояние между всички обекти.

Стратегията за търсене определя начина, по който комбинация от свойства се избират и проверяват по даден критерий. При голям брой параметри *изчерпателното* търсене обикновено изисква много изчислителни ресурси, поради което се използуват алтернативни стратегии:

*Едномерно търсене на "най-добрите" свойства.*Често броят на свойствата е толкова голям, че едновременната оценка на повече от едно свойство е трудоемка и единственият осъществим метод за пресмятане е индивидуален анализ на всяко свойство. За всяко свойство се пресмята неговия критерий. След това свойствата се подреждат в намаляващ ред. Избират се първите *n* свойства, които формират най-доброто множество. Сериозен недостатък на метода е неотчитането на корелации между свойствата.

Последователно избиране е стратегия за търсене, която предлага компромис между бързина и ефективност. Първоначално множеството избрани параметри е празно. На всяка стъпка се избира параметъра, който има най-висок критерий, и се добавя към множеството избрани параметри, докато броя им достигне *n* от всички *N* възможни.

Последователното отхвърляне е аналогично на последователното избиране, но първоначално са избрани всички свойства и след това се премахват едно по едно докато не останат *n* свойства.

Branch & Bound Ако параметрите не отговарят на условието за монотонност, т.е. изтриването на едно свойство не води до намаляване критерия за избор, то може да се приложи оптимизационният алгоритъм Branch & Bound.

Като **критерий за край** често се използува предварително зададен брой на параметрите.

Въпреки че методът за избор на свойства може да отдели значимата от несъществената информация, той има следните недостатъци: първо, влиянието на дадено свойство е абсолютно, т.е. то или се взема предвид или не. Комбинацията на даден параметър с други параметри може да генерира ново свойство, което е по-значимо от първоначалните. Второ, метриката на разстоянията между класовете е ориентирана по осите. Това означава, че промяната на координатната система на пространството на образите може да доведе до друг избор на параметри. Предполага се, че най-добрата стратегия е да се комбинира избора на параметри с

извличане на параметри, т.е. да се трансформира и/или разшири набора параметри и след това да се избере най-доброто множество.

От гледна точка на моделирането на химични съединения изборът на параметри има предимството, че избраните параметри имат физически смисъл и построеният модел може да се използува за обяснение на механизма на биологичния ефект.

Извличане на свойства

Извличане на свойства наричаме метод, при който първоначалните *N* свойства се комбинират, за да се получат нови *n* свойства. Следователно, извличането на свойства представлява трансформация M, която дефинира преобразуването на първоначалния *N*-размерен вектор **x** в нов *n*-размерен вектор **y**=*M(***x***)*. В идеалния случай преобразуването запазва и дори увеличава разделящата информация, с едновременно намаляване на размерността на вектора на свойствата.

Метод на главните компоненти (Principal Component Analisys)

Най-важният метод за линейно извличане на свойства е методът на главните компоненти, известен също и като преобразуване на Karhunen-Loeve[23,31,102,105]. При този метод принадлежността към даден клас не се взема под внимание.

1. Намира се *N*-размерният вектор $\boldsymbol{\mu}= \mathbf{E\{x\}}$ (средното за всички точки от множеството данни)
2. Оценява се глобалната *N* x *N* ковариационна матрица $\boldsymbol{\Sigma}=E\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\}$ за всички *N* точки.
3. Намират се собствените вектори и собствените стойности, като решение на уравнението
$$(\boldsymbol{\Sigma} - \lambda\mathbf{I})\,\mathbf{x} = \mathbf{0},$$
където λ е скаларна величина, **I** е единичната матрица, **0** е нулев вектор. Решението $\mathbf{x}=\mathbf{e}_i$ и съответния скалар $\lambda=\lambda_i$ се наричат собствени вектори и съответно собствени стойности. Ако матрицата Σ е реална и симетрична, то съществуват *N* решения $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N\}$ и *N* собствени стойности $\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$
4. Собствените вектори се подреждат по намаляващ ред на собствените стойности.
5. Избират се първите *k* собствени вектора, съответстващи на най-големите собствени стойности
6. Построява се *N* x *k* матрица **A,** чиито колони се състоят от *k* собствени вектора. Представянето на данните с главните компоненти има смисъл на проекция на данните върху *k*-размерно пространство, съгласно
$$\mathbf{x'} = \mathbf{F_1(x)} = \mathbf{A^t(x-\mu)}$$

Методът на главните компоненти е преобразуване запазващо информацията, ако се използуват всички ненулеви собствени вектори. Новите *k* свойства са ортогонални, т.е. не съществува корелация между тях. Дисперсията на новите

параметри са стойностите λ_j. Отсичането на най-малките собствени стойности не въвежда голяма грешка. Новата размерност k може да се определи въз основа на грешката за реконструкция

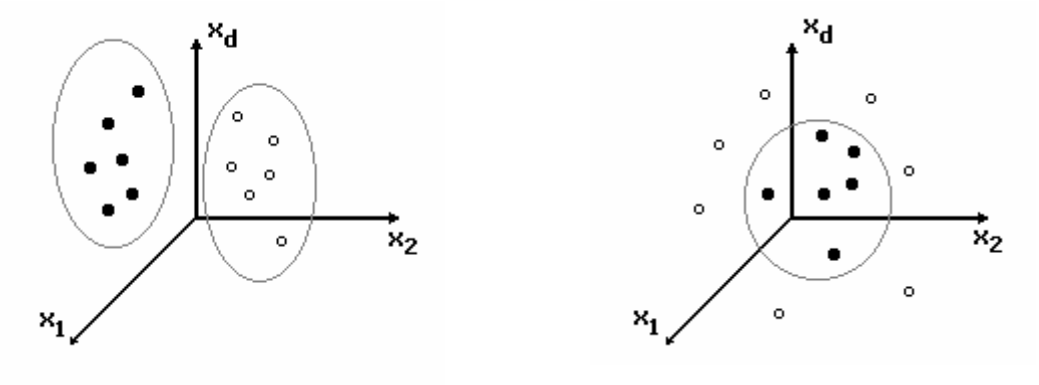
Методът на главните компоненти е най-широко използуваният метод за редукция на информацията при изследванията на активността на химичните съединения.

Обобщен принцип за редукция на информацията

Ако се обобщи принципът за намаляване на размерността, то трябва да се дефинира какво свойство се очаква да се постигне при намаляване на размерността, и след това да се търси начин да се намери тази проекция на данните, която има това свойство [85]. От тази гледна точка, методът на главните компоненти има за цел намиране на проекция, при която дисперсията на данните е най-висока. Въпреки широкото използуване на метода, очевидно не за всички проблеми най-съществения и единствен критерий за описание на данните е високата дисперсия [85,102]. Пример за друг критерий е ентропията, дефинирана от Шенон [85], както и методът, известен като карта на Sammon [82,102], който има за цел да запази разстоянията между обектите. В резултат на нелинейно преобразуване, многомерните данни могат да се преобразуват и визуализират в 2- или 3-мерно пространство. Разстоянието между две точки в оригиналното и редуцираното пространство се запазва. Известни са също методите Kernel Principal Component Analisys, Nonlinear Component Analisys, Independent Component Analisys[23, 31, 69, 80].

1.3.3. Дискриминантен анализ

Изследването на структурата на данните, използувани в QSAR анализа, е довело до разделянето им на симетрична и асиметрична структура [70,107]. Симетричната структура имат данните, които образуват линейно разделими области от активни и неактивни съединения. В този случай могат да се прилагат известните от разпознаването на образи методи за линеен дискриминантен анализ.



Симетрична структура Асиметрична структура

Фигура 1.1

Линейната дискриминационна функция *d(x)* се дефинира чрез *(N+1)*-размерен тегловен вектор **w**, където w_0 има смисъл на праг, а $x_0=1$. Тя представлява *N-1* размерна хиперравнина в *N* размерното пространство на обектите и разделя пространството на две области.

$$d(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0 = 0$$

Вземането на решение за принадлежността към клас се определя като:

векторът **x** *се класифицира към клас* ω_1, *ако* $d(\mathbf{x}) > 0$
векторът **x** *се класифицира към клас* ω_2, *ако* $d(\mathbf{x}) < 0$.

В случай на по-голям брой класове се дефинират частични линейни дискриминантни функции $d_i(\mathbf{x})=\mathrm{w_i}^{\mathbf{t}}\mathbf{x}$, с различни тегловни вектори $\mathbf{w}_i$ за всеки клас. Линейната машина разделя пространството на образите на линейни области и се задава със следното решаващо правило:

векторът **x** *се класифицира към клас* ω_i, *ако* $d_i(\mathbf{x}) > d_j(\mathbf{x}), i, j = 1...c, i \neq j$

Недостатък на този вид класификация е, че не може да се определи степента на принадлежност към дадения клас. Точките в края и центъра на даден клас се разглеждат по един и същ начин.

За намиране на свободните параметри на линейните модели (теглата w_i) се използуват оптимизационни методи, най-разпространен от които е методът на най-бързото спускане. Методите, които използуват само неправилно класифицираните образи за да коригират теглата, се наричат методи за корекция на грешката. Пример за такъв метод е перцептронът [74,81], един от ранните модели на изкуствени невронни мрежи. Друга група методи използуват всички обучаващи

образи за адаптиране на теглата. Целевата функция се дефинира като средната квадратична грешка между желаната класификация *desired(**x**$_p$)* и резултата от решаващото правило *d(**x**$_p$) = decided(**x**$_p$)* за всички *n* образа:

$$MSE = \frac{1}{n}\sum_{p=1}^{n}(desired(\mathbf{x}_p) - decided(\mathbf{x}_p))^2$$

Обучаващото правило има смисъл на итеративно спускане по градиента с цел минимизиране на функция, пропорционална на MSE. Започва се със случайни тегла и на всяка стъпка се добавя корекция, която зависи от градиента на грешката. При достигане на минимум градиентът става нулев и съответното множество от тегла е намерено.

В зависимост от това дали ценовата функция е дефинирана усреднено по всички образи с цел минимизация, или на всяка стъпка образът се избира по случаен начин (стохастично обучение), се извеждат различни обучаващи правила, известни като *delta rule, ADALINE rule, Widrow-Hoff rule, LMS (Least Mean Square) rule* [74, 81, 102, 105].

Обобщени дискриминантни функции

В реалния случай не винаги е възможно класовете да се разделят с линейни функции. Пример за това е т.нар. асиметрична структура на данните[70,107], използувани при моделиране активността на химичните съединени (Фигура 1.1), която се среща в преобладаващата част от случаите. Възможен подход за обобщение е въвеждането на дискриминационни функции от вида

$$d(\mathbf{x}) = \sum_{r=0}^{R} w_r f_r(\mathbf{x}) = 0,$$

където *f(x)* са базисни функции, *R* е броят на базисните функции. При достатъчно голямо *R* и подходящо избрани базисни функции решаващите функции имат желаното свойство на универсални апроксиматори.

За целите на класификацията всеки клас w_i се представя с неговата решаваща функция $d_i(x) = \sum_{r=1}^{R} w_{ri} f_r(x)$, където изборът на клас се осъществява според най-голямата стойност в *c*-размерния вектор *d(**x**)*. Теглата w_{i0} са коефициентите за съответния клас на базисната функция $f_0(\mathbf{x})$, като обикновено $f_0(\mathbf{x})=1$.

Работата на различни методи за класификация (невронни мрежи, нелинейна статистика) може да се моделира чрез използуване на обобщени дискриминационни функции. Примери за това са:

Полиноми[102]	$f_r(x) = x_1^{r_1} x_2^{r_2} \ldots x_D^{r_D}, r_j > 0$
Многослоен перцептрон [31,74,81]	$f_r(\mathbf{x}) = g(\mathbf{w}_r^T \mathbf{x})$, където *g(.)* е нелинейна активираща функция
Radial basis function [31, 73]	$f_r(\mathbf{x}) = \exp(\lVert \mathbf{x} - \mathbf{x}_r \rVert^2 / \sigma_r^2)$, където x_r е локален център, σ_r^2 е е дисперсия

Таблица 1.1

Известни са 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*) [22, 23, 31, 69, 80], но те засега рядко се използуват при анализа на химични съединения.

1.3.4. Вероятностен подход за класификация

Да се класифицира даден обект $\mathbf{x}=(x_1,\ldots,x_N)^t$, означава че е определена неговата принадлежност към един от възможните *c* класове. Нека *d(x)* да е правило за вземане на решение, което определя към кой клас от възможните *c* класа, принадлежи обекта:

$$d : R^N -> \Omega, \qquad \Omega = \{\omega_0, \omega_1, \ldots, \omega_c\}.$$

По този начин $d(\mathbf{x})$ е дефинирана като функция. Ще използуваме означението $d(\mathbf{x})=w_i$, за да обозначим, че *x* е класифициран към клас ω_i. За случаите, когато не може да се вземе решение е въведен специален клас ω_0.

Правилото на Бейс за вземане на решения е теоретичен оптимум за качеството на класификация, като практическите реализации могат само да се доближават до тази граница [102]. Нека да разглеждаме околната среда като случаен генератор на обекти с вероятност *P(.)* и съответната плътност на вероятността *p(.)* в случай на непрекъснати стойности. Генерира се вектор от стойности **x** и неговото значение (принадлежност към клас ω), които се разглеждат като случайни променливи. Нека **x** да е многомерна непрекъсната случайна величина, а ω е дискретна случайна вели-

чина. Величината, необходима за класификация е $p(\omega \mid \mathbf{x})$, т.е. условната вероятност обектът да принадлежи на клас ω, при условие, че има свойства **x** (апостериорна вероятност). Според теоремата на Бейс:

$$p(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{\sum_{i=1}^{c} p(\mathbf{x} \mid \omega_i)P(\omega_i)},$$

където $P(\omega)$ е априорната вероятност за клас ω, c е броят на класовете.

Ако е известен законът (съвместната плътност на вероятността), по който се генерират обектите $p(\mathbf{x}, \omega)$, то класификацията е тривиална.

Правилото на Бейс за класификация с минимална грешка гласи, че

векторът **x** *се класифицира към клас* ω_i, *ако*
$$P(\omega_i \mid x) > P(\omega_j \mid x), j = 1...c, i \neq j$$

което означава, че се избира клас, който има най-високата апостериорна вероятност за **x**. По този начин се минимизира средната вероятност за грешка. За грешка при класификация се счита вземането на решение за принадлежност към клас ω_j, въпреки, че апостериорната вероятност за друг клас ω_i е по-голяма. **Правилото за класификация с минимална грешка и възможност за отхвърляне** се въвежда аналогично, като се дефинира праг за отхвърляне.

В теоремата на Бейс решаващото правило не зависи от нормиращия знаменател $p(\mathbf{x})$. Тогава **правилото на Бейс за класификация с минимално правдоподобие** може да се запише като:

векторът **x** *се класифицира към клас* ω_i, *ако*
$$P(\mathbf{x} \mid \omega_i)P(\omega_i) > P(\mathbf{x} \mid \omega_j)P(\omega_j), j = 1...c, i \neq j$$

Условната плътност на вероятността $p(\mathbf{x}|\omega_i)$ се нарича още правдоподобност на класа ω_i при даден вектор **x**, а отношението $p(\mathbf{x}|\omega_i)/ p(\mathbf{x}|\omega_j)$ се нарича отношение на правдоподобност. В случаите, когато априори вероятността на класовете е неизвестна, класификацията може да се извърши само по стойностите за правдоподобност.

Правило на Бейс за класификация с минимален риск е най-общият случай за вземане на решение по Бейс. Рискът е очакваната стойност на цената на решението. Очакваната загуба или условен риск за класификатора при взето решение, че **x** принадлежи на ω_i е:

$$R_i(\mathbf{x}) = \sum_{j=1}^{c} \lambda(\omega_i \mid \omega_j) P(\omega_j \mid \mathbf{x}), i = 1 \ldots c$$

Условният риск за вземане на решение е:

$$R(x) = \sum_{j=1}^{c} \lambda(d(\mathbf{x}) \mid \omega_j) P(\omega_j \mid \mathbf{x})$$

Средният риск по цялото пространство на обектите е математическото очакване $E\{\mathrm{R}(\mathbf{x})\}$ на условния риск $R(\mathbf{x})$. Ако правилото $d(\mathbf{x})$ е оптимално, то трябва да минимизира средния риск. Интегралът се минимизира, като винаги се избира минималната стойност на $R(\mathbf{x})$ за всяка възможна стойност на **x**. Това се постига с решаващото правило на Бейс за минимален риск:

векторът **x** *се класифицира към клас* ω_i, *ако*
$$R_i(\mathbf{x}) > R_j(\mathbf{x}), j = 1 \ldots c, i \neq j, i \in \{0..c\}$$

1.3.5. Видове класификатори. Обучение с учител

Известни са различни модели на класификатори и методи за оценка на свободните параметри. Някои модели се опитват да оценят апостериорните вероятности $p(\omega_i|\mathbf{x})$ директно от данните, например класификация по най-близкия съсед. При други методи се прави оценка на условната плътност на вероятността $p(\mathbf{x}|\omega_i)$. Оценка на плътността на вероятността p означава намиране на подходяща функция $f(\mathbf{x})$, която да описва наличните данни. Методите за оценка са параметрични и непараметрични. Параметричните методи изискват предположение за вида на функцията на плътност на вероятността, и последваща оценка на параметрите на функцията. Поради това параметричните методи крият опасността от допускане на грешка първо при предположението на вида на функцията, и второ, 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.

Параметрични методи

При използването на параметрични методи се предполага че условната вероятност има вида на някое от известните разпределения (нормално, Поасоново, експоненциално, Г,β). Класически пример е използването на многомерно гаусово разпределение. Ако параметрите на разпределенията са известни или са коректно оценени, то може да се достигне оптимална класификация по Бейс, тъй като класификацията по Бейс е независима от вида на разпределнието.

Оценката на свободните параметри се прави въз основа на обучаващите данни $\{(x_p, \omega_p)\}$, $p=1,\ldots,n$, чрез методи за максимално правдоподобие [22, 23, 31, 85, 105]. Очевидният недостатък на параметричните класификатори е възможното несъответствие на данните с избраното разпределение. В този случай изчислената апостериорна вероятност ще бъде невярна. Друг проблем е свързан с евентуалния голям брой свободни параметри при наличието на малък брой данни.

Непараметрични методи

Хистограмата е най-простият начин за оценка на вероятностната плътност. Ако е дадена извадката $x_1, x_2, \ldots, x_n$, при зададено начало x_0 и ширина на интервала h, хистограмата се построява чрез разделяне на интервали $[x_0 + mh, x_0 + (m+1)h)$, където m е цяло число[84, 85]. Хистограмата се дефинира като

$$f_H(x) = \frac{1}{nh} \nu_i ,$$

където ν_i е броя на x_i в този интервал, на който принадлежи x.

Недостатък на хистограмата като подход за представяне и анализ на данните е необходимостта от предварителен избор на началото x_0 и ширината на интервала h. Една приближена оценка за ширината на интервалите е $h = 3.5\sigma n^{-1/3}$, където σ е оценка за дисперсията[10,85]. Друг недостатък, е , че хистограмите, построени въз основа на двумерни и тримерни данни са трудни за възприемане и не позволяват лесно да се построят контурни диаграми.

Naive estimator

Ако случайната променлива *X* има плътност на вероятността *f*, то :

$$f(x) = \lim_{h->0} \frac{1}{2h} P(x-h < X < x+h)$$

За всяко *h* може да се оцени *P(x-h<X<x+h)*, като пропорцията на точките, принадлежащи на интервал *(x-h,x+h)*. По този начин, естествената оценка на плътността на вероятност се получава при избор на малка стойност на *h* и по формулата (*naive estimator* [85]):

$$\hat{f}(x) = \frac{1}{2hn}[\textit{броя} \quad \textit{на} \quad X_1,...X_n \quad \textit{принадлежащи} \quad \textit{на} \quad (x-h, x+h)] \qquad (1.1)$$

Ако дефинираме тегловна функция *w,* такава, че:

$$w(x) = \begin{cases} \frac{1}{2}, & \textit{ако} \quad |x| < 1 \\ 1, & \textit{ако} \quad |x| > 1 \end{cases} \qquad (1.2)$$

то формулата 1.1 може да се запише като:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} w(\frac{x - X_i}{h}) \qquad (1.3)$$

От (1.2) следва, че оценката на плътността на вероятност се получава чрез "поставяне" на правоъгълник с ширина *2h* и височина *(2nh)*$^{-1}$ във всяка точка и сумирането им. Недостатък на получената оценка е, че функцията не е непрекъсната.

Оценка на плътността на вероятност с кернел функции

При заместване на тегловната функция *w* с кернел функция φ, която удовлетворява условието

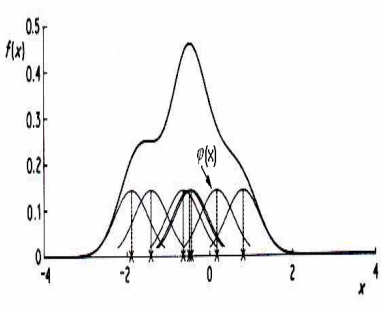
$$\int_{-\infty}^{\infty} \varphi(x)dx = 1,$$

получаваме изследвания метод за оценка на плътността чрез кернел функции (kernel estimator, Parzen Windows) [22, 23, 85, 102], дефиниран като:

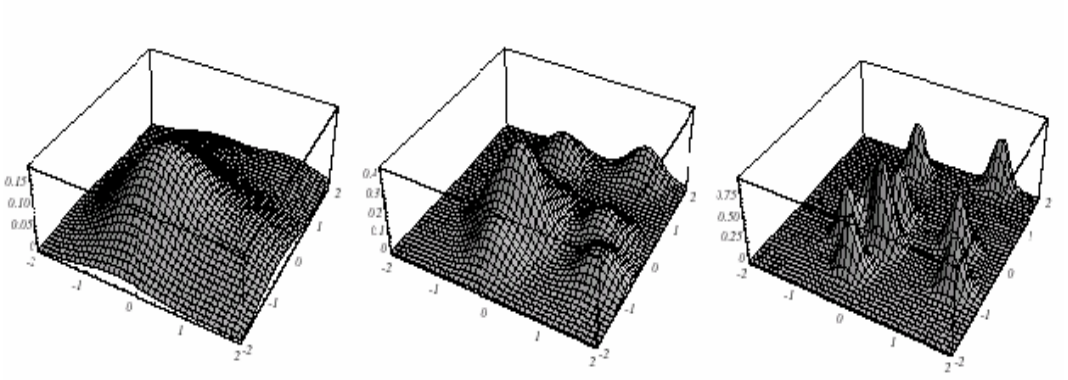
$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} \varphi\left(\frac{x - X_i}{h}\right), \qquad (1.4)$$

където *h* е изглаждащ параметър (*window width, smoothing parameter, bandwidth)*.

Оценката с кернел функции представлява смес от *n* плътности, центрирани в точките от извадката (Фигура 1.2). Това означава, че за голям размер на извадката директното изчисление ще бъде много трудоемко. Ефективен метод за изчисление на оценката с кернел функции е предварително разделяне на интервали и използуване на бързо преобразуване на Фурие[2, 3, 85].



Фигура 1.2



Фигура 1.3

Оценката с кернел функции има естествено разширение в двумерния и многомерния случай (Фигура 1.3):

$$\hat{f}(x,y) = \frac{1}{h_x h_y} \sum_{i=1}^{n} \varphi\left(\frac{x - x_i}{h_x}, \frac{y - y_i}{h_y} \right)$$

Кернел функциите могат да се разделят на две групи – с неограничена и локална дефиниционна област на прозоречната функция φ. Кернел функции с неограничена дефиниционна област са Гаусовият, експоненциален и лоренцов кернел. Прозоречната функция има ненулеви стойности по цялото пространство на обектите и дори точки далеч от центъра допринасят за глобалната плътност.

Кернел функции с локална поддръжка са хиперкубичен, хиперсферичен и хипертриъгълен кернел.

Тип на кернел функцията		Метрика
Хиперкубичен	$\begin{cases} \varphi(x, x_{ik}) = 1, ако \quad \delta_C(x, x_{ik}) < \rho \\ \varphi(x, x_{ik}) = 0, в \quad противен \quad случай \end{cases}$	Разстояние на Чебишов $\delta_C(\mathbf{x}, \mathbf{x}_{ik}) = \max\{\lvert \mathbf{x}_j - \mathbf{x}_{ikj} \rvert\}, \; j = 1,...,D$
Хиперсферичен	$\begin{cases} \varphi(x, x_{ik}) = 1, ако \quad \delta_E(x, x_{ik}) < \rho \\ \varphi(x, x_{ik}) = 0, в \quad противен \quad случай \end{cases}$	Евклидово $\delta_E(x, x_{ik}) = \sqrt{(x_j - x_{ik})^2}$
Гаусов	$\varphi_G(x, x_{ik}) = [q_i (2\pi\rho)^D]^{-\frac{1}{2}} \exp(\frac{-\delta_E(x, x_{ik})}{(2\rho^2)}),$ $q_i = \frac{1}{n_i}\sum_{j=1}^{D}(x_{ikj} - \mu_{ij})^2; \mu_i = \frac{1}{n_i}\sum_{k=1}^{n_i} x_{ik}$	Квадратично $\delta_Q(x, x_{ik}) = (x_j - x_{ik})^T Q(x_j - x_{ik}), \; j = 1,...,D$
Експоненциален	$\varphi(x, x_{ik}) = \frac{1}{2\rho}\exp(\frac{-\delta_E(x, x_{ik})}{\rho^2})$	Евклидово
Лоренцов	$\varphi(x, x_{ik}) = \frac{\pi}{1 + (\frac{\delta_E(x, x_{ik})}{\rho})^2}$	Евклидово
Хипертриъгълен	$\begin{cases} \varphi(x, x_{ik}) = 4(\rho - \delta_E(x, x_{ik})), ако \quad \delta_E(x, x_{ik}) < \rho \\ \varphi(x, x_{ik}) = 0, в \quad противен \quad случай \end{cases}$	Евклидово

Таблица 1.2

Недостатък на метода е необходимостта от избор на свободните параметри - вида на функцията φ *и* изглаждащия параметър *h*. Изследванията показват, че видът на функцията не е от голямо значение (оптималната кернел функция е кернел на Епанечников [85], но всички други типове кернел функции са близки до оптималния). Параметърът *h* може да се разглежда и като размер на "прозореца" на кернел функцията. Ако той има голяма стойност, то влиянието на всеки обект ще се разпространява на повече съседни обекти, докато ако е с ниска стойност, обектът ще има голямо влияние върху стойността на плътността (Фигура 1.3). За преодоляване на този проблем са разработени множество методи за избор на изглаждащия параметър *h*, на базата максимално правдоподобие, информационни оценки, кръстосана проверка за достоверност и други [23,85,57,84]. Съвременните методи включват използуване на нееднакви изглаждащи параметри за всяка точка (*adaptive kernel estimation*), избор на wavelet, сплайни и други в качеството на кернел функции. Непараметричните методи за оценка на вероятностната плътност

намират широко приложение при непараметричен дискриминантен анализ, клъстерен анализ, класификация и други [10, 23, 55, 80, 84, 85, 57].

Класификация по най-близкия съсед

Интуитивният подход за класификация е да се изчисли разстоянието от непознат образ до всички останали и да се избере класът на тази точка, която е най-близо. Решаващото правило за класификация по най-близкия съсед е [102]:

векторът **x** *се класифицира към клас* w_i, *ако*

$$\delta(\mathbf{x}, \mathbf{x}_p) < \delta(\mathbf{x}, \mathbf{x}_q), \mathbf{x}_p \in \omega_i, p, q = 1,\ldots n, p \neq q$$

Класификаторът по най-близкия съсед оценява директно апостериорната вероятност $P(\omega_i|\mathbf{x})$. Предимството му е неговата простота и ограничената грешка, чиято максимална стойност е два пъти грешката при оптималната класификация по Бейс[102]. Това е теоретическа граница и може да бъде достигната само при безкраен размер на обучаващото множество.Въпреки това, класификаторът дава добри практически ресултати и може да бъде използуван за качествена оценка на входните данни. Например, ако класификаторът по най-близък съсед показва грешка от 15% при достатъчно количество данни, то оптималният класификатор по Бейс никога няма да постигне безгрешна класификация. Като недостатъци могат да се отбележи необходимостта да се сравнява неизвестния обект със всички останали, което може да забави класификацията при голям брой данни.

Класификаторът по най-близкия съсед може да бъде разширен до класификатор по най-близките *k*-съседа. Изчислява се принадлежността към клас на най-близките *k*-съседа на неизвестния обект **x**. Най често срещания клас в тези *k*-съседа се взема за клас на неизвестния обект. Този класификатор трябва да се използува само ако броят на обектите е достатъчно голям.

Класификация по най-близкия прототип

В този случай се използуват x_r представители на всеки клас ω_i и се извършва класификация по най-близкия съсед. Проблемът се свежда до определяне на прототипите. Пример за подобен алгоритъм е Learning Vector Quantization (LVQ) на Кохонен [47].

1.3.6. Обучение без учител и клъстериране

В разгледаните до сега методи се предполага че са известни броят на класовете и принадлежността на обучаващите данни към даден клас.В повечето реални задачи тази информация не е известна и трябва да се открие въз основа структурата на данните. Областите с данни, които съдържат по-голям брой обекти от останалите се наричат клъстери. Клъстерите играят ролята на класове в обучението без учител.Целта е да се намери описание на данните като смес от различни условни разпределения за всеки клъстер $p(\mathbf{x}|\omega_i)$. При c различни клъстера $\omega_1..\omega_C$:

$$p(\mathbf{x}) = \sum_{i=1}^{c} p(\mathbf{x} \mid \omega_i) P(\omega_i)$$

Задачата, която трябва да се реши е разделяне на модите на функцията за плътност на вероятността. Съществуват директни и индиректни (клъстериране) методи за разделяне на модите. Клъстериращите методи могат да бъдат итеративни и йерархични [102].

Пример за метод на клъстериране е т.нар. *c-means clustering*, който представлява динамична клъстериращa процедура, където всеки клъстер се представя с неговото средно μ_i. Принадлежността на точка x_{ik} към клас, може да се променя в процеса на работа на алгоритъма. Процедурата за клъстеризация намира оптималните клъстери като минимизира критерия за качеството на клъстеризацията [105,23]:

Алгоритъм
1)задават се броят на точките n , брой на клъстерите c ,изчисляват се центровете на клъстерите $\mu_1, \mu_2, \ldots \mu_c$
2)класифицират се всички n точки според близостта им до центровете μ_i
3) изчисляват се отново центровете на клъстерите $\mu_1, \mu_2, \ldots \mu_c$ според новата класификация
4)повтаря се 2) докато центровете μ_i се променят
5) клъстерите се определят според намерените центрове $\mu_1, \mu_2, \ldots \mu_c$

Клъстериране може да се използува като предварителна обработка на данните. По-сложен клъстериращ алгоритъм е алгоритъмът *ISODATA* [105]. Той позволява изтриване на клъстери, ако в клъстера няма достатъчно точки и сливане на клъстери, ако те са достатъчно близо. Обучение без учител може да се постигне и с методи от изкуствените невронни мрежи, например самоорганизиращи се мрежи [74].

Клъстериране на химични съединения

Поради възприетата хипотеза в молекулното моделиране, че съединенията с подобно биологично действие имат близки стойности на стереоелектронните параметри, създаването на ефективни методи за клъстериране на голям брой химични съединения е от особено значение. Известните методи за клъстериране могат да се прилагат при избрана метрика за сходство между съединенията. Дефинирането на метрика за сходство между съединенията, само по себе си е сложна задача. Предложени са различни методи, както въз основа на теория на графите, така и въз основа на пространствено разделяне [104]. Сложността им варира от O(n) до $O(n^3)$. Някои от популярните методи за клъстериране на химични съединения са:

*Геометрични методи.*Разглеждат се набор от молекулни дескриптори и се разделя пространството на дескрипторите на интервали. Съединенията се разпределят по съответните интервали според стойностите на техните дескриптори. Тези методи имат сложност O(n) и са подходящи за големи бази данни, но клъстерирането зависи от избраното разделяне на интервали

Метод на Jarvis-Patrick.

> За всеки обект се намират неговите най-близки *J* съседа. Това изисква N^2 сравнения при наличие на *N* обекта. Два обекта ще принадлежат на общ клъстер, ако са изпълнени условията:
> а)Всеки от тях е в списъка на другия обект за най-близките *J* съседа
> б)имат *K* общи съседа от техните *J* най-близки съседа.

Методите за разделяне на Jarvis-Patrick и Ward разглеждат всички разстояния между структурите и извеждат списък на подобните съединения, които формират клъстерите. Сложността на тези методи е $O(n^2)$ и не са подходящи за много големи множества от съединения.

1.3.7. Системи с дърво на решенията

Построяването на дърво на решението (*Decision Tree*) е стандартен метод в системите за извличане на знания и класификация. Известни са алгоритмите *CART*, *ID3*, *C4.5* [32, 34, 36, 8]. 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.

Дървото на решение представлява дърво с корен, обикновено двоично, с прости класификатори във всеки вътрешен възел и решения за класифициране в крайните възли. За да се оцени дадено дърво T по отношение на входни данни $\mathbf{x}$, всеки класификатор в дървото се оценява според аргумента $\mathbf{x}$. В някои алгоритми за мярка на информативността на класификатора се използува информационната ентропия, според дефиницията на Шенон. Резултатът от класификатора определя уникалния път от корена до крайния възел. Във всеки вътрешен възел пътят следва левия клон, ако резултатът от функцията е $+1$ или десния, ако резултатът е -1. Този път се нарича път за оценка. Стойността на функцията $T(x)$, получена в крайния възел е резултатът от класификацията.

Дървото на решението се получава от обучаваща процедура, която започва от корена на дървото и постъпково избира разделяне на данните, което максимизира дадена ценова функция. След избора на разделяне, се намира съответствието на данните към двата клона и процедурата продължава рекурсивно докато се удовлетвори зададен критерий за край. След това дървото се подлага на "окастряне" чрез процедура за търсене от листата към корените с цел премахване на излишните възли.

Обикновено класификаторите във вътрешните възли на дървото представляват сравнения на входните параметри с праг. По този начин изходното пространство се разделя на области паралелни на координатните оси. Стандартният алгоритъм за този начин на обучение е C4.5 [32]. Друга възможна стратегия, която е реализирана в OC1 е да се използуват хиперравнини в произволно положение.

Системите с дърво на решенията са привлекателни за приложение при изследванията на активността на химичните съединения. Като пример за реализация може да се посочи системата Oncologic [72], която дава оценка за потенциалната канцерогенност на няколко групи химични съединения. Потребителят избира подходящата група (органични съединения, метали и т.н.) и отговаря на

последователност от въпроси на системата, които водят до заключение за канцерогенността на веществото. 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.

1.3.8. Генетични алгоритми

Генетичните алгоритми са разработени по аналогия с естествения еволюционен процес. В "Произхода на видовете" Дарвин отбелязва, че живите организми се адаптират към околната среда и предполага че това се дължи на следния механизъм. Всеки добре приспособен към средата организъм живее по-дълго и ражда по-добро потомство, което наследява неговата приспособеност. С течение на времето, приспособеността към средата нараства.

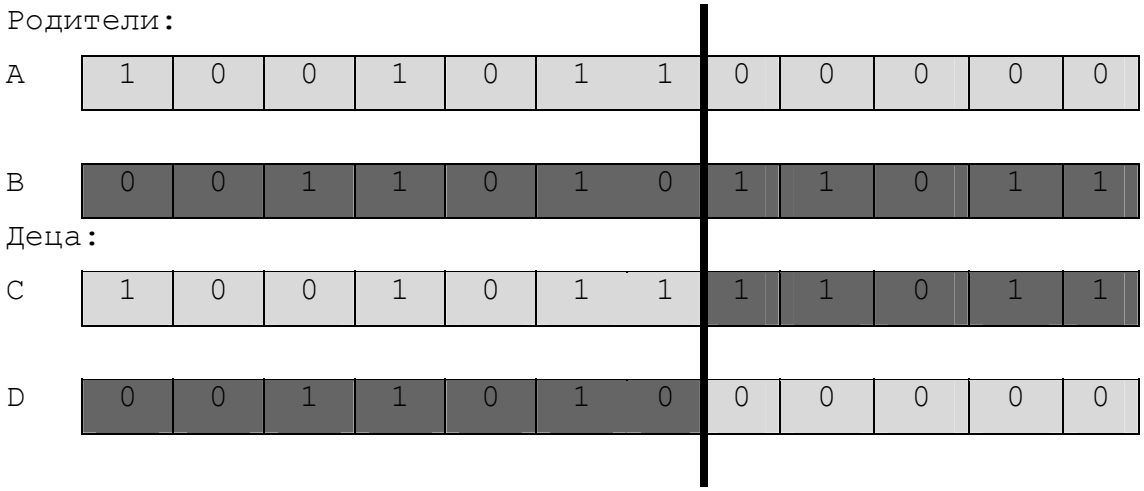
Генетичният алгоритъм е итеративна процедура, която се прилага върху популация от зададен постоянен брой индивиди. Всеки индивид се представя с низ от символи, наречени *геноми*, кодиращи дадена точка (решение) в пространството на задачата. Пространството на задачата съдържа всички възможни решения на проблема. Генетичните алгоритми се прилагат обикновено за задачи, където пространството за търсене е твърде обширно, за да бъде проведено изчерпателно търсене. Низовете обикновено се състоят от двоични символи, но могат да се използуват и други представяния.

Стандартният генетичен алгоритъм се състои от следните стъпки:

- Генерира се първоначална популация, като индивидите се генерират случайно или чрез подходяща за проблема евристика;
- Всяка еволюционна стъпка, наречена *генерация*, се декодира и оценява според зададен критерий (оценъчна функция);
- Индивидите, които ще образуват новата популация (следващата генерация), се *избират* според стойността на оценъчната функция. Използуват се различни процедури за избор, като най-разпространената е избор с вероятност пропорционална на стойността на оценъчната функция.

Изборът не въвежда нови индивиди в популацията (т.е. нови точки в пространството на решенията). Те се генерират чрез *генетични операции*, известни

като *кръстосване* и *мутация*. Кръстосването се извършва с вероятност p_{cross} между два избрани индивида, наречени *родители*, чрез размяна на части от техния геном, при което се получават два нови индивида, наречени *деца*. При най-простата форма на кръстосване, поднизовете се разменят след разцепване на случайно избрано място. На Фигура 1.4. е илюстрирана на операцията "кръстосване":



Фигура 1.4.

- Операцията *мутация* се въвежда за да се избегне попадането в локален минимум, чрез въвеждане на случайни нови точки от пространството на решенията. Тази операция се извършва, като се променят символите по случаен начин с (малка) вероятност p_{mut}. Генетичните алгоритми са стохастични итеративни процеси, чиято сходимост не е гарантирана. Условието за край трябва да се зададе като максимален брой генерации или стойност на оценъчната функция

Генетичните алгоритми предоставят недетерминистично решение на комбинаторни оптимизационни проблеми. Приложенията на генетични алгоритми в конформационния анализ имат за цел да оптимизират генерацията на конформери според зададени геометрични/стерични условия. Повечето от приложенията на генетични алгоритми в конформационния анализ намират само ротамери. За да се

изследва гъвкавостта на циклични фрагменти Glen и Payne[75] включват и обръщане (отражение) на т.нар. "свободни ъгли", които включват ротации около 2 връзки в цикъла.

1.4. 3D QSAR

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. Върху получените огромен брой независими променливи не може да се приложи регресионен анализ. За целта обикновено се прилага извличане на главните компоненти, като по-нататък се търси тяхната корелация с биологичната активност.

1.4.1. Метод CoMFA - Comparative Molecular Field Analisys

Методът *CoMFA* [48] е представител на 3D QSAR методите. Получената зависимост се нарича *CoMFA* модел. Молекулите се сравняват чрез налагането им в пространството и съпоставянето на техните полета в 3D решетка. Необходимото изискване при моделиране е множеството от молекули, подлежащи на анализ, да взаимодействуват с подобни рецептори по сходен начин. От тях се избира подмножество от молекули, което ще се използува като обучаващо множество за извеждането на CoMFA модела. Останалите структури се използуват за оценка на валидността на изведения модел.

Генерират се една или повече нискоенергетични конформации и се изчисляват зарядите на атомите. Прави се предположение за фармакофора, за да се намери подходящо разположение и взаимна ориентация на всички структури.

> Използуването на общ фармакофор за серия съединения не означава, че съединенията трябва да бъдат структурно сходни. Напротив, най-важното предимство на CoMFA метода е, че молекули с еднакви фармакофори, но различна топология, могат да бъдат анализирани едновременно. Пространственото съпоставяне на всички молекули се извършва според еквивалентни функционални групи на ръка или с подходящ алгоритъм за силово поле, като се цели максимално припокриване. Подходящ начин за съпоставяне е реализиран в програмата *SEAL* [48]. Той дефинира "коефициент на подобие" A_F между молекулите A и B при произволна взаимна ориентация. За тази цел се изчисляват стойностите на подобие на атоми между всеки от m-те атома на молекулата A с всеки от n-те атома на молекулата B. Сумата на тези стойности по всички двойки атоми дава коефициента на подобие A_F:

$$A_F = -\sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij} e^{-\alpha r_{ij}^2},$$

$$w_{ij} = w_E q_i q_j + w_S v_i v_j + \ldots$$

където r_{ij} е разстоянието между атомите i и j, α е дефиниран от потребителя коефициент, w_E и w_S са дефинирани от потребителя коефициенти за отчитане на електростатичните и стерични ефекти, q_i е зарядът на атом i, v_i е Ван дер Ваалсовия радиус на атом i. Към дефиницията на w_{ij} могат да бъдат добавени и други параметри, ако е необходимо да се отчете техния ефект при съпоставяне на молекулите. Горното равенство дефинира камбановидна Гаусова крива. Поради експоненциалната зависимост, най-доброто сходство ще се получи, ако съответните атоми от двете молекули са най-близо един до друг. Ако някои атоми са вече поставени един до друг, то малки измествания с цел останалите атоми да се изравнят няма да доведат до големи промени в коефициента на подобие. Този индекс позволява автоматично съпоставяне, без да е необходима предварителна ориентация и/или дефиниране на фармакофор.

След изравняване на молекулите в простанството около тях се дефинира решетка, като обикновено стъпката на решетката е около 2 ангстрьома (1Å= m^{-10}). Тази стойност е твърде висока, като се има предвид, че дори разстояния от порядъка на десети от Å водят до различие във взаимодействието между атомите. Използуването на по-ситна мрежа, обаче, изисква много по-големи изчислителни ресурси. Във всеки възел на мрежата се поставя пробен атом и се пресмята стойността на полето в тази точка.

След генерацията на набор от конформации за всяка от тях се изчисляват зарядите на атомите. Те се използуват за да се изчисли за всяка молекула поотделно електростатичното поле във всички точки на мрежата, използувайки зареден пробен атом. В метода CoMSIA вместо електростатичното поле се изчислява коефициентът на подобие с пробния атом.

Тези стойности се записват в таблици, които обикновено съдържат няколко хиляди колони и трябва да се намери зависимостта между тях и дадена биологична активност. За тази цел се използува частично линейна регресия (PLSR - Partial Least Squares Regression). Качеството на предсказване на модела се оценява с кръстосана проверка за достоверност.

Извеждане на 3D QSAR модела

Както и в регресионния анализ, при PLS анализа корелационният коефициент нараства с нарастването на броя на извлечените вектори. В зависимост от броя на компонентите може да се достигне перфектна корелация, поради големия брой на независимите променливи. Следователно, коефициентът на корелация не е критерий за валидността на PLS модела. В повечето случаи се използува кръстосана проверка за достоверност, като една от структурите се премахва от обучаващото множество и PLS моделът се извежда от останалите структури. Този модел се използува за да се предскаже биологичната активност на изключената структура. Тази процедура се повтаря за всяка една структура от обучаващото множество. Сумата

$$PRESS = \sum (y_{predicted} - y_{observed})^2,$$

се използува като мярка за предсказващата способност на изведения модел. За големи множества от данни се препоръчва k-leave-out процедура, за да се постигне по-добра стабилност на модела.
По аналогия с коефициента на корелация се дефинира

$$Q^2 = \frac{(SD - PRESS)}{PRESS},$$

$$SD = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

Стойностите на Q^2 са винаги по-малки, отколкото стойностите на r^2, при включване на всички обекти. Докато в PLS модела се включват значиими вектори, Q^2 се увеличава, а намаляването на Q^2 означава overprediction. В по-сериозни случаи на overprediction могат да се получат отрицателни стойности за Q^2, което означава, че моделът е по-лош, отколкото ако за предсказани се вземат средните стойности на y. Резултатите от PLS анализа могат да бъдат трансформирани в регресионни коефициенти на независимите променливи.

Резултатът от анализа е регресионно уравнение с голям брой коефициенти. Обикновено той се илюстрира с контурни карти, които показват благоприятните и неблагоприятните райони за съответните заместители. Предсказването за тестовите съединения може да се направи или качествено чрез визуална оценка на контурните карти или количествено чрез пресмятане на полетата за тези съединения и заместването на тези стойности в PLS модела.

Предимства и недостатъци

- Съпоставянето на структурите създава проблеми, особено при изследване на молекули, различаващи се по своята топология
- При метода *CoMFA* се генерират огромен брой дескриптори за всяка молекула, докато обикновено съществуват данните за биологичната активност за 10 - 30 структури. Методът *PLS* може да обработва такива данни, но е чувствителен към зашумени данни. Освен това не всички параметри могат да имат отношение към изследваните свойства. За QSAR изследванията е необходим метод, които да анализира данни с огромна размерност и да не взема предвид дескрипторите, които не са свързани с изследваното свойство.
- За отчитане на конформационната гъвкавост трябва предварително да се генерира набор от конформери.

Към предимствата могат да се отбележат възможността за визуализация на стерични и електростатични потенциали.

1.5. Молекулен дизайн

Основната задача на QSAR анализа е откриване на общите характеристики, обуславящи реакционната способност, с цел предсказване на съответната биологична активност. Резултатите от изследванията на връзката между химичната структура и активността дават възможност за:

- Изясняване на механизмите водещи до конкретната биологична активност;
- Предсказване на токсичност и други биологични свойства и оценка на замърсяването на околната среда;
- Търсене на структури с предварително зададени свойства (лекарства, агрохимикали, други) сред известните химични съединения;
- Компютърно генериране на нови структури със зададени свойства и последващото им експериментално синтезиране (de novo design).

Компютърният дизайн на нови лекарства и други химични съединения с предварително зададени свойства включва следните задачи: оптимизация и генериране на лидери. Първият, или наричан още корелационен SAR (структура-активност) анализ, се състои в постъпкова модификация на известни естествени или изкуствени съединения, имащи желания ефект, което ограничава подхода в намирането на структури близки до изходната структура. Генерирането на лидери, от своя страна, включва два главни подхода. Първият е търсене в бази данни на съществуващи структури, които да удовлетворяват зададени структурни изисквания за активност, докато при втория задачата е да се генерират нови структури при тези условия. Методите с които се генерират нови структури са известни под името de novo методи [12].

1.5.1. Методи за генериране на нови структури

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). Наличието на информация за структурата на рецепторите въз основа на кристалографски анализ или ядреномагнитен резонанс е важен стимул за разработването им. При търсенето в бази данни целта е да се открие структура, която да съответствува най-добре на рецептора. В случая на пораждане на нови структури са разработени различни алгоритми за построяване на новата структура, така че също да съответствува най-добре на рецептора. Намирането на най-добро съответствие е известно като "прикачване" (docking). Прикачването се дефинира (по аналогия на акостирането на кораб в док) като поставяне на структурата в подходяща конфигурация за взаимодействие с рецептора и има за цел 1)изучаването на геометрията на комплекса; 2)изучаване свързването на една структура към няколко макромолекули; 3) дава възможност за предлагане на модификации на структурата; 4) възможност за оценка на съпоставимостта на структурата и рецептора.

Методите за генериране, изискващи наличието на информация за пространствената структура на съединението и рецептора се наричат *директни*. Тази информация обаче често не съществува или не е достъпна, тъй като е собственост на фармацевтични или други компании. От друга стране, оптималното съответствие не гарантира че съответната структура ще има максимална биологична активност и минимални странични ефекти.

Индиректнте 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 с разгледаните в предишните точки методи се извлича информация за фармакофора, или за образа на реакционната способност, който по-нататък се използува за търсене или генериране на нови структури.

Всяка генерирана структура подлежи на *оценка*, доколо е подходяща. Използуват се различни типове оценъчни функции (енергетични и основани на правила) и не съществува съгласие коя е най-добра. *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 извър-

шва 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 достатъчна информация, като 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 от тях се избира 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 предизвиква нежеланието да се синтезират новите структури само защото дадена структура е генерирана от компютърна програма. Другата съществена причина е, че изследванията се правят от фармацевтични компании и изследователите не публикуват най-интересните резултати и рецептори. Съществуват малък брой публикувани резултати на синтезирани химични съединения, предложени от de novo дизайн.

Методи за молекулен дизайн основан на анализ на активните центрове

Активният център е съвкупност от атоми или атомни групи в рецептора, осигуряващи взаимодействието между рецептора и химичното съединение. Методът има за цел да определи кои атоми и функционални групи от евентуалната нова структура ще реагират най-добре с активните центрове в рецептора и да намери предпочитаното им разположение. По този начин не се получават нови структури, а само позициониране на прости фрагменти в близост до активните центрове. Примери за

реализация са програмите GRID [29], GREEN, HSITE, MCSS, някои Monte-Carlo или методи за симулирано закаляване, HINT, BUCKETS.

Методите, анализиращи активните центрове, се основават на идеята, че малък брой от подходящо разположени фрагменти могат да осигурят значителна част от енергията на свързване, следователно е разумно първо да се намерят тези фрагменти и техните позиции. При увеличаването на разнообразието на началните фрагменти се увеличава вероятността да се намери синтезируема структура.

Методи, които прикачват цели молекули

Тези методи също не генерират нови структури, а само съпоставят структурата с активния център. Те обаче са съществена част от генерирането на нови структури и могат да се използуват за тяхната оценка. Всяка структура (конструирана чрез друг метод или прочетена от база данни) се съпоставя с модела на фармакофора или се прави опит да се разположи в активния център на рецептора. Първата програмна реализация на такъв метод е DOCK [12,91]. DOCK използува съпоставяне по форма, като търси голям брой различни ориентации на лигандите в активния център, но може да бъде и конфигуриран за да използува и електростатична оценка. DOCK е разработена за търсене в бази данни, съдържащи хиляди съединения. Известни са реализации на метода, които отчитат гъвкавостта, като позволяват конформациите да се променят при прикачването, докато други работят с набор от предварително записани конформации и твърдо прикачват всяка от тях. AUTODOCK[12] е програма, която прави опит за гъвкаво съпоставяне на молекулата с рецептора чрез прилагането на комбинации от Monte Carlo и simulated annealing методи за търсене. В последните години се използуват генетични алгоритми за гъвкаво прикачване [41]. Те са обикновено по-бавни от използуваните твърди структури, но успешно се справят с намирането на нискоенергетични активни конформации.

Предимството на тези методи е, че могат да се тестват известни или подлежащи на синтез съединения за възможността да взаимодействуват с рецептора.

Като **недостатък** може да се посочи, че негъвкавото съпоставяне пропуска много добри кандидати, тъй като записаните конформации са само част от възможните.

При добавяне на конформационно търсене методът става по-реален, но твърде бавен. Очевидно разработването на бързи алгоритми за прикачване е от особено значение

Методите, които свързват фрагменти и атоми, позволяват конструирането на нови структури чрез свързване на атоми и/или фрагменти. Могат да бъдат разделени на четири категории, в зависимост от използувания подход:

Методи, основани на свързване на фрагменти в активните центрове.

Използуват се методите за позициониране на фрагменти, като за разлика от тях фрагментите се съединяват и се конструира цяла структура. При известна 3D структура на рецептора програмата *LUDI* [12] изчислява точките на взаимодействие, които би трябвало да съответствуват на идеалната търсена структура. С изчислените области се сравняват фрагменти от неголяма база данни и на всеки от тях се приписва тежест според зададени правила или енергетични функции. *CLIX* [12] използува резултатите за позициите на центровете от *GRID* [12] и след това търси в Cambridge Structural Database [20] за молекули, удовлетворяващи тези центрове.

Предимство на тези методи е тяхното бързодействие. Анализират се само центрове, които са близо един до друг и така се ограничава областта на 3D търсенето. Молекулите могат да се конструират последователно (по фрагменти) или да се използуват за намиране възможните свързващи групи между два отделни фрагмента. Броят на центровете се избира от потребителя, което позволява да се използува по-слаб или по-силен критерий за свързване на структурата с рецептора.

Недостатък на метода е, че конструирането на структури зависи от подходящо разположение на атомите в активните центрове. Ако центровете трябва да съвпадат перфектно, повечето фрагменти ще ги пропуснат и ще бъдат отхвърлени. Друг недостатък е липсата на гъвкавост на индивидуалните фрагменти. Както при всички типове 3D търсене, полезните резултати понякога се пропускат ако фрагментите се разглеждат като твърди структури. Възможно решение е да се позволяват гъвкави фрагменти или множество конформации за фрагмент.

Методи за свързване на предварително позиционирани фрагменти.

Работят с предварително позиционирани фрагменти и намират "свързващия скелет" за съединяване на фрагментите в цяла структура, без да променят местоположението им. Примери за тези методи са програмите CAVEAT [91], HOOK, SPLICE, NEWLEAD, PRO_LIGAND, ELANA [12] и други. Използуват се различни 3D бази данни съдържащи известни съединения. В някои случаи се генерират БД от нови структури, въвеждат се на ръка или се генерират в процеса на свързване.

Като **предимства** на метода може да се посочи, че ако се разполага с множество от кандидат-фрагменти, разположени в активните центрове, то методите за свързване позволяват те бързо да бъдат съединени в цяла структура. По този начин могат да бъдат тествани хипотези за фармакофори. Друго предимство е, че е налице разнообразие от свързващи скелети и много от методите отчитат гъвкавостта.

Недостатъци на тези методи е, че 1)търсенето е бавно, ако се отчита гъвкавостта; 2)скелетът може да бъде отхвърлен ако се припокрива с рецептора; 3) генерираните молекули са твърде сложни.

Последователно построяване

Структурите се конструират чрез последователно свързване на атоми и/или фрагменти. Поредният фрагмент може да бъде поставен на произволно позволено място в структурата. Примери за методи изграждащи структурите атом по атом са програмите LEGEND [71], GenStar, GROWMOL [5] и други. В програмите GROW и SPROUT [27] въз основа на (прикачен към рецептора) скелет се добавят функционални групи, за да се конструира структура специфична за активния център. Gemini, GroupBuild, LeapFrog [12] изграждат структурите фрагмент по фрагмент. Цитираните програми са твърде различни по използуваните подходи. Важно значение има наборът от използуваните фрагменти и/или атоми. Някои програми използуват конформационно търсене за ориентиране на фрагментите, докато други ги поставят в произволна ориентация.

Методите за последователно свързване, както и повечето от методите за de novo дизайн, трябва да правят компромис между скоростта на генерация на структури и

тяхното разнообразие. **Предимство** на метода е, че 1)структурите, построени по този метод ще бъдат по-малки и по-"ефективни", отколкото структурите, получени от свързване на предварително позиционирани фрагменти, тъй като всеки фрагмент се избира с оглед на неговия принос към взаимодействието с ензима; 2)възможно е да се извърши по-детайлен конформационен анализ и да не се пропускат потенциалните решения, тъй като всеки фрагмент се добавя последователно.

Най-същественият **недостатък** е т.нар. "пресичане на мъртви зони" – област в пространството, където няма контакт с ензима. Поради оценяването на степента на близост на структурата с ензима на всяка стъпка, възможно е тази област никога да не бъде избрана. Друг проблем е комбинаторният "взрив", или огромният брой различни начини атомите да бъдат разположени в активиния център на рецептора. Въпреки това, тъй като се работи с ограничен набор от атоми и фрагменти, съществува и ограничение в разнообразието на конструираните структури. Съществен недостатък е и въпросът за синтезиране на новите структури, тъй като генерирането им от последователно свързване на фрагменти не е гаранция за успешния им синтез. За неговото преодоляване трябва да се въвеждат подходящи правила, които да оценяват структурата в процеса на генериране.

Случайно свързване на фрагменти и атоми.

Методите за случайно свързване използуват аналогични подходи на методите за последователно съединяване на фрагменти, но включват и различни техники за създаване и разкъсване на връзки в процеса на конструиране на структурата. Използуват се и различни подходи за конформационно търсене. При системите CONCEPTS и CONCERTS [76], активният рецептор се запълва с атоми, които възприемат подходящата ориентация и нискоенергетична конфигурация при прилагането на молекулна динамика. MCDNLG [26] е Monte Carlo метод за генериране на атоми в активния център и случайната им модификация с цел създаване на нова структура. AutoDock е опит за гъвкаво съпоставяне на молекулите към рецептор, използувайки Monte Carlo Simulated Annealing алгоритъм. Процедурата позволява да се извърши търсене в конформационното пространство без предварително зададени ограничения от изследователя. Подобен

праг обикновено оказва влияние на ориентацията и конформацията на молекулата в активния обект. Първоначалната конформация обикновено е заемащата енергетичния минимум, докато свързаната може да бъде с относително висока енергия. Предимствата на този метод са очевидни. На практика обаче намирането на всички възможни конформации на структури с ротация от около повече от 10 торзионни ъгъла изисква твърде дълго време за изчисление. Друг недостатък е че моделирането предполага че рецепторът има твърда структура.

Голяма част от методите започват работа с набор от атоми или фрагменти, които могат да бъдат несвързани или да съществуват връзки между част или всички от тях. Независимо от първоначалния избор на елементите за конструиране, всички методи от тази група генерират структури чрез създаване и разкъсване на връзки между тези елементи, което може да се контролира от молекулна динамика или Monte Carlo методи. Ако първоначалните елементи са само атоми, то е позволено те да се създават или унищожават, да се сменя типа и връзките на атома. Във всички случаи обаче е разрешена модификация на връзките между елементите, но при различни условия за различните методи. Приемането или отхвърлянето на дадена модификация зависи от нейната оценка или може да стане по случаен начин.

Важен клас методи за de novo дизайн са използуващите генетични алгоритми за конструиране на структури. Набор от предварително избрани структури (случайно или не) се разполага в активния център. Всяка молекула се описва с низ от символи, по аналогия с ген. По-нататък гените на двойка структури се кръстосват и се получава ново поколение структури. Новите структури се оценяват според тяхното съответствие към рецептора и най-добрите стават основа за следваща генерация. Освен това индивидуалните структури могат да бъдат подложени на случайна "мутация". Няколко групи са публикували методи за de novo дизайн с използуване на генетични алгоритми [21,28,75].

Привлекателността на тези методи се дължи на случайността, която използуват, за разлика от традиционните детерминирани методи, които конструират ограничен набор от структури. Случайното свързване и разкъсване би могло по принцип да доведе до получаване на произволна структура, следователно способността на тези

методи да "покрият" пространството на възможните структури и да генерират по-разнообразни структури е по-голяма. От друга страна, последователните методи също имат възможност да генерират твърде разнообразни съединения чрез систематично търсене. Генетичните алгоритми заемат междинно място, като позволяват недетерминираност в определени граници.

Разработени са и хибридни методи, които представляват комбинации от изброените подходи.

Съществува разнообразна литература, както и комерсиален и безплатен софтуер за молекулно моделиране.

1.6. Изводи

Въпреки наличието на голям брой методи за разпознаване на образи и клъстеризация, прилагането им в молекулното моделиране среща редица проблеми, като:

- 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, т.е. при реалистичното допускане, че едно химично съединение (една и съща топологична структура) може да има представители с различна пространствена структура.

Едно възможно решение въз основа на вероятностни оценики е предложено и реализирано в системата COREPA (Common Reactivity PAttern), описана в глава 2.

- Разделяне (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 използуваният 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 [102]. Например, според изследвания, цитирани в [107], 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.

Предложеният в дисертационната работа метод COREPA оценява вероятностното разпределение на данните и използува класификатор на Бейс за определяне критериите за принадлежност към даден клъстер.

- Предсказване на свойства.

За предсказването на свойства на химичните съединения е необходимо да се построи модел на изследваното свойство.

За тази цел е разработен формален език (глава 2.2) за описание на специфични изисквания за химичните съединения, които включват наличие или отсъствие на фрагменти, стерични и електронни условия и други. По този начин получените от COREPA критерии (и/или от прилагането на други методи) се систематизират в дърво на решенията, което да се използува за предсказване на свойствата на други химични съединения.

- Конформационен анализ.

Отчитането на гъвкавостта на молекулите изисква намирането на представително множество от конформерите на всяка от разглежданите молекули. Намирането на всички възможни конформери е комбинаторен проблем. От друга страна,

стохастичните методи създават проблеми при необходимостта за генериране на конформери при зададени ограничения.

В дисертационната работа е предложен и реализиран генетичен алгоритъм за намиране на множество конформери най-добре покриващо конформационното пространство (глава 3.1). Алгоритъмът дава възможност за генериране на конформери при наложени ограничения. Тези ограничения се описват чрез създадения формален език описан в точка 2.2.

- Генериране на нови съединения.

Молекулният дизайн има за цел намирането на нови или съществуващи съединения с предварително зададени свойства. За идентифицирането на съединения със зададени свойства в съществуващи бази данни се използува получения модел на свойството.

За получаване на нови съединения със зададени свойства е разработен генетичен алгоритъм на две нива, описан в глава 3.2. Той дава възможност и за отчитане гъвкавостта на съединенията в хода на генерацията, което го отличава от известните алгоритми.

2. МЕТОДИ ЗА МОДЕЛИРАНЕ И АНАЛИЗ НА ДАННИ

Извличането на знания е нетривиален процес за идентифициране на валидни, нови, потенциално полезни и разбираеми образи (шаблони) от зададено множество от данни. Данните съдържат описание на изследваните обекти. В повечето случаи обектът се определя с наредена *n*-торка от свойства, като могат да се използуват и релации между обектите или между части от тях. Свойствата могат да бъдат дискретни (подредени или неподредени) или непрекъснати, т.е. подмножество от реални числа. Свойството на даден обект се задава чрез атрибут и неговата стойност (Цвят – атрибут, Цветът е син – свойство; Токсичност – атрибут, Лекарството е токсично – свойство). Всеки конкретен обект може да се използува за обучение (построяване на модела) или върху него може да се приложи вече построеният модел (предсказване). Построяването на модел има за цел да обобщи или резюмира набора от данни, с цел описание и предсказване. Моделът може да бъде класификатор, регресор, визуализация на данните за интерпретация от човека, обработка на данните за последващ процес на извличане на знания, откриване на общите характеристики на група обекти. Откриването на шаблони изисква дефинирането на понятието "шаблон" или "общ образ". Шаблоните могат да бъдат представени чрез релации. В задачите за класификация се търсят релациите между отделните обекти от дадена област и съответните класове (маркери за класовете). В задачите за регресия се търси съответствие между обекти и стойности в зададено метрично пространство (например непрекъснато). или стойности на някаква функция. Целта е от зададеното множество данни $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_k, y_k)\}$, генерирани от неизвестна функция $f = \mathbf{x} \rightarrow y$, да се намери функция $f' = \mathbf{x} \rightarrow y$, такава че f' апроксимира f. Всеки обект от D се дефинира като наредена n-торка $\mathbf{x} = x_1 x_2, \ldots x_n$, а x_j съответствува на конкретно свойство.

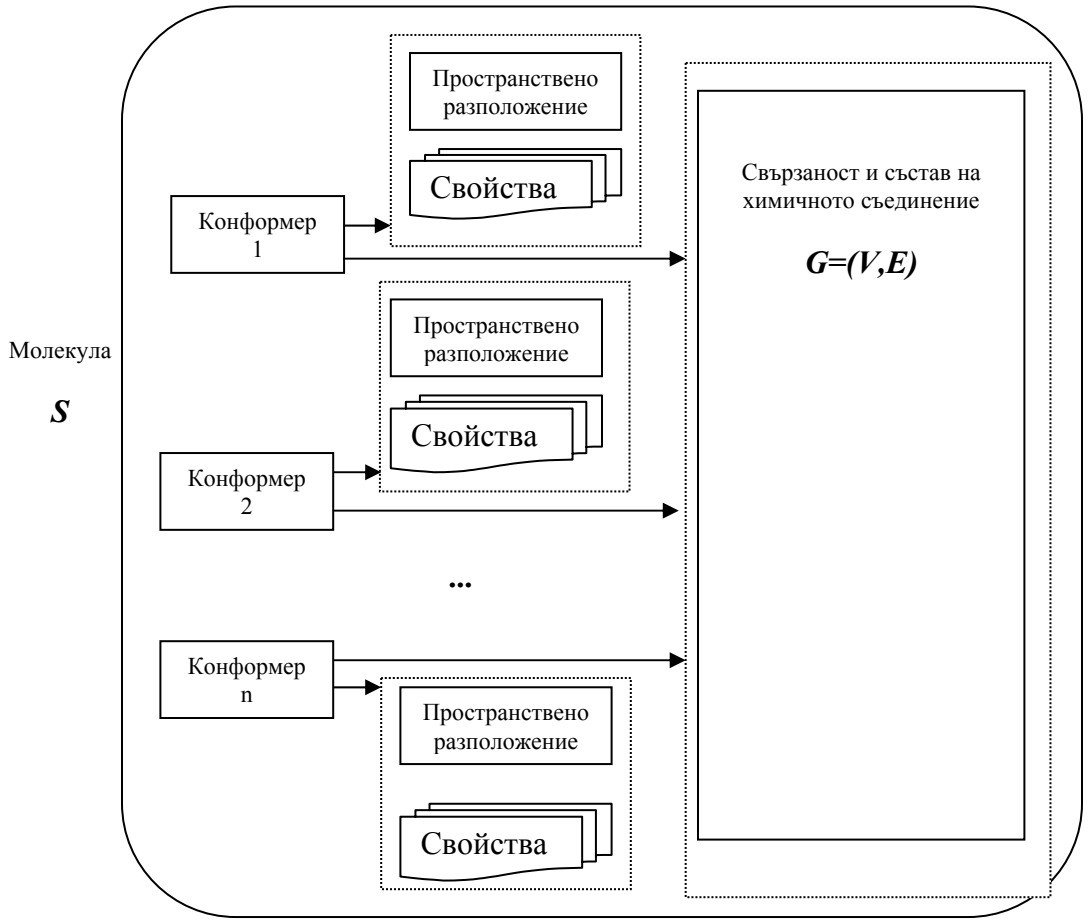
Обектът на изследване представлява гъвкава молекула, което означава, че тя има повече от едно пространствено разположение. Пространственото разположение оказва влияние върху част от свойствата на молекулата (например енергия на

конформера, разстояния между атомите). Друга част от свойствата на молекулата не зависят от пространственото разположение и са общи за цялата молекула (например свързаност и състав на молекулата, молекулна маса). Свойствата, специфични за конформерите, както и общите свойства за молекулата също могат да бъдат получени експериментално или изчислени.

Съвкупността от конформери и свойства, показана на Фигура 2.1, представлява цялостен сложен обект, представящ химичното съединение. За целите на молекулното моделиране е необходима възможност за сравнение между такива обекти. Както се вижда, молекулата е сложен обект и не е достатъчна класическата дефиниция на обекта като наредена n-торка.

Структурата на химично съединение може да бъде зададена като неориентиран граф $G=(V,E)$, където с V означаваме върховете (атомите), а с E – ребрата на графа (връзките между атомите). За всеки връх от множеството на върховете $V(G)$ разполагаме с набор от пресметнати или измерени стойности, които ще наречем локални дескриптори и ще означим тяхното множество с $L=\{(v,u)|v\in V(G),\ u\in R^m\}$. Разполагаме също и с набор от пресметнати или измерени стойност за цялата молекула, които ще обозначим с $z\in R^n$. Ще дефинираме обекта C, с който ще описваме множеството от конформери на различни химични съединения. $C=\{(g,d,z) \mid g\in\{G\},\ \mathbf{d}\subset L,\ \mathbf{z}\in R^n\}$. Химичното съединение, като съвкупност от конформери, представлява сложен *вариантен обект*, s_i, където свързаността на всички c_i представители на обекта $s_{i,1}, s_{i,2}\ldots; s_{i,j}\in C$ се описва с един и същ граф G_i.

$$
\begin{aligned}
&\textit{молекула}\quad s_i \in M\\
&c_i\ \textit{конформери}:\ s_{i,1}, s_{i,2}, \ldots s_{i,c_i}\\
&\textit{представяне}:\ C(s_{i,j}) = (G_i, \mathbf{d}_{ij}, \mathbf{z}_{ij});\\
&G_i = (V_i, E_i); \mathbf{d}_{ij} \subset L; \mathbf{z}_{ij} \in R^n
\end{aligned}
$$



Фигура 2.1

В базата данни от химични съединения гъвкавата молекула се задава с краен брой конформери. Данните за пространственото разположение на конформерите могат да бъдат получени експериментално или генерирани с подходящ алгоритъм. Броят на различните пространствени положения на дадена молекула е безкраен, като обикновено се търсят конформери, съответствуващи на локалните минимуми на енергията на образуване на химичното съединение. Техният брой обикновено е огромен и зависи от големината на молекулата, степените на свобода и др.

Проблемът за представянето на гъвкавата молекула с краен брой конформери, изисква метод за намирането и избора на тези конформери, които най-добре отразяват пространственото разнообразие на структурата.

В настоящата работа (т.3.1) е предложен генетичен алгоритъм за намиране на зададен брой максимално различни конформери.

Предложена е съвкупност от методи, решаващи следните взаимосвързани задачи:

- Анализ и обобщение на масиви данни, съдържащи информация за химични съединения, чрез откриване на общи шаблони (метод *COREPA - Common Reactivity Pattern*– т.2.1). 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 правила (*RDL* – т.2.2), което осигурява универсален метод за записване на условия, зависещи от структурата на химичните съединения. Построеният модел може да се използува многократно за автоматично предсказване на моделираното свойство, търсене на молекули с зададени свойства(т. 4.1.3) или задаване на ограничения при компютърно конструиране на нови химични структури (т.3.2).
- Генериране на множество от най-различаващите се конформери (молекули с една и съща свързаност и състав, но различно разположение в пространството) чрез генетичен алгоритъм - *Genetic Algorithm Search* - т. 3.1;
- Генериране на нови химични съединения със предварително зададени свойства, чрез генетичен алгоритъм на две нива(*LEADGEN – Lead Generation* - т. 3.2). Свойствата се задават чрез техния модел, записан като правила в езика *RDL* – т.2.2. Генетичният алгоритъм на първо ниво конструира нови химични структури, удовлетворяващи зададените условия, а на второ ниво за всяка нова молекула се прилага генетичният алгоритъм от т. 3.1 за получаване на конформерите на молекулата..

2.1. Откриване на шаблони в масиви данни - COREPA

В дисертационния труд е разработен метод, наречен COREPA за откриване на шаблони в масиви данни. За тази цел се използува вероятностен подход. Множеството данни се разделя от потребителя на групи. Намират се параметрите,

които осигуряват максимално различие между вероятностните плътности на групите.

2.1.1. Алгоритъм

COREPA работи върху множество от химични съединения, зададени с тяхната структура, пространствени координати и произволен брой изчислени или експериментално получени свойства (обучаващо множество). Възможно е данните за свойствата да са непълни, т.е., не за всички съединения да има данни за всички свойства. Всяко химично съединение е представено с краен брой конформери, като броят ми за различните съединения може да бъде различен. Целта е да се построи модел на зададено свойство – обикновено това е биологична активност – например токсичност, мутагенност, възможност за свързване към рецептор и др., данните за което са получени чрез експериментално измерване.

1. Обучаващото множеството се разделя на две групи от активни и неактивни съединения, според стойностите на активността и преценката на потребителя.
2. Избира се множество параметри, които потребителят предполага, че могат да имат отношение към активността. Това могат да бъдат всички параметри, за които има стойности, или произволно тяхно подмножество. Освен това, потребителят може да дефинира параметри, за които в базата данни няма записани стойности, но те могат да бъдат намерени въз основа на информацията за всяко химично съединение (например разстояние между кислородните атоми в молекулата, или заряд на определен атом, както и по-сложни условия). Това дава възможност на химика да дефинира хипотези за влиянието на фрагменти от молекулата върху нейната активност.
3. За всички параметри се оценява плътността на вероятността за всяка от двете групи (по описания по-долу начин) и се избира параметъра, за който разстоянието между плътността на вероятността на активната и неактивна групи е най-голямо. Определя се интервалът за този параметър, при който молекулата може да се класифицира към една от групите с най-голяма вероятност и минимална грешка.
4. Ако дървото е празно (обучаващото множество е първоначално зададеното), то се създава корен на дървото на решението, като за решаващо условие се задава принадлежност към избрания интервал.
5. В противен случай, се създава нов възел на дървото на решенията,, като за решаващо условие се задава принадлежност към избрания интервал.
6. Обучаващото множество се разделя на две множества според критерия в новия възел на дървото
7. За всяко от редуцираните множества се продължава с т.3, ако в съответното множество се съдържат съединения и от двете групи (активна и неактивна). Ако множеството съдържа съединения само от една група, или не може да се направи по-нататъшно разделяне (например не е значимо според χ^2 статистика), то изпълнението на алгоритъма завършва.

.

2.1.2. Оценка на параметрите

Изборът на параметър, който води до най-добро разделяне между двете групи се прави въз основа на оценка на плътността на вероятността на всяка от групите и сравнението им.

Нека е дадено множество от вариантни обекти $M=\{S_1, S_2, ...S_n\}$, и множество от класове $\omega=\{\omega_1,..., \omega_k\}$. Нека означим с x произволен дескриптор на обекта S, такъв, че $x \in R$.

Търсим вероятността за принадлежност към клас ω_i, при известни стойности на x за обекта $S_i=(g_i,\mathbf{d}_i,\mathbf{z}_i)$. Според формулата на Бейс:

$$p(\omega_i|x) = \frac{p(\omega_i)p(x|\omega_i)}{p(x)} = \frac{p(\omega_i)p(x|\omega_i)}{\sum_{j=1}^{k} p(\omega_j)p(x|\omega_j)} \qquad (2.1)$$

Ако предположим, че класовете ω_i са равновероятни, то неизвестни остават само вероятностите $p(x|\omega_i)$. Оценка на плътността на вероятността p означава намиране на подходяща функция $f(x)$, която да описва наличните данни.

Оценка на плътността на вероятността с кернел функции

В настоящата задача не съществуват основания за избор на конкретен вид на функцията, затова се прилага непараметричен метод – модифициран вариант на публикувания в [85] метод за оценка на плътността на вероятността с кернел функции (*kernel density estimation*). Функцията $f(x)$ се изчислява по формулата:

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} \varphi\left(\frac{x - x_i}{h}\right) \qquad (2.2)$$

където n е броят на точките, h е изглаждащ параметър, φ е кернел функция. За φ може да се избере произволна функция за плътност на вероятността. Тази оценка представлява смес от n плътности, центрирани в точките от извадката. Следователно, за голям размер на извадката са изчислението ще е много трудоемко. Оценката на плътността всъщност е конволюция на неизвестната плътност на разпределение и кернел функцията [85][3][2]. Ако $f_n(x)$ е оценката на $f(x)$ в n-тата точка, то усредената оценка по всички точки е:

$$\overline{f_n(x)} = \sum f_n(x) = \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{1}{nh}\sum_{i=1}^{n}\varphi\left(\frac{x - x_i}{h}\right)\right] =$$

$$= \int \frac{1}{nh}\varphi\left(\frac{x - v}{h}\right)p(v)dv = \int \delta_n(x - v)p(v)dv$$

Ефективният алгоритъм за изчисляването на конволюция се основава на факта, че Фурие преобразуването на конволюцията е произведение от Фурие преобразуванията на участвуващите в конволюцията функции. От това следва, че е достатъчно да се намерят Фурие преобразуванията на кернел-функцията δ и плътността $p(v)$, да се умножат, и да се намери обратното Фурие преобразуване на произведението.

Дискретното преобразуване на Фурие може да се приложи, ако отчетите на точките са през равни интервали. За да може да се използува бързото преобразуване на Фурие, то данните трябва да се приведат в подходящ вид чрез предварително разделяне на интервали. Нека интервалът $[a,b]$ обхваща всички точки. За да се ограничи граничният ефект при препбразуването на Фурие, интервалът се разширява. Например, при използуване на Гаусова кернел функция с параметър h, то задаваме

$a < min(X_i) - 3h$ и $b > max(X_i) + 3h$.

Избираме $M = 2^r$, където r е цяло число. $r = 7$ или $r = 8$ дава добри резултати, но r може и да бъде избрано според необходимата точност на получаване на дискретното представяне на плътността на вероятността. Оценката на плътността ще бъде получена за M точки в интервала $[a,b]$. Нека дефинираме

$$\delta = (b - a)/M$$
$$t_k = a + k\delta, \quad k = 0,1,...M - 1$$

Дискретизираме дните по следния начин: ако точката X попада в интервала $[t_k, t_{k+1}]$, то тя се разделя на тегла $n^{-1}\delta^{-2}(t_{k+1} - X)$ в t_k и $n^{-1}\delta^{-2}(X - t_k)$ в t_{k+1}. Тези тегла се акумулират по всички данни X_i, като се получава последователност от тегла (ξ_k), чиято сума е δ^{-1}. За $-½\ M \le l \le ½\ M$, дефинираме Y_l като дискретното Фурие преобразуване:

$$Y_l = \frac{1}{M}\sum_{k=0}^{M-1} \xi_k e^{i\frac{2\pi kl}{M}}$$

Модифициран метод за оценка на плътността на вероятността с кернел функции за вариантни обекти

Оценка на вероятностната плътност чрез кернел функции за всеки клас ω_i може да се получи съгласно формула (2.1)

$$p(x|\omega_i) = \frac{1}{n_i h} \sum_{k=1}^{n_i} \varphi\left(\frac{x - x_{ik}}{h}\right),$$

където n_i е броя на точките в класа ω_i. Този запис обаче *не отчита наличието на вариантни обекти* както бяха дефинирани по-горе, тъй като предполага че всички точки x_{ik} са равноправни. Ще изведем формулата за оценка на плътността, като използуваме понятия от конформационния анализ за вероятностното разпределение на конформерите на една молекула.

Нека означим с C_{sj} събитието, представляващо появата на j-я конформер на молекулата s. Оценката на вероятността е

$$p(x|C_{sj}) = \frac{1}{N_{sj} h} \sum_{k=1}^{N_{sj}} \varphi\left(\frac{x - x_{sjk}}{h}\right),$$

където N_{sj} е броя на стойностите на параметъра x. Нека вероятността за наличието на конформер j в множеството от R_i конформери за молекулата S_i се изразява с $p(C_{ij})$. Тогава

$$p(x|S_i) = \sum_{j=1}^{R_i} p(C_{ij}) p(x|C_{ij}) =$$

$$= \sum_{j=1}^{R_i} \left[p(C_{ij}) \frac{1}{N_{ij}} \sum_{k=1}^{N_{ij}} \varphi(x - x_{ijk}) \right] = \sum_{j=1}^{R_i} \sum_{k=1}^{N_{ij}} \frac{p(C_{ij})}{N_{ij} h} \varphi\left(\frac{x - x_{ijk}}{h}\right) \qquad (2.3)$$

Нека към класа ω_m принадлежат M_m молекули. Ще считаме вероятността $p(S_i)$ за принадлежност на молекулата S_i към класа ω_m за еднаква за всички молекули и равна на $1/M_m$. Следователно:

$$
\begin{aligned}
p(x|\varpi_m) &= \sum_{i=1}^{M_m} p(S_i)p(x|S_i) \\
&= \sum_{i=1}^{M_m}\left[p(S_i) \sum_{j=1}^{R_i} \sum_{k=1}^{N_{ij}} \frac{p(C_{ij})}{N_{ij}h} \varphi\left(\frac{x - x_{ijk}}{h} \right)\right] = \\
&= \frac{1}{M_m} \sum_{i=1}^{M_m}\left[\sum_{j=1}^{R_i} \sum_{k=1}^{N_{ij}} \frac{p(C_{ij})}{N_{ij}h} \varphi\left(\frac{x - x_{ijk}}{h} \right)\right] \\
&= \frac{1}{M_m} \sum_{i=1}^{M_m} \sum_{j=1}^{R_i} \sum_{k=1}^{N_{ij}} \frac{p(C_{ij})}{N_{ij}h} \varphi\left(\frac{x - x_{ijk}}{h} \right)
\end{aligned}
\qquad (2.4)
$$

За определяне на $p(C_{ij})$ ще използуваме методите на конформационния анализ. Известно е, че термодинамическите функции на произволна система могат да бъдат описани със статистическата сума Z [101]

$$Z = \sum_i g_i e^{(-E_i / k_B T)}, \qquad (2.5)$$

където E_i са енергетичните нива на енергията на системата, k_B е константата на Болцман, T е абсолютната температура и g_i е кратността на изродените нива с енергия E_i. Важно свойство на статистическата сума е нейната мултипликативност, т.е. ако системата се състои от две части А и В с нива на енергията E_i и E_j, то $Z^{A+B} = Z^A Z^B$. Тъй като физическият смисъл на статистическата сума представлява вероятността да се реализира даденото състояние, то от вероятностна гледна точка мултипликативността означава, че съставните части на системата могат да се разглеждат като независими събития.

Нека разгледаме съвкупността от конформери на дадена молекула S_i. Тогава статистическата сума за 1 мол от конформер j може да се запише като $Z_j = n_j e^{(-E_j / k_B T)}$, където n_j е броя на конформерите с енергия E_j. В равновесно състояние вероятността за всички конформери е еднаква, т.е $Z_1 = ... = Z_j = ... = Z_n$. Пропорцията на конформер j в съвкупността от конформери на дадена молекула се изразява с

$$\frac{n_j}{\sum_{m=1}^{N} n_m} = \frac{Z_j e^{(-E_j / k_B T)}}{\sum_{m=1}^{N} Z_m e^{(-E_m / k_B T)}} = \frac{1}{\sum_{m=1}^{N} \frac{e^{(-E_m / k_B T)}}{e^{(-E_j / k_B T)}}} =$$
$$= \frac{1}{\sum_{m=1}^{N} e^{(E_m - E_j)/RT}} = \frac{1}{\sum_{m=1}^{N} e^{(E_j - E_{\min})/k_B T - (E_m - E_{\min})/k_B T}} =$$
$$= \frac{1}{\sum_{m=1}^{N} \frac{e^{-\Delta E_m / k_B T}}{e^{-\Delta E_i / k_B T}}} = \frac{e^{-\Delta E_j / k_B T}}{\sum_{m=1}^{N} e^{-\Delta E_m / k_B T}} \qquad (2.6)$$

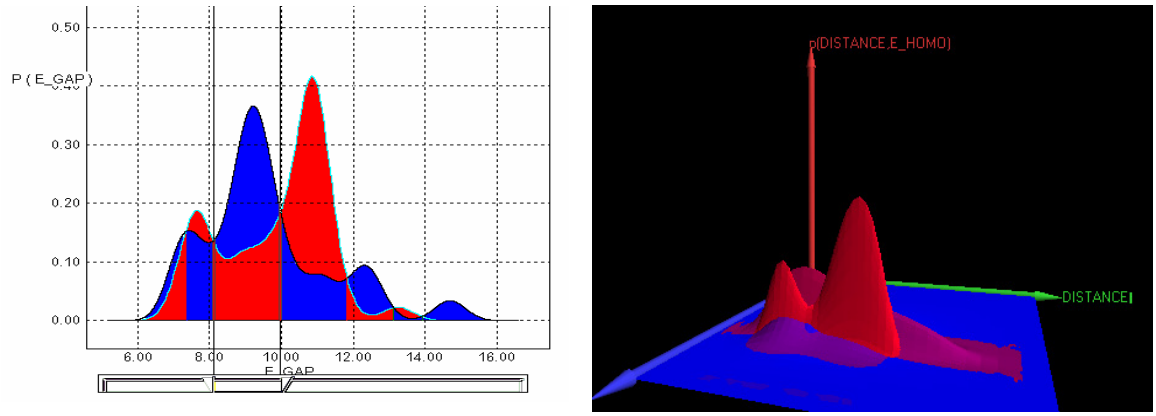
където E_{min} е енергията на конформера с минимална енергия. Формула (2.6) показва, че пропорцията на даден конформер зависи от разликата в енергиите на конформерите. Следователно $p(C_{ij})$ може да се изчисли от наличните данни за енергията на конформерите на всяка молекула и ще я считаме за известна величина. Нека обозначим коефициента пред кернел функцията в (2.4) с α_{ij} (за *j-я* конформер на *i-а* молекула*)*

$$\alpha_{ij} = \frac{p(C_{ij})}{N_{ij}} = \frac{e^{-\Delta E_{ij} / k_B T}}{N_{ij} \sum_{m=1}^{R_i} e^{-\Delta E_{im} / k_B T}} \qquad (2.7)$$

От (2.7) и (2.4) и получаваме

$$p(x|\varpi_m) = \frac{1}{M_m} \sum_{i=1}^{M_m} \sum_{j=1}^{R_i} \sum_{k=1}^{N_{ij}} \frac{\alpha_{ij}}{h} \varphi\left(\frac{x - x_{ijk}}{h}\right) \qquad (2.8)$$

Формулата (2.8) представлява сума по всички точки, описващи обектите от даден клас. При подходящо прилагане на метода с предварително разделяне на интервали за изчисляване може да се използува бързо преобразуване на Фурие[85].



(а) (б)

Фигура 2.2

Фигура 2.2а е илюстрация на оценката на вероятностната плътност за набор от химични съединения по параметър E_GAP. Червената графика съответствува на групата на активните съединения, а синята графика - на групата на неактивните съединения. Описаният метод може да се обобщи и за оценка на многомерна плътност на вероятност. На Фигура 2.2б е показана оценката на вероятностна плътност за пет групи съединения по параметри "разстояние между специфични атоми" и E_HOMO.

След като вероятностните плътности за всеки клас $p(\omega_i|x)$ са изчислени, то могат да бъдат решени следните задачи:

- Оценка на подобието между класовете ω_I;
- Оценка на подобието между даден обект и клас;
- Оценка на подобието между два обекта;
- Класифициране на обекти;
- Извличане на признаци за класифициране на обекти.

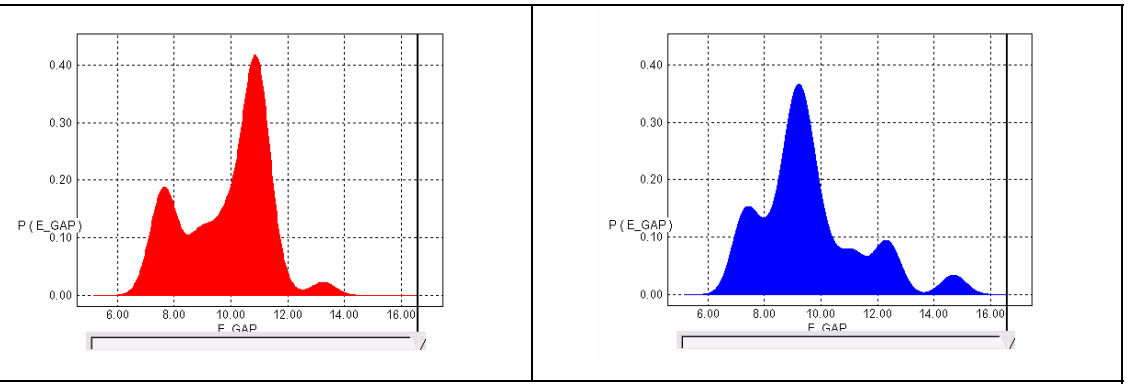
За тези цели се оценява разстоянието между съответните вероятностните плътности, като могат да се използува някое от известните вероятностни разстояния (*Kullback-Leibler divergence* [31], разстояния [85] на *Чернов, Bhattacharyya* [9], *Matusita, Mahalanobis, Patrick-Fisher, Hellinger* [55]). Квадратът на разстоянието по Hellinger между вероятностните плътности p_1 и p_2 е

$$D^2(p_1, p_2) = \int (\sqrt{p_1(x)} + \sqrt{p_2(x)})^2 dx \,, \qquad (2.9)$$

като при еднакви p_1 и p_2 има стойност 0 и максимална стойност 2 при несъвпадащи функции.

Оценка на подобието между класовете

Степента на подобие между класовете ω_i дефинираме като разстоянието по Hellinger (формула (2.9)) между вероятностните плътности за всеки два класа ω_i и ω_j : $D_{ij}(x) = D(p(\omega_i|x), p(\omega_j|x))$. Това разстояние е важна оценка на качеството на параметъра x като критерий за класификация. Ниска стойност на D_{ij} означава че параметърът x е лош критерий за разделяне на класовете ω_i и ω_j и обратно. Тази метрика може да се използува и за оценка на степента, в която отделните съединения влияят върху общия образ. Стабилността на образа се оценява чрез последователна оценка на плътността на вероятността за всички редуцирани множества от обекти, където редуцираните множества съдържат всички обекти, с изключение на един ("*leave one out*" процедура). Фигура 2.3 показва оценената вероятностна плътност съответно за групата активни (червена графика) и неактивни (синя графика) съединения.

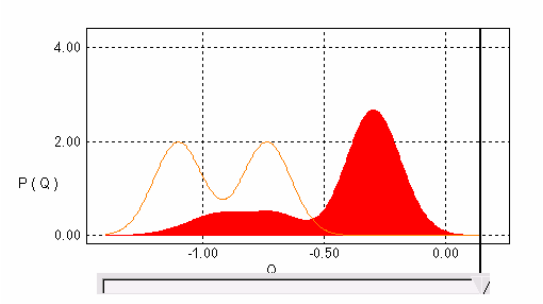


Фигура 2.3

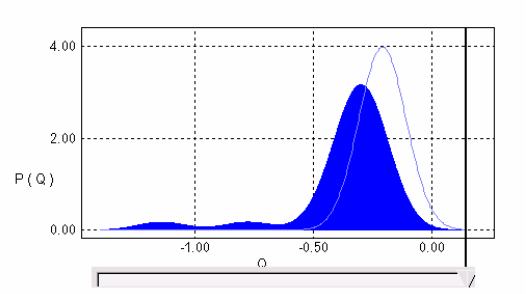
Оценка на подобието между даден обект и клас

Степента на подобие между даден обект и клас ω_i дефинираме като $D_i(O) = D(p(\omega_i|x), p(O|x))$. Това разстояние може да се интерпретира като степента на

принадлежност на обекта към даден клас. Обектът се класифицира към клас ω_I, ако стойността на $D_i(O)$ е най-голяма в сравнение със $D_j(O)$, $i \neq j$. На Фигура 2.4 е показана оценената вероятностна плътност за класа на активните съединения (червена плътна графика) и с тънка линиа плътността, съответствуваща на един обект (едно активно съединение). Аналогичен пример е показан на Фигура 2.5 , където синята плътна графика показва вероятностната плътност на класа неактивни съединения.



Фигура 2.4



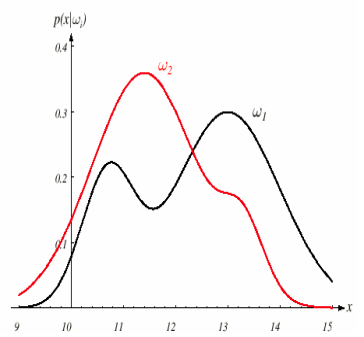
Фигура 2.5

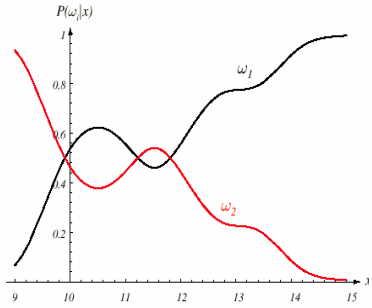
Оценка на подобието между два обекта;

Степента на подобие между два обекта O_i и O_j дефинираме $D(O_i,O_j) = D(p(O_i|x), p(O_j|x))$. По този начин сме дефинирали мярка за сходството между двата (вариантни) обекта, което е от голямо значение, ако подчертаем, че обектите бяха въведени за да описват гъвкави химични съединения. Понятието сходство между химични съединения (дори и при представяне само с един конформер, т.е. негъвкави) не е строго дефинирано в литературата и често се оценява със съпоставяне на пространствената структура на съединенията. При наличие на повече от един представител на съединенията, задачата става още по-сложна. Предложената количествена оценка на сходството между гъвкави химични съединения дава възможност да се избегне субективизмът при определянето на степента на сходство и необходимостта от трудоемки и нееднозначни алгоритми за съпоставяне на тримерните структури.

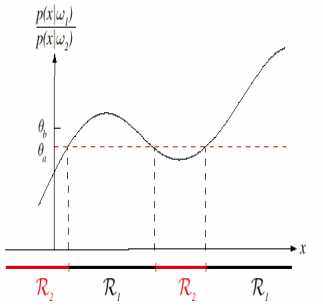
Класифициране на обекти

Класифициране на нови обекти може да се извърши като се намери класът с най-голяма стойност $p(\omega_i|x)$ за дадена стойност на параметъра x. В общия случай, обектът може да има множество стойности за параметъра x и класификацията може да се извърши, като се намери класът, за който стойността на $D_i(O)$ е максимална.





Фигура 2.6



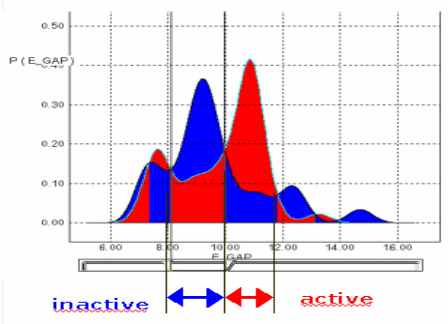
Фигура 2.7

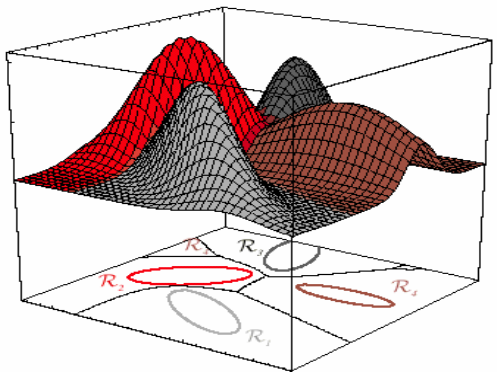
Извличане на признаци за класифициране на обекти.

При наличие на два класа ω_1 и ω_2, отношението на правдоподобие L (Фигура 2.7) определя принадлежността към клас при дадена стойност x. Освен това, L може да се разглежда и като решаваща функция, която разделя пространството, в което е дефинирано x, на области съответно където $L>1$ и $L<1$.

$$L = \frac{p(\omega_1|x)}{p(\omega_2|x)}, \text{ където}$$

обектът се класифицира към клас ω_1, ако $L>1$,
обектът се класифицира към клас ω_2, ако $L<1$,
класификацията е произволна, ако $L=1$ (2.10)



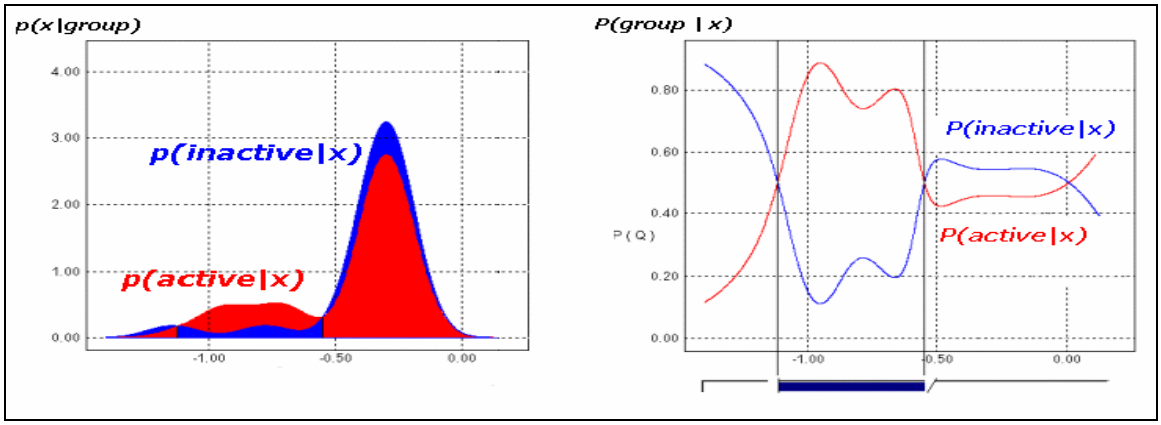


Едномерен случай Двумерен случай

Фигура 2.8

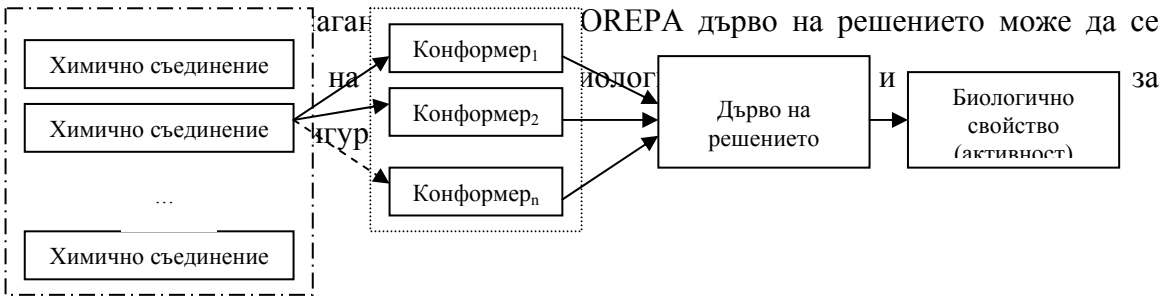
В едномерния случай областите се свеждат до интервали, т.е. могат да се определят интервали за параметъра **x**, които да служат като критерий за класификация. В многомерното пространство областите могат да бъдат по-сложни, но с известна точност могат да бъдат сведени до съвкупност интервали за параметрите. Това е важен резултат за разглежданата задача за откриване на общ реакционен образ на химичните съединения, тъй като получените интервали могат да бъдат тълкувани от гледна точка на химични, електронни и други критерии и да се достигне до хипотеза за механизма или обяснение на изследвания ефект (биологично свойство). Отношението на правдоподобие $L(\omega_l)$ може да се използува по аналогичен начин за намиране на областите на x, където $L(\omega_l)$ има максимална стойност. Тъй като $L(\omega_l)$ приема стойности от 0 до 1, могат да се определят областите на x, където отношението на правдоподобие е над зададен праг.

$$L(\omega_i) = \frac{p(\omega_i|x)}{\sum_{j=1}^{k} p(\omega_j \mid x|)} \qquad (2.11)$$



Фигура 2.9

Фигура 2.10



аган OREPA дърво на решението може да се на иолог и за игур

2.2. Построяване на модели, описващи данните

Методът COREPA позволява извличане на общите характеристики на обекти, принадлежащи към съответен клас. При оценяване на многомерна вероятностна плътност чрез COREPA, общите характеристики представляват области на параметри, при които обектът е от търсения клас със вероятност по-голяма или равна на зададена от потребителя. В едномерния случай тези характеристики са във вид на интервали на параметрите. Получените области могат да се използуват за по-нататъшно класифициране на непознати обекти. Обикновено обектите са описани с голям брой параметри, и нерядко има липсващи стойности или параметри за част от обектите. По тази причина понастоящем COREPA се използува главно за оценка на едномерна вероятностна плътност и извличане на общите характеристики във вид на интервали. За да се отчете разнообразието на параметрите, въз основа на съвкупността от получени общи характеристики се построява дърво на решението. Дървото на решението се използува за класифициране на непознати обекти, търсене в бази данни от химични съединения, генериране на нови съединения със зададени свойства. За тази цел е необходимо описание на дървото на решението, което да позволява:

- автоматизирано вземане на решение - прилагането на вече създадено дърво на решението към друго множество обекти чрез компютърна програма;
- универсалност на описанието - да не се изисква създаването на нова компютърна програма за всеки нов проблем;
- разбираемост за потребителя - описанието да е в термини близки до тези, с които работи експертът химик.

Задачата се усложнява от факта, че получените общи характеристики представляват параметри, зададени по сложен начин, с отчитане свързаността и тримерното разположение на атомите в молекулата. Така например, като общи характеристики могат да бъдат зададени:

- Молекулни характеристики (единична стойност за цялата молекула) от вида на изброените в т.1.1.4;
- Характеристики на атоми, т.е. по една стойност за атом и съответно множество стойности за цялата молекула;
- Характеристики на специфично зададени атоми (например зарядите на всички кислородни атоми в молекулата). Атомите могат да бъдат зададени и по сложен начин, например атом, принадлежащ на определен фрагмент;
- Характеристики на специфични зададени атоми, при зададени ограничения на други параметри на молекулата (стойности, параметри атоми, фрагменти и други);
- Взаимно разположение на фрагменти, където за параметър се приема разстояние между центровете на фрагментите;
- Неколичествени характеристики, като общ подфрагмент за набор от молекули. Условието за вземане на решение в този случай е наличие или отсъствие на подфрагмента.

За решаване на поставените цели е разработен формален език за описание на логически правила за класифициране и търсене в бази данни от химически съединения [45, 43]. Правилата се основават на дефинирани от потребителя условия, които включват съпоставяне на структурни подфрагменти с ограничения върху стойностите на молекулните параметри и взаимното разположение на химичните групи. Ограниченията се записват, използвайки разширен SMILES[97] код. Правилата представляват изрази от вида на описаните ограничения, свързани с булеви оператори. Има възможност за присвояване на стойности на параметри на молекулата. По този начин може да бъде построено произволно дърво на решението.

SMILES е стандарт за обозначаване на химични съединения със символен низ. SMILES е абревиатура на *Simplified Molecular Input Line Entry Specification*. Опростено описание на SMILES може да се даде със следните правила:

1)Атомите се представят с техните атомни символи

2)Двойните връзки се означават с ‘=’, тройните – с ‘#’.

3)Разклоненията се отбелязват със скоби

4)Циклите се обозначават с двойки от съвпадащи цикли

Подробно описание на SMILES е дадено в приложение I.

2.2.1. *Синтаксис на езика Rule Description Language - RDL*

Формално описание на езика RDL с форми на Bacus-Naur

Фигура 2.11 съдържа формалното описание на езика RDL в форми на Bacus-Naur. Категориите са заградени в < >. Ключовите думи са показани с удебелен шрифт.

Структура на RDL програмата

RDL съчетава означението SMILES с някои разширения и структури по подобие на програмните езици. RDL програмата се състои от три части - *define*, *rule* и *apply*. Първите две секции съдържат дефиниции на разглежданите молекулни характеристики, докато последната може да се разглежда като тяло на програмата от химически правила. RDL програмата се изпълнява върху дадена молекула, като в резултат на разглежданата молекула се присвояват стойности на параметрите, както е зададено в RDL програмата. В този смисъл, контекстът, в който се изпълнява RDL програмата винаги е една (текущата) молекула. Присвояваните стойности обикновено имат смисъл на класифициране на молекулата към даден предварително дефиниран клас.

<ruleprogram> ::= **defines**
<definessection>
rules
<rulessection>
apply
<applysection>
end.
<definessection> ::= <defineline>*
<defineline> ::= <property> **:** <smiles>
<rulessection> ::= <ruleline>*
<ruleline> ::= <rule> **:** <boolexpr>
<boolexpr> ::= **"**<rule>**"** | **"**<property>**"** |**(**<boolexpr>**)** | **not** <boolexpr> |
<boolexpr> **and** <boolexpr> | <boolexpr> **or** <boolexpr>
<applysection> ::= <listofstatements>
<listofstatements> ::= <statementsemicolon>*

<statementsemicolon> ::= <statement>;
<statement> ::= <assignmentstatement> | <ifstatement> |
<compoundstatement>
<assignmentstatement> ::= <selector> := <realvalue>
<ifstatement> ::= <ifthen> | <ifthenelse>
<ifthen> ::= **if** <rule> **then** <statement>
<ifthenelse> ::= **if** <rule> **then** <statement> **else** <statement>
<compoundstatement> ::= **begin** <listofstatements> **end**
<property> ::= <id>
<rule> ::= <id>
<id> ::= <letter>[<alphanum_>*]
<letter> ::= A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | a | b | c | d | e | f | g | h | I | j | k | l | m | n | o | p | q | r | s | t | u | v | w | x | y | z
<alphanum> ::= <letter> | <num> | _
<num> ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
<smiles> is a SMILES string.
<realvalue> is a (possibly signed) real value.

Фигура 2.11

RDL програмата се записва в текстов файл с разширение **.rul* . Текстовия файл е организиран по следния начин:

defines
text of *defines* section,
rules
text of *rules* section
apply
text of *apply* section
end.

Ключовите думи **defines**, **rules, apply**, и **end.** трябва да бъдат на отделни редове.

Секция *Define*

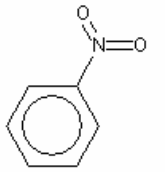
defines секцията се състои от **дефиниции на молекулни характеристики**, всяка от които на нов ред.

propertyid : SMILESstring

propertyid трябва да бъде валиден идентификатор, SMILESstring трябва да бъде валиден SMILES низ. Дефинициите се използуват по-нататък, за построяване на дефиниции на правила в секция *rules*.

Пример:
Hal:I,Br,Cl,F
CSP3: C{sp3}

Молекулните характеристики се записват като SMILES и функционират като заявка за търсене на подфрагменти на двумерни структури. В примера по-горе дефиницията *Hal* означава наличие на халогенен атом в молекулата - йод, бром,хлор,флуор. Дефинициите могат да включват и по-сложни фрагменти. Например, SMILES стрингът "c1ccccc1N=(O)=O" задава показаната на Фигура 2.12 структура и означава заявка за търсене на такъв подфрагмент. (подразбира се, че атомите, които нямат буквено означение на рисунката на фрагмента са въглеродни атоми - C).

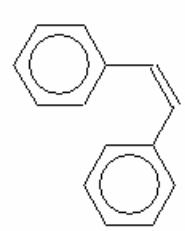
Дефиниция	SMILES	Фрагмент
Fragment: c1ccccc1N=(O)=O	c1ccccc1N=(O)=O	

Фигура 2.12

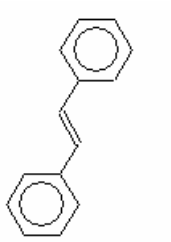
Дефинициите могат да задават допълнителни изисквания към молекулните фрагменти, които се търсят с отчитане само на свързаността на молекулата (2D ниво), или с отчитане на тримерното разположение в пространството (3D ниво), или засягат различни параметри на химичното съединение. Това изисква разширение на стандартния SMILES, чрез който могат да се задават само ограничения на 2D ниво. За тази цел са въведени *квалификатори*, които представляват низове, заградени в {} скоби и се разполагат вътре в SMILES низа, до съответния атом. Описание на квалификаторите е дадено в приложение II. Квалификаторът е асоцииран със съответния атом или връзка. Атом или връзка може да има повече от един квалификатор. За означение на специфични условия за атом, като йонизация, хибридизация, принадлежност към цикъл, брой на съседните протони, четност и др. се използуват запазени квалификатори. Пример:

c1c{H}c(C{sp3})c{H}cc1N=(O)=O.

Аналогично, запазени квалификатори се използуват, за да се зададе специфична конфигурация на съответната връзка, както е илюстрирано на Фигура 2.13. (Двете структури се различават само по взаимното разположение на бензолното ядро спрямо двойната връзка)



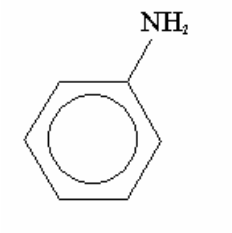
c1ccccc1-C={c}C-c1ccccc1



c1ccccc1-C={t}C-c1ccccc1

Фигура 2.13

Друг тип квалификатор е въведен за задаване числени ограничения върху произволен параметър. Този квалификатор се състои от име на параметъра и допустимите числени стойности във вид на горна и/или долна граница. Квалификаторът може да се отнася за цялата молекула, или за специфичен атом. В последния случай, квалификаторът се отнася за този атом или връзка, до който е в съседство. Например, на Фигура 2.14 е зададено ограничение зарядът на азотния атом да бъде по-голям от *-0.3* .

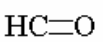


c1ccccc1N{H2}{-0.3<Q}

Фигура 2.14

Пример за квалификатор, отнасящ се за цялата молекула е {E_LUMO<0.5}, което означава енергията на LUMO орбиталата на молекулата да бъде по-малка от 0.5eV. Използуват се също запазени имена за т.нар. *динамични параметри*, които зависят едновременно от позицията на квалификатора и от контекста на RDL програмата. Такова запазено име е например *enumerate*, което означава, че трябва да се намери броят подфрагменти, зададени чрез дефиниция. Върху тези динамични параметри, също могат да се задават ограничения. Например, зададеното ограничение на

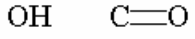
Фигура 2.15, означава да се намерят само молекули с повече от една група, като показаната на фигурата (алдехидна група).



C{H}=O{1<enumerate}

Фигура 2.15

В стандартния SMILES не могат да се задават несвързани химични структури. За целите на RDL програмата обаче, може да се наложи едновременно да се зададат два или повече несвързани фрагмента. Това се постига, чрез въвеждането на разделителя "_" в SMILES, който формално свързва две SMILES дефиниции. Заявка за намиране на подфрагмент, зададена чрез такъв *съвместен* SMILES означава, че всички зададени фрагменти трябва да присъствуват в обработваната молекула (контекста), за да бъде удовлетворена заявката. В смисъла на SMILES, разделителят "_" може да се разглежда като изкуствена (празна) връзка, или като липса на химична връзка. В RDL програмата обаче, тази изкуствена връзка може да има зададен квалификатор. Например, запазеното име *distance* се използува, за да се означи разстоянието между геометричните центрове на зададените фрагменти на текущата молекула. Показаната на Фигура 2.16 дефиниция, означава намиране само на тези молекули, където разстоянието между OH (хидроксилна) група и C=O (карбоксилна) група е между 1.5 и 2.6 Å.



O{H}_{1.5<DISTANCE<2.6}C{sp2}=O

Фигура 2.16

При използуване на запазеното име *tweak*, се търси конформер на текущата молекула, който да удовлетворява зададеното условие. Използува се насочено търсене на конформер, по метода, описан в [37].

Няколко дефиниции, разделени със запетая, образуват *съставна* дефиниция. Молекулата удовлетворява такава съставна дефиниция, ако удовлетворява поне

една от съставните и дефиниции. Всяка дефиниция, независимо проста или съставна, се приписва на *идентификатор* на дефиницията. Идентификаторът може да се използув по-нататък като съставна част от друга дефиниция. В смисъла на SMILES, идентификаторът може да се разглежда като предефиниран тип атом. На практика, при анализа на дефинициите се търси най-дългият подниз, съответствуващ на вече присвоен идентификатор. Ако такъв не се намери, поднизът се проверява за стандартните означения на атоми. Идентификаторът има смисъл на дефинирана константа, тъй като не може да му се присвоява стойност втори път. На Фигура 2.17 е показана извадка от *define* секцията на RDL програма. Текстът, заграден в (' … ') се възприема от RDL програмата като коментар.

```
XHalo: F,Cl,Br,I                    ('съставна дефиниция Xhalo, съдържаща халогенни
                                    атоми')
Benz:c1ccccc1                       ('дефиниция за молекули, съдържащи бензолно
                                    ядро')
Polyarom:Benz_C{ar}                 ('молекули с бензолно ядро и друга ароматна част')
Cunsat:C{sp1},C{sp2},C{ar}          ('ненаситен C-atom')
ABunsat:Cunsat-C{sp3}{acy}-
Xhalo
PolyarSide:Benz_C{sp2}{sc}
```

Фигура 2.17

Секция *Rules*

rules секцията се състои от дефиниции на правила, всяка от които на нов ред. Синтаксисът на дефинициите на правила е следният:

```
rule_identifier : булев израз
```

където *rule_ identifier* е **идентификатор на правило**. То се дефинира в секция *rules* и се използува в секция *apply* , или при дефиницията на следващи правила в секция *rules*.

Идентификатор е всяка поредица от букви, цифри и символа '_' ,която започва с буква.
Пример:
- RDD1
- RX_1
- Rabc

```
Q1
Q_4_2a
```

Булевият израз е израз, който се състои от идентификатори на правила и дефиниции на свойства, свързани с булевите оператори ‘*and*’, ‘*or*’ и ‘*not*’.

```
Пример:
defines
. . . .
 RX1:. . . .
 RX2:. . . .
. . . .
rules
 RDD3 : “RX1” and “RX2”
 RDD4 : “RX1” and not “RX2”
apply
 if “RDD4” then
  Selector := 4
else
 Selector := 0;
end.
```

В примера правилото *RDD3* ще получи стойност "истина" тогава и само тогава, когато и двете свойства *RX1* и *RX2* са изпълнени за текущата молекула. Съответно, *RDD4* ще има стойност "истина" ако *RX1* има стойност "истина" и *RX2* има стойност "лъжа".

Синтаксисът на булевия израз в *RDL* е аналогичен на този синтаксис в програмния език Pascal. Могат да се използуват вложени изрази и скоби.

Секция *Apply*

В *apply* секцията се задава последователността от правила, които трябва да се приложат върху данните (върху текущата молекула). Тя се състои от последователност от **оператори**, разделени с точка и запетая.

```
Пример:
if "rDist1" then
  if "rRDD1"   then ER_Binding:= 30
  else
    if "rRDD4" then  ER_Binding:= 28
    else  ER_Binding:= 10
else
```

```
 if "rDist2" then
   begin
     if "rRDD2"   then ER_Binding:= 26;
     if "rRDD4" then  ER_Binding:= 24 else ER_Binding:= 9;
     if "rDelta"      then ER_Binding:=5
     else ER_Binding:= 0;
   end;
end.
```

Операторите могат да бъдат прости и съставни. Простите оператори могат да бъдат следните:

if оператор
оператор_за_присвояване

Операторите се изпълняват последователно. Синтаксисът на if-оператора е следният:

```
if "ruleid" then statement1
```

ruleid трябва да бъде валиден идентификатор, а *statement1* може да бъде произволен оператор. При изпълнение на *apply* секцията, ако правилото *ruleid* получава стойност "истина" за текущата молекула то се изпълнява операторът *statement1*.

```
Пример 1:
if "rRDD4" then ER_Binding:= 26

Пример 2:
if "rRDD4" then
 begin
   ER_Binding:= 26;
   if "rRDD4" then ER_Binding:= 24;
 end
```

По-сложната версия на if- оператора позволява да се използува *else* клауза:

```
if "ruleid" then statement1 else statement2
ruleid must be a rule identifier; statement1 and statement2 must be statements.
```

ruleid трябва да бъде валиден идентификатор, а *statement1* и *statement2* са произволни оператори. При изпълнение на *apply* секцията, ако правилото *ruleid*

получава стойност "истина" за текущата молекула то се изпълнява операторът *statement1*, в противен случай се изпълнява операторът *statement2*.

Пример:
```
if "rRDD4" then  ER_Binding:= 24 else ER_Binding:= 9
```

Могат да се използуват вложени *if*-оператори:

```
if "rRDD4" then
  if "rRDD8" then ER_Binding:= 24
  else if "rRDD8" then ER_Binding:= 24
      else ER_Binding:= 9
```

else клаузата се съпоставя на последния срещнат *if* - горният пример е еквивалентен на:

```
if "rRDD4" then
  begin
    if "rRDD8" then ER_Binding:= 24
    else begin
          if "rRDD8" then ER_Binding:= 24
          else ER_Binding:= 9
        end
  end
```

Операторът за присвояване има следния синтаксис:

```
selector := realvalue
```

selector трябва да бъде валиден идентификатор на глобален параметър на молекулата (тези идентификатори са дефинирани в общ файл, валиден за всички програмни приложения, използуващи RDL), а *realvalue* е реално число.

Примери:
```
ER_Binding := 4
A_Max := -0.488
```

Съставният оператор, или оператор *'begin…end'* представлява списък от оператори, поставени между ключовите думи *begin* и *end*:

```
begin
   ER_Binding:= 26;
   if "rRDD4" then  ER_Binding:= 24 else ER_Binding:= 9;
   if "rDelta"      then ER_Binding:=5
   else ER_Binding:= 0;
end;
```

2.2.2. Реализация

Rule Interpreter (RI) е софтуерната среда за разработка и изпълнение на правилата. Правилата се записват чрез т.нар. език за описание на правила (Rule Description Language - RDL). Те дефинират различни механистични правила и основани върху тях правила за вземане на решения. Правилата се изпълняват върху бази данни от химични съединения, като се поддържат различни файлови формати. Правилата се записват като текстови файлове с разширение *.rul* .

RI е разработен на Borland Delphi, и е 32 битово Windows приложение. Използувани са голяма част от модулите на ниско ниво (структури данни, обработка на химичната информация) от OASIS QSAR системата, описана в [46,67]. RI предоставя текстов редактор, с възможности за работа с програмните файлове на езика (*.rul* файлове). В текстовия редактор програмата се показва разделена в три прозореца за всяка от секциите *define, rule, apply*. Има възможност за анализ на кода (*parse*) и изпълнение на програмата (*apply*). Анализът представлява синтактична проверка на кода и компилация до специфично вътрешно представяне. Обновяват се структурите, съдържащи дефинициите и идентификаторите на правила. При намиране на синтактична грешка курсорът се позиционира върху сгрешения ред и се дава индикация за вида на грешката. SMILES, използувани в кода могат да се визуализират.

Изпълнение на програмата е възможно след успешен анализ. При изпълнението е необходим входен файл с химична информация и изходен, където се записват модифицираните структури.

Освен като самостоятелно приложение, RI е вграден във всички останали приложения, разработени в лабораторията по Математична Химия в Университет "проф. А.Златаров" - Бургас.

2.3. Изводи

Предложеният метод COREPA (Common Reactivity PAttern) [62,64] позволява да се извличат правила за оценка на биологичната активност на различни по структура

химични съединения. COREPA оценява вероятностното разпределение на данните и използува класификатор на Бейс за определяне критериите за принадлежност към даден клъстер. Методът COREPA не изисква наличието на детайлно описание на рецептора, нито тримерно наслагване на структурите. Предимството на метода е, че работи с ограничен набор стереохимични параметри и позволява извеждането на правила, обясняващи механизма на действие. Въз основа на извлеените критерии се построява дърво на решението , което представлява модел на дадено (биологично) свойство. Във всеки възел на дървото е поставено условие, задаващо критерий, свързан със структурата на молекулата или други изчислени или експериментално получени параметри. Правилата се записват чрез средствата разширен SMILES и на разработения формален език (т. 2.2) и могат да се прилагат многократно върху различни бази данни.

Дървото на решението всъщност представлява класификатор, който класифицира химичното съединение в една от няколко групи (например на активни и нактивни съединения), въз основа на изведения модел за механизма на действие на биологичното свойство.

След като моделът е построен, и съединенията класифицирани, то се прилага метод за количествена оценка на активността (например 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.

Правилата се записват чрез средствата на разработения формален език *Rule Description Language* и могат да се прилагат многократно върху различни бази данни. *RDL* позволява задаване на множество различни условия за химичното съединение, в това число условия за свързаността, пространственото разположение, както и електронни узловия. Функционалните групи се задават чрез стандартно SMILES [97] означение. Разширение на SMILES позволява задаване на стереоинформация за фрагментите и молекулите. Чрез ключовата дума DISTANCE могат да се дефинират разстояния между произволни фрагменти и да се задават

ограничения за тях (например {DISTANCE < 11}). Чрез ключовата дума TWEAK се дефинират разстояния между фрагменти, които се търсят чрез насочен tweak. Ограничения могат да се задават и на произволен параметър на структурата (заряд, EHOMO и др.). Езикът позволява комбинирането на дефинираните условия в правила чрез използуване на логически оператори *and, or, not,* което осигурява гъвкавост и приложимост на подхода в множество различни ситуации.

Правилата позволяват компютърна реализация и могат да бъдат достатъчно подробни, за да бъдат използувани за предсказване и търсене. Предложеният подход е общо решение на проблема, тъй като правилата позволяват комбиниране на подфрагментно търсене с молекулни дескриптори, стереохимични конфигурации и взаимно 3D разположение на функционалните групи.

3. Методи за генериране на нови обекти

Химичните съединения се описват с тяхната свързаност (молекулен граф, където на атомите съответстват върхове на графа, а на връзките между атомите - ребра на графа) и пространствено разположение на атомите. Пространственото разположение на атомите и неговото влияние на свойствата на веществата се изучава от науката стереохимия. Конформационният анализ е важен раздел от стереохимията и се занимава с изучаване на молекулните конформации и превръщанията между тях. Обикновено с термина "конформация" се наричат неидентичните положения на атомите в молекулата, които се получават при въртене на една или повече връзки без те да се прекъсват. В по-широк смисъл, терминът се употребява за да обозначи структури, получени след промяна на торзионни ъгли. Терминът "конформер" се използува за множеството конформации, съответстващи на локален енергиен минимум. В биологична среда обаче, конфигурацията на молекулата зависи силно от обкръжаващата я среда, което се отчита при изследване и моделиране на биологичните свойства. В тази връзка се допуска съществуването на безкрайно множество конформери с енергия под даден праг, вместо дискретно множество конформери, съответстващи на локалните енергийни минимуми [38,62,63,64,87,88]. Тази хипотеза се използува в моделирането и гъвкавото търсене в бази данни [16,40, 75]. Множеството възможни конформации се нарича конформационно пространство.

В конформационния анализ генетични алгоритми се използуват за намиране на потенциалните нискоенергетични състояния. Предложеният метод има за цел оптимално покритие на конформационното пространство, като генерира множество от различни конформери, удовлетворяващи зададени енергетични условия.

3.1. Генетичен алгоритъм за намиране на множество конформери, най-добре покриващи конформационното пространство (Genetic Algorithm Search)

Разработен е генетичен алгоритъм, който намира малък брой конформери с ниска енергия, максимално покриващи конформационното пространство. Вместо изчерпателно търсене на всички конформери се използува 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 тридименсионално евклидово пространство се дефинира като [86]:

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

където N е броят на точките на X и Y.

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

Основни принципи

В генетичните алгоритми обектите (индивидите) се представят с хромозоми. Хромозомите представляват последователности от битове (гени), които съдържат информация за свойствата на индивидите. Приспособеността (годността) на индивида зависи от тези свойства. В молекулния дизайн хромозомите определят молекула или конформер, докато критерият за годност обикновено оценява разположението и ориентацията на функционални групи, разпределението на заряди, молекулна форма електростатично поле и др. Генетичният алгоритъм започва със запълването на случайна популация с размер N_p, (размер на "постоянната" популация). Постоянната популация се разширява с N_c нови индивиди (деца). От разширената популация с (N_p+N_c) индивида според зададените критерии се избират N_p представителя, които ще образуват следващата популация. Разширението на популацията и следващото селективно избиране на постоянна популация представлява отделна еволюционна стъпка. Еволюцията е итеративен процес, който се повтаря докато не се достигне критерий за край. Критерият за край включва тест за сходимост, който проверява дали годността не се подобрява за последните 2 или повече поколения. Генерирането на нови индивиди се осъществява с генетичните операции мутация и кръстосване.

Описанието на предложения генетичен алгоритъм за генериране на конформери се състои от:

- Описание на кодировката на хромозомите;
- Блок-схема на алгоритъма;
- Описание на запълването на първоначалната популация;
- Описание на критерия за оценка на популацията
- Описание на генерирането на нова популация
- Описание на критерия за край
- Анализ на резултатите, получени от работата на генетичния алгоритъм

Кодиране на хромозомите

При предложения метод се използуват 4 типа гени, кодиращи ротационните връзки, четността на стереоцентровете, свободните ъгли и пирамидите. Гени от четирите различни типа се комбинират в хромозома. Промяната на който и да е от тези гени, води до промяна в конфигурацията на структурата. Информацията, записана в гените, влияе върху конформационните степени на свобода, както и върху стереохимичните свойства. Гените, които модифицират стереохимичните свойства могат да бъдат забранени или не от потребителя. Ако стереохимичните гени са позволени, то 3D структурите могат да генерират различни конфигурации. Тук ще използуваме термина конформер в по-широк смисъл, като пространствена структура, която има уникални хромозоми.

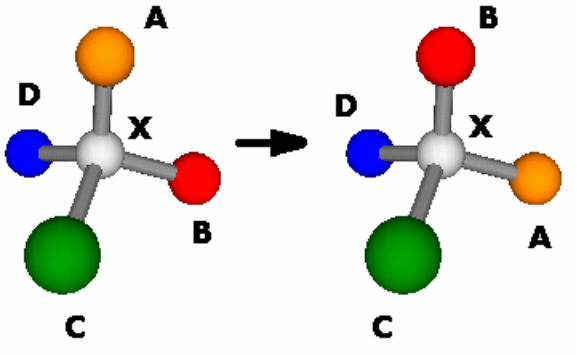
Торзионни ъгли

Всички торзионни ъгли, които се образуват от единични и двойни връзки и не принадлежат на цикли се кодират с гени. Гените за различни торзионни ъгли имат различни на брой възможни стойности. По подразбиране броят възможни стойности се задават според броя на бариерите (минимумите) на периодичните торзионни потенциали за съответната връзка. Например, ако генът има 3 възможни стойности, то той задава ъгли 120°, 240°, 360°. Броят им може да бъде променян от потребителя. Генът, кодиращ тези връзки е с дължина *8N*, където *N* е броя на ротационни връзки в молекулата.

Стерео центрове

Стереоцентровете се кодират с гени, които могат да се инвертират, без да е необходимо прекъсване на връзка. Разглеждат се всички атоми с четири връзки и поне три различни съседа. От тях се избират само тези, за които изтриването им води до отделяне на поне една от двойките съседи. Генът има две възможни стойности и описва четността на стереоцентъра. Четността се определя от знака на смесеното векторно произведение на векторите *X-A, X-B* и *X-C* (Фигура 3.1). Мутация, приложена към този ген, инвертира четността на стереоцентъра.

Инверсията се получава геометрично чрез завъртане на 180° на *A* и *B* и всички свързани към тях атоми. *A, B* и *X* могат да принадлежат към общ цикъл, но нито *C,* нито *D* не трябва да принадлежи към същия цикъл. По подразбиране инверсия на стереоцентрове е позволена само за посочените във входната структура стереоцентрове.



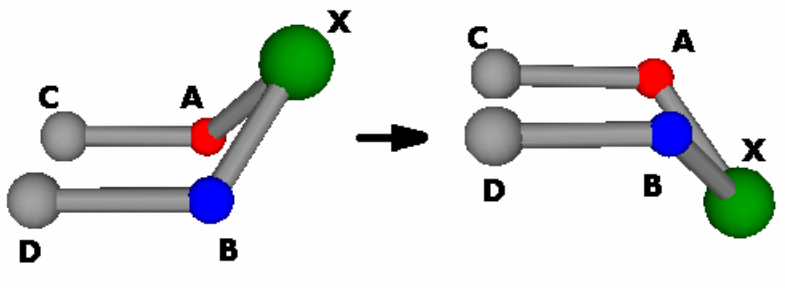
Фигура 3.1

Свободни ъгли

Свободните ъгли са дефинирани от Payne & Glen [75] като двусвързани циклични атоми, намиращи се извън равнината на цикъла, свързани с прости връзки и имащи за съседи само двусвързани атоми. Свободният ъгъл е илюстриран на Фигура 3.2. Всичките показани атоми принадлежат на общ цикъл, а свободният ъгъл е фрагментът *A*-*X*-*B*. С изключение на двойката *C* и *D*, никои други две двойки не трябва да принадлежат на друг общ цикъл. Обръщането може да се разглежда като две едновременни ротации в противоположни посоки, на приблизително 120° около връзките *C*-*A* и *D*-*B*. От гледна точка на *X* и свързаните към него фрагменти извън свободния ъгъл, трансформацията е ротация около ос, прекарана през *A* и *B*. Следователно обръщането на свободен ъгъл не влияе върху стереохимията. Свободният ъгъл е идеален, ако позволява обръщане, което променя само торзионния ъгъл. Това е възможно само ако C-A и D-B са успоредни (и определят равнина).

В предложената версия е възприето по-слабо ограничение за свободни ъгли. Позволява се обръщане на "неидеална равнина", като се въвежда степен на

равнинност, съответно позицията на свободния ъгъл извън равнината се проверява според зададен праг за равнинност. Очевидно е, че по-слабата дефиниция позволява по-детайлен анализ на конформационното пространство за полициклични структури. Дължините на връзките и ъглите могат да се коригират чрез ограничено силово поле.



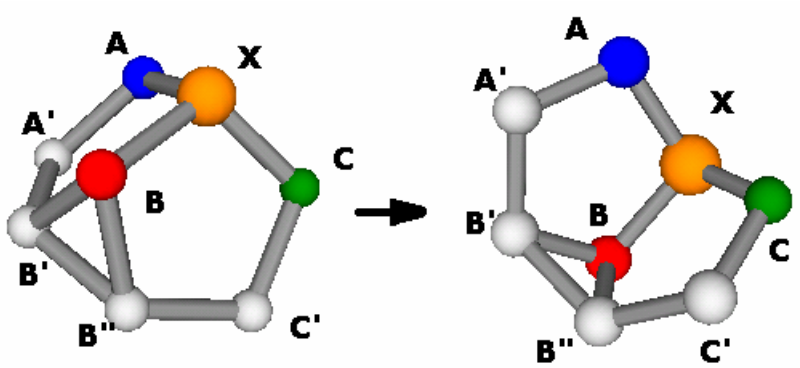
Фигура 3.2

Пирамиди

Обръщането на свободен ъгъл е евристично решение, което все още не решава напълно проблема за гъвкавостта на циклични структури. В настоящата работа е предложено още една модификация, наречена "обръщане на пирамиди". Нека X принадлежи на поне два кондензирани цикъла, заедно с първите си съседи A, B, C (Фигура 3.3). Всички атоми показани на схемата принадлежат на един полицикличен фрагмент. Атомите A, B, C дефинират основата на пирамидата. Отражението на пирамидата е огледално отражение спрямо основата на пирамидата. Отражението включва X и всички части, свързани към X, A, B или C, без показаните. Всички потенциални стереоцентрове, включително A, B и C се инвертират от тази операция. Ако четността на A, B, C или X е фиксирана, то отражението е забранено. Отражението се извършва винаги преди други операции върху структурите.

В общия случай отражението влияе върху валентните ъгли на A, B и C, такива като X-A-A' и X-B-B'. Тези валентни ъгли не се нарушават само ако цикличният втори съсед на X, т.е. A', B', B" и C" са точно в равнината на основата на пирамидата. Освен специфичната свързаност, 3D структурата трябва да отговаря на условието

за равнинност на A-A' и B-B'. За тази цел потребителят може да избере праг на равнинност.



Фигура 3.3

Списъкът от четности на стереоцентровете позволява генетичния алгоритъм да бъде ограничен до стереохимията на входния конформер. Всяко обръщане на цикъл използува един бит, всяко обръщане на пирамида използува също един бит.

Описание на алгоритъма

На

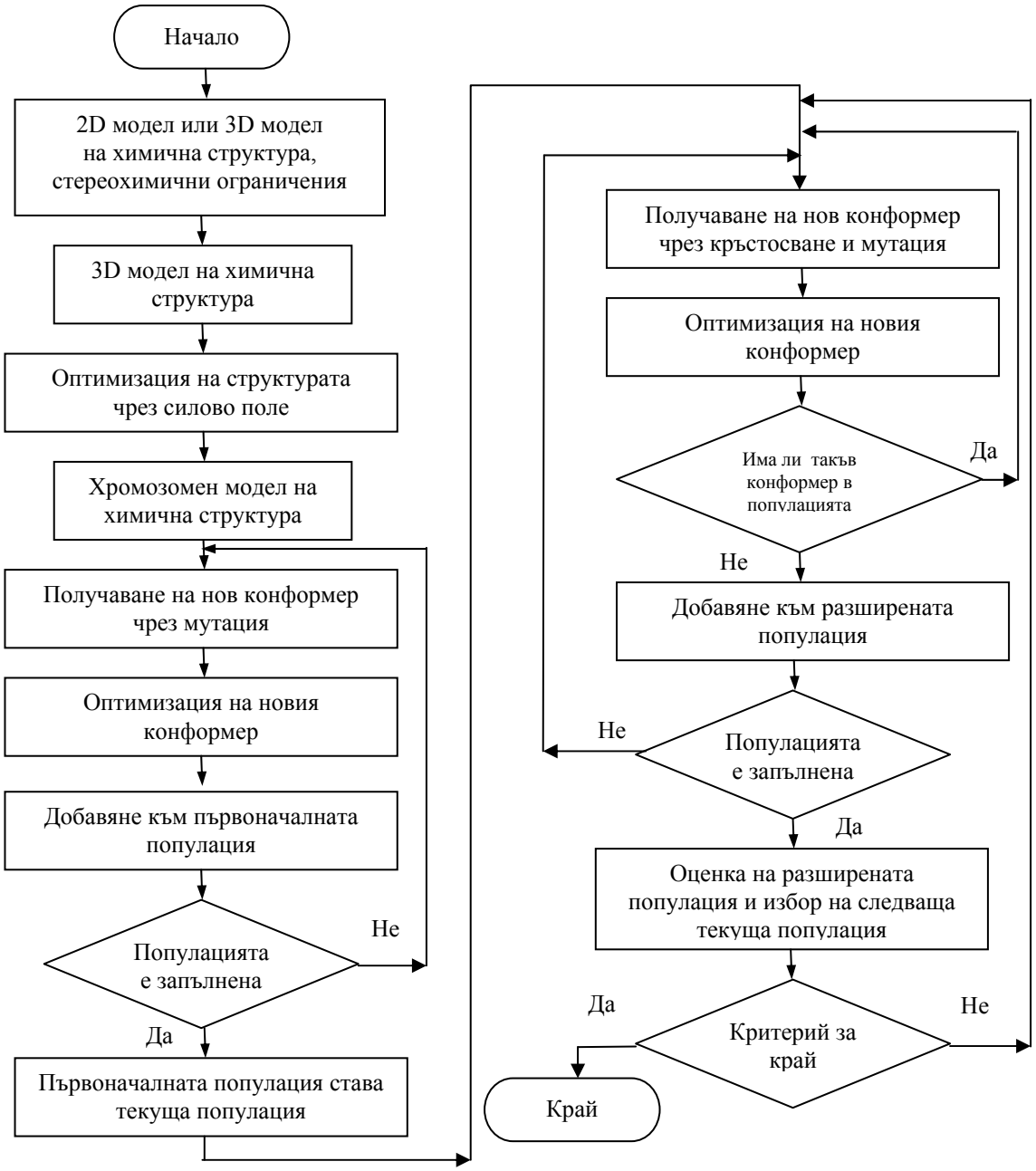
Фигура 3.4 е показана обобщена блок-схема на предложения алгоритъм за генериране на конформери. Входът на алгоритъма представлява 3D молекулен модел (поддържа се както специфичен файлов формат[67], така и разпространените SDF, MOL, MOL2 файлови формати [19]) или разширен SMILES[97]. Стандартният SMILES обозначава само молекулната свързаност, докато разширената версия съдържа и стереохимични ограничения. При вход от SMILES се прилага конвертиране до 3D. При 3D вход всички стереохимични свойства са фиксирани. След това се прилага силово поле (Pseudo Molecular Mechanics) [38] за намаляване на потенциалната енергия.

Преди започване на генерацията, всички стереохимични свойства и конформационни степени на свобода се превръщат в хромозоми и се показват, така че гените да могат да бъдат избрани или забранени от потребителя. Ако е използуван 3D молекулен модел, то всички стереохимични свойства са забранени за модификация. В противен случай намерените, но не указани в SMILES са

позволени за модифициране. Всеки конформер се представя от хромозоми и тримерен (3D) молекулен модел. Генетичните операции се извършват на ниво хромозоми За минимизиране на потенциалната енергия, новата 3D структура може да бъде подложена на оптимизация със силово поле. По този начин се освобождават напрегнатите структури. При прилагане на оптимизация е необходим 3D модел, който се конструира от хромозомния модел. Това се извършва чрез сравнение на хромозомите с тези на изходната структура и прилагането на съответните структурни модификации. Конформерът може да бъде отхвърлен ако потенциалната енергия надвишава даден праг. Ако химичният граф има елементи на симетрия, различните хромозоми могат да доведат до една и съща конформация. Еднакви конформери, които имат различни хромозоми се наричат изродени. Новите конформери се проверяват за израждане. Това се постига чрез групата на симетрия Π на химичния граф. Последователността на хромозомите се преобразува според всички пермутации $p \in \Pi$, запазвайки инвариантен химичния граф и се проверява за идентичност със съществуващите конформери.

Генериране на първоначалната популация

Първоначалната популация се получава от изходната 3D структура чрез случайни мутации на хромозоми. За да се осигури еднаква вероятност за различните модификации, множеството гени и стойностите им се избират случайно. Мутиралите хромозоми се конвертират в 3D модел. Конформерът може да се отхвърли, ако е изроден или напрегнат. Изходната 3D структура също се включва в популацията.


Начало
2D модел или 3D модел на химична структура, стереохимични ограничения
3D модел на химична структура
Оптимизация на структурата чрез силово поле
Хромозомен модел на химична структура
Получаване на нов конформер чрез мутация
Оптимизация на новия конформер
Добавяне към първоначалната популация
Популацията е запълнена
Не
Да
Първоначалната популация става текуща популация
Получаване на нов конформер чрез кръстосване и мутация
Оптимизация на новия конформер
Има ли такъв конформер в популацията
Да
Не
Добавяне към разширената популация
Популацията е запълнена
Не
Да
Оценка на разширената популация и избор на следваща текуща популация
Критерий за край
Да
Не
Край

Фигура 3.4

Оценка на популацията

Според традиционния генетичен алгоритъм, след декодирането на хромозомите, новите структури се оценяват индивидуално според зададените условия. В дадения случай генетичният алгоритъм има за цел да генерира множество конформери, които покриват най-добре конформационното пространство. Това води до идеята за едновременна оценка на цялата популация, а не на индивидуалните и членове, както в стандартните генетични методи. Общият брой членове на популацията е N_{total}, от които A на брой родители и (N_{total} - A) деца. За тази цел се използува *RMS* [86] разстояние между конформерите. Новото поколение структури, генерирано на дадена стъпка на еволюцията се оценява, като се пресмята разстоянието rms_{ij} между i-тата нова структура и j-тия член на цялата популация. При размер на популацията N_{total}, броят на необходимите сравненията е $N_{total}*(N_{total}-A)/2$. След това се решава комбинаторната задача за избиране на подмножество от A конформера от общо N_{total}, които имат максимална средна сума от разстояния rms_{ij}.

$$\mathrm{rms}_{ij} = \min_{p \in \Pi} \left\{ \sqrt{\frac{1}{m} \sum_{k=1}^{m} (r_{ik} - r_{ip_k})^2} \right\} \qquad (3.1)$$

Където r са радиус вектори, m е броя на атомите и $p=\{p_1, p_2, \ldots, p_m\}$ са пермутациите на номерата от 1 до m. Вземат се под внимание всички пермутации от групата на симетрия Π на химичния граф, без тривиалната пермутация $p_k=k$. Минимизацията за дадена $p \in \Pi$ се намира числено, чрез линейно търсене в пространството на ротационните степени на свобода на молекулата. Индивидуалната оценка на конформера i е сума на RMS разстоянията до всички останали индивиди от популацията.

$$\mathrm{rms}(i) = \sum_{\substack{j \in S \\ j \neq i}} \mathrm{rms}_{ij} \qquad (3.2)$$

Множеството S от N конформера се оценява с т.нар. средна разлика (average dissimilarity – AD) , която се дефинира като сума на rms за всички двойки конформери, разделено на броя на двойките:

$$\mathrm{rms}(S) = \frac{1}{N(N-1)} \sum_{\substack{j \in S \\ j \neq i}} \mathrm{rms}_{ij} \qquad (3.3)$$

Генериране на новата популация

Според по-горе описаната процедура се избира подмножеството с най-голямо разнообразие, което се използува за родителска популация при следващта еволюционна стъпка. *Изборът* на подмножеството, както и последващите *кръстосване* и *мутация* са генетичните операции, използувани за конструиране на следващата популация. При операцията *кръстосване* двете родителски хромозоми се разделят на едно и също случайно избрано място и съответните части се съединяват за да се получат хромозомите на децата. Това е т.нар. едноточково кръстосване. В предложения метод се използува 4-точково кръстосване, тъй като 4-те различни гена се кръстосват поотделно с използуване на едноточково кръстосване. По този начин хромозомите на новополучения индивид винаги съдържат части от хромозомите на двамата родителя. Вероятността конформерът да бъде избран за родител е пропорционална на неговата стойност за RMS спрямо популацията.

Преди извършването на кръстосване, родителските хромозоми се подлагат на *мутация*. Мутациите са средство за предотвратяване на последователно генериране на едни и същи индивиди. Вероятността за мутация се задава от потребителя и определя вероятността, с която родителските гените ще бъдат модифицирани преди да се извърши кръстосване. Стойността по подразбиране е 1%.

Полученият нов конформер не се добавя автоматично към разширената популация. Той може да бъде отхвърлен ако е изроден или с неприемлива енергия. Максималният брой на опитите да се получат неизродени конформери с ниска енергия се задава от потребителя. Нови конформери се генерират докато се запълни разширената популация или се надвиши този брой. В последния случай алгоритъмът се прекъсва. Ако разширената популация се запълни успешно, то от нея се избират N_p индивида според следния критерий: подмножеството S_p от N_p от общо $N_p + N_c$ индивида трябва да съдържа най-различните конформери .

Комбинаторната задача за избора им се решава с известния branch-and-bound алгоритъм.

Генерираните конформери се подлагат на оценка на енергията, за да се избегнат структурите със стерични конфликти. Всички конформери, с енергия по-голяма от зададен от потребителя праг се елиминират като неуспешни опити, или в зависимост от потребителска конфигурация могат да бъдат оптимизирани.

Новото поколение индивиди се проверява за израждане, като се елиминират структурите, различаващи се по торзионни ъгли под зададен праг. Броят на опитите да се получи енергетично стабилен индивид също се задава от потребителя.

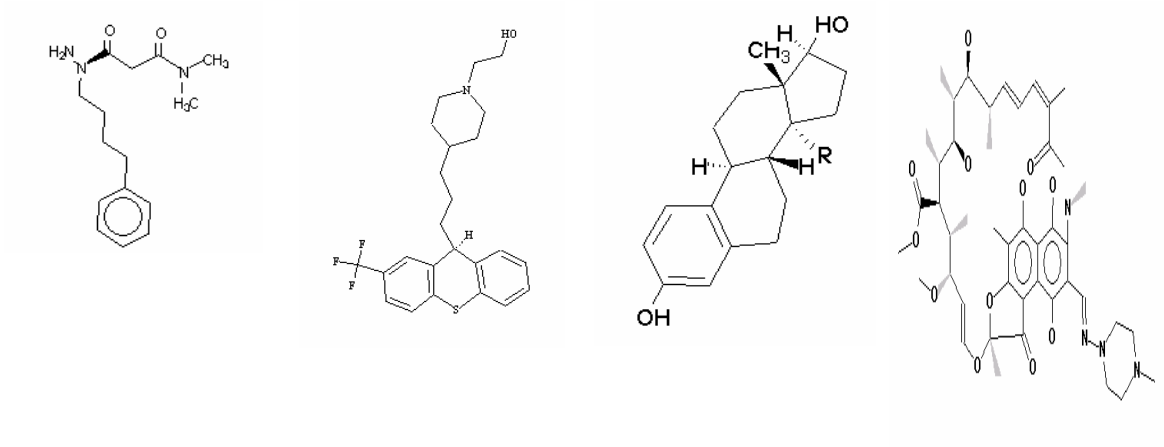
Критерий за край

При развитие на еволюционния процес, популацията на индивидите се подобрява като се натрупват индивиди с високо разнообразие. Използуват се 3 критерия за спиране на процес. Първият задава максималнният брой еволюции. Вторият следи сходството на стойностите на средното различие на популацията и алгоритъмът спира ако разликата между стойностите е по-малка от дефинирана от потребителя стойност. Третият задава максималната стойност на RMS разстоянието.

Резултати

Изследвани са три качествени критерия за оценка на изпълнението на алгоритъма: стабилност, възпроизводимост и покритие на конформационното пространство. Генетичният алгоритъм е приложен към четири тестови молекули с различна гъвкавост (Фигура 3.5) и са проведени няколко експеримента с различни параметри. Структура 1 има нециклична част с ротационни степени на свобода. Структура 4 подлежи на структурни деформации както в цикличната, така и в нецикличната част. Структура 3 представлява конформационно твърда молекула.

Структура 4 притежава голям гъвкав цикъл. Изображенията на фигура 3.1.4 са получени от разширен SMILES , с вградения в софтуера конвертор.



Структура 1 Структура 2 Структура 3 Структура 4

Фигура 3.5

Както е показано в **Таблица 3.1** - Таблица 3.4 при различните изпълнения на алоритъма са използувани различни управляващи параметри – размер на постоянната популация N_p, брой дъщерни индивиди N_c, т.нар. праг на израждане, който представлява ъгъла, над който два торзионни ъгъла се считат различни; отношение мутация/кръстосване (*m/c* отношение); праг на потенциалната енергия *E, kJ/mol*. В експериметите не е използувана оптимизация със силово поле, с цел да се намали времето за изпълнение.

Параметър	I	II	III	IV	V	VI	VII	VIII
m/c	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8
wdW-E	100	200	100	200	100	200	100	200
P-popul	10	10	10	10	10	10	10	10
C-popul	5	5	5	5	5	5	5	5
degener	60	60	60	60	30	30	30	30
Параметър	IX	X	XI	XII	XIII	XIV	XV	XVI
m/c	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9
wdW-E	100	100	100	100	100	100	100	100
P-popul	20	30	10	14	20	5	7	10
C-popul	5	5	7	7	7	5	7	10
degener	60	60	60	60	60	60	60	60

Таблица 3.1

Параметър	I	II	III	IV	V	VI	VII	VIII
M/c	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8
wdW-E	100	200	100	200	100	200	100	200
P-popul	10	10	10	10	10	10	10	10
C-popul	5	5	5	5	5	5	5	5
degener	60	60	60	60	30	30	30	30
Параметър	IX	X	XI	XII	XIII	XIV	XV	XVI
m/c	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9	0.1/0.9
wdW-E	100	100	100	100	100	100	100	100
P-popul	20	30	10	14	20	5	7	10
C-popul	5	5	7	7	7	5	7	10
degener	60	60	60	60	60	60	60	60

Таблица 3.2

Параметър	I	II	III	IV	V	VI	VII
m/c	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8	0.1/0.9	0.1/0.9	0.2/0.8
wdW-E	500	1000	500	1000	500	1000	500
P-popul	3	3	3	3	3	3	3
C-popul	1	1	1	1	2	2	2
degener	30	30	30	30	30	30	30
Параметър	VIII	IX	X	XI	XII	XIII	XIV
m/c	0.2/0.8	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8	0.1/0.9	0.1/0.9
wdW-E	1000	500	1000	500	2000	500	1000
P-popul	3	4	4	4	4	4	4
C-popul	2	1	1	1	1	2	2
degener	30	30	30	30	30	30	30

Таблица 3.3

Параметър	I	II	III	IV	V	VI
m/c	0.1/0.9	0.1/0.9	0.2/0.8	0.2/0.8	0.1/0.9	0.1/0.9
wdW-E	500	1000	500	1000	500	1000
P-popul	5	5	5	5	5	5
C-popul	3	3	3	3	3	3
Degener	60	60	60	60	30	30
Параметър	VII	VIII	IX	X	XI	XII
m/c	0.2/0.8	0.2/0.8	0.1/0.9	0.2/0.8	0.1/0.9	0.1/0.9
wdW-E	500	1000	1000	1000	1000	1000
P-popul	5	5	7	7	10	10
C-popul	3	3	3	3	3	5
Degener	30	30	60	60	60	60

Таблица 3.4

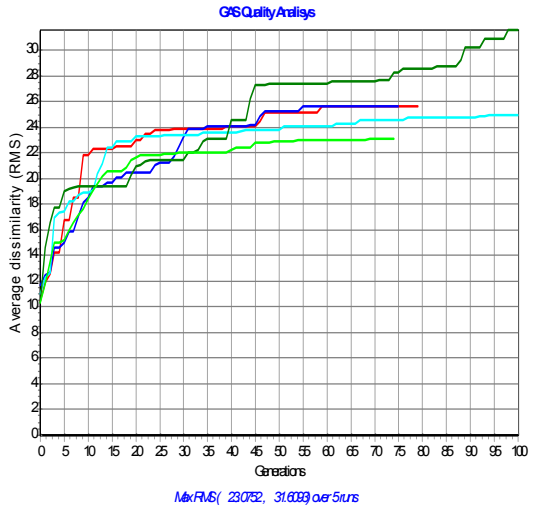
Броят опити за намиране различни дъщерни индивиди с ниска енергия е фиксиран на 100 и е увеличен само в някои случаи. Броят итерации е постоянен и равен на

100 за всички експерименти. В повечето случаи генерацията завършва поради достигане на сходимост. Това означава, че средната разлика между конформерите (*Average Dissimilarity, AD)* нараства с по-малко от 0.1Å за 20 итерации.

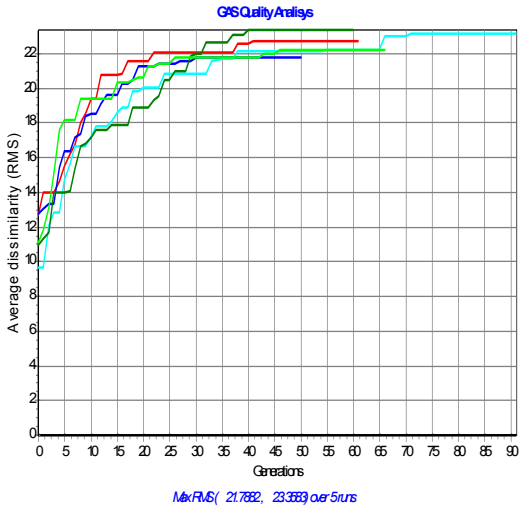
За структура 1 четностите на стереоцентрове не се запазват. За структура 2 се запазва само четността на един атом. За структури 3 и 4 се запазват само четностите на стереоцентровете, показани на фигура Фигура 3.5.

Както е показано в таблицата, управляващите параметри се различава за четирите структури. За структури 3 и 4, които са по-малко гъвкави от 1 и 2, е използуван по-висок праг на енергията, както и по-малък размер на постоянната популация N_p и по-малък брой деца N_c.

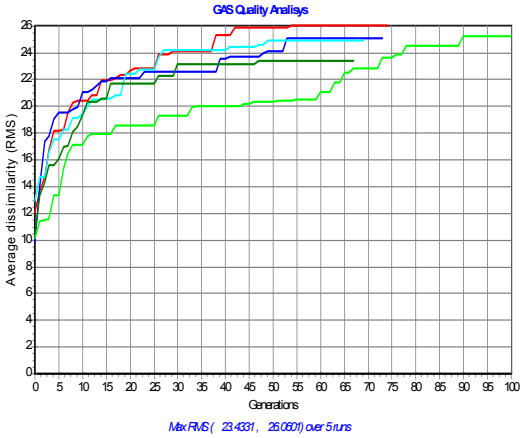
Експериментите включват 5 последователни изпълнения за структури 1,2 и 4 и 10 последователни изпълнения за структура 3. Всичко са направени 360 изпълнения при 58 различни набора параметри. За всички експерименти, измененията на средното различие в последователните генерации е показано на Фигура 3.6-Фигура 3.9. съответно за структури 1-4. Графиките илюстрират анализа на резултатите, даден по-долу.



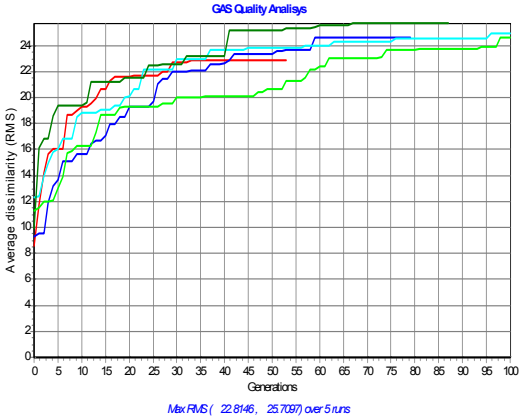
I



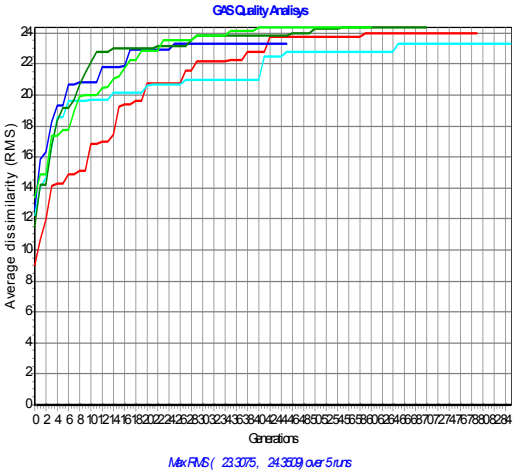
II


GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (23.4331 , 26.0601) over 5 runs

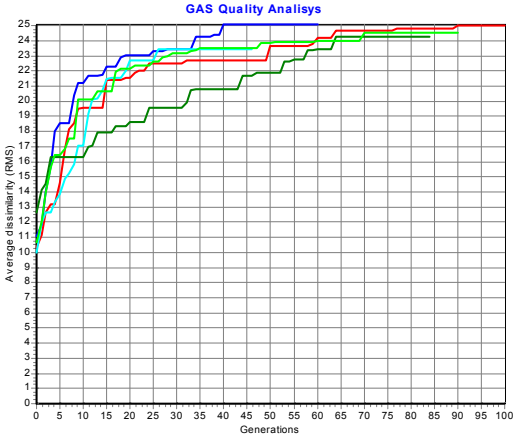
III


GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (22.8146 , 25.7097) over 5 runs

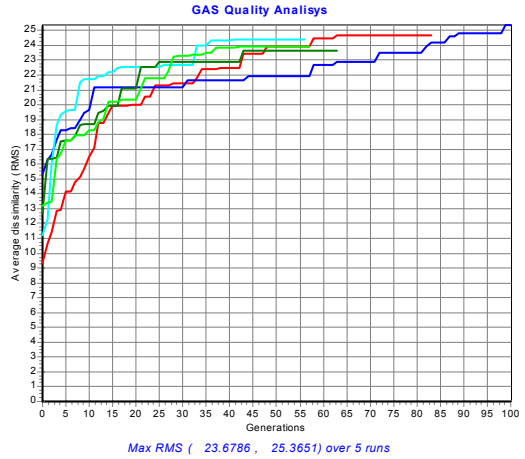
IV

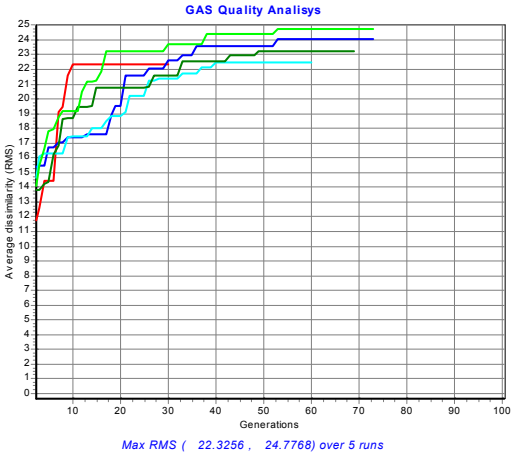

GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (23.3075 , 24.3609) over 5 runs

V

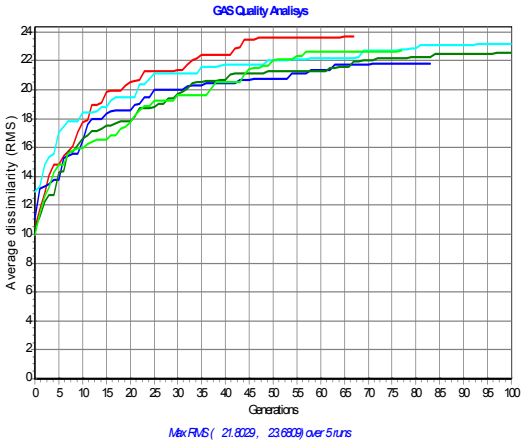

GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (23.4820 , 25.0810) over 5 runs

VI

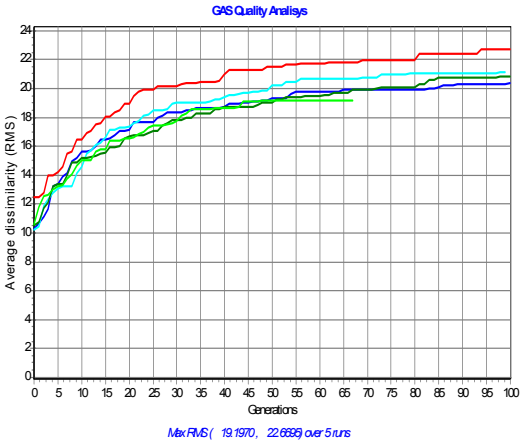

GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (23.6786 , 25.3651) over 5 runs


GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (22.3256 , 24.7768) over 5 runs

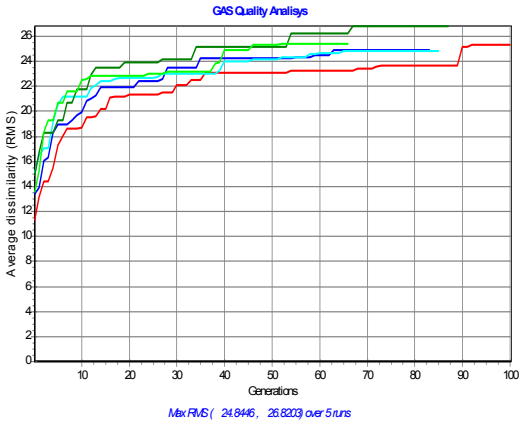
VII



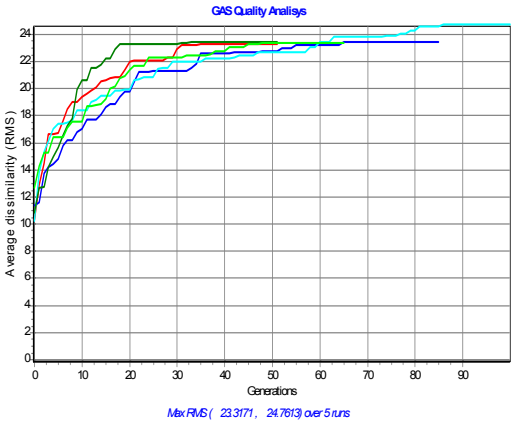
VIII



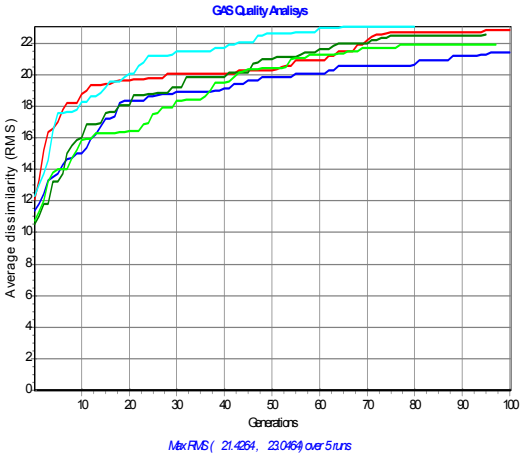
IX



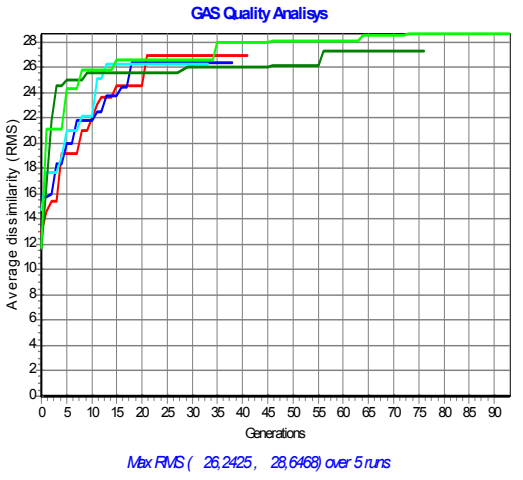
X



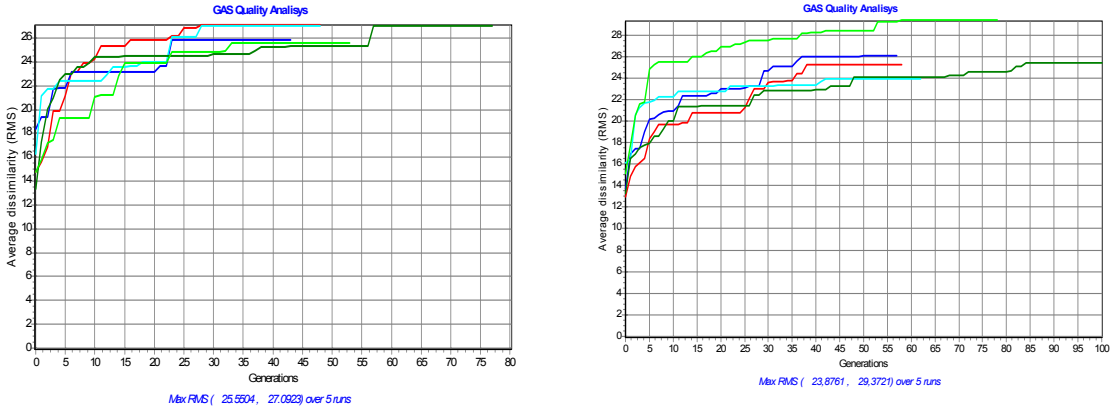
XI



XII

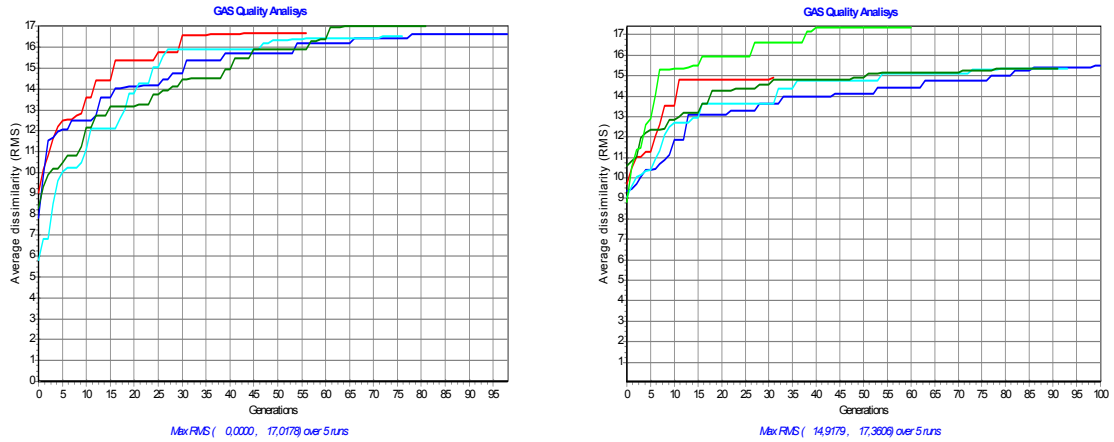


XIII XIV

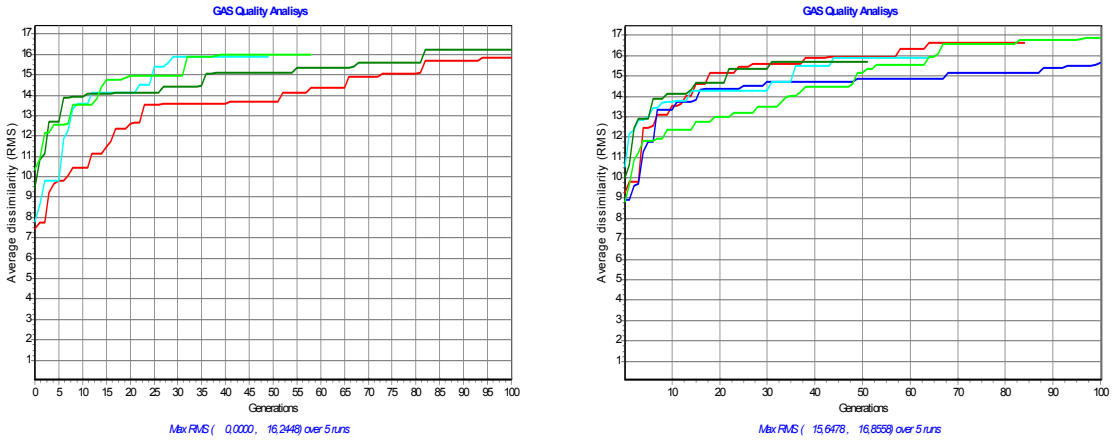


XV XVI

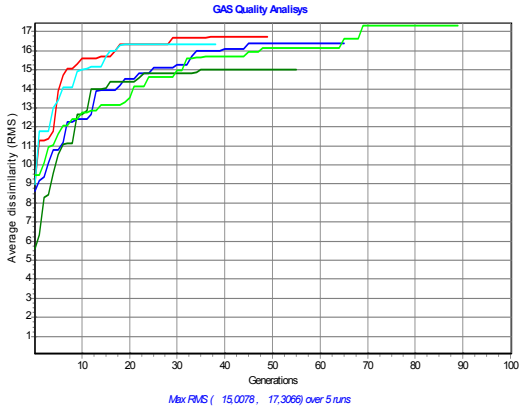
Фигура 3.6



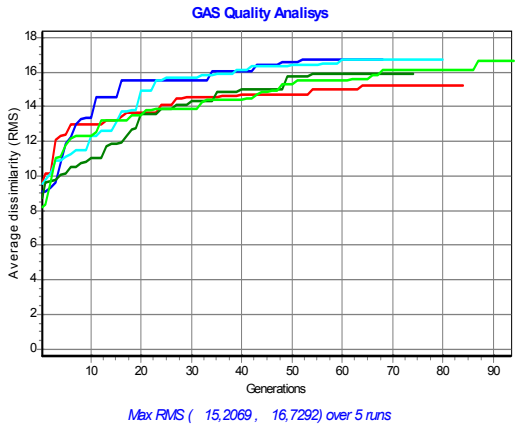
I II



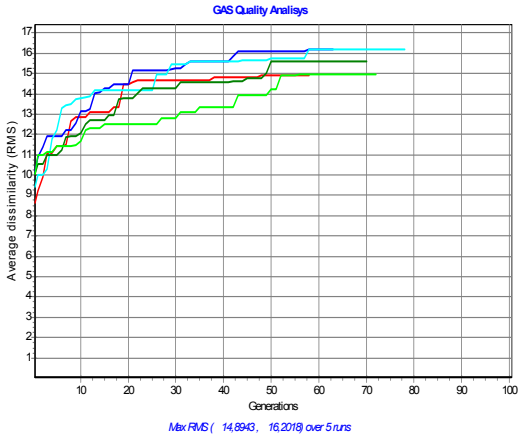
III



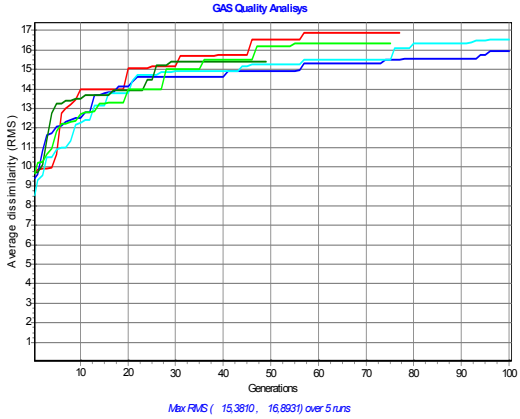
IV



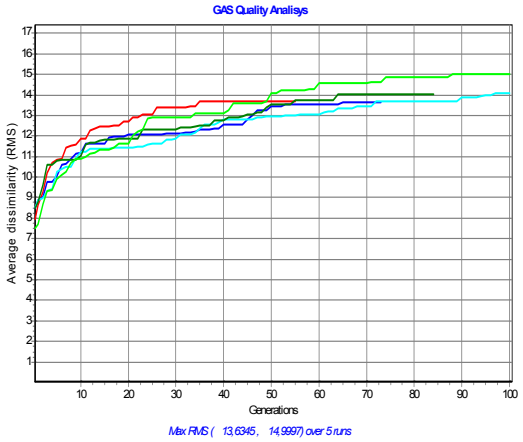
V



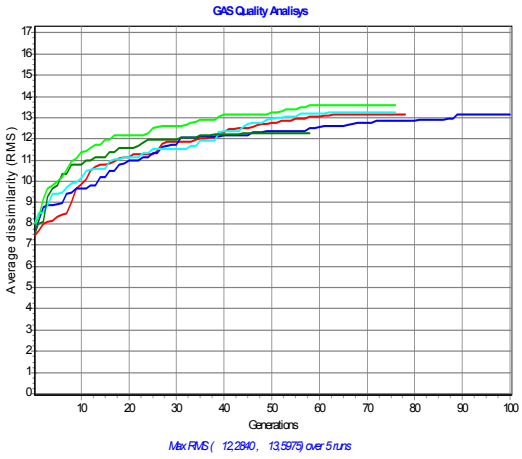
VI



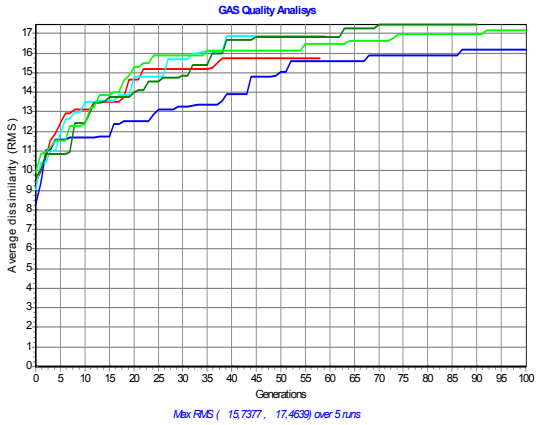
VII



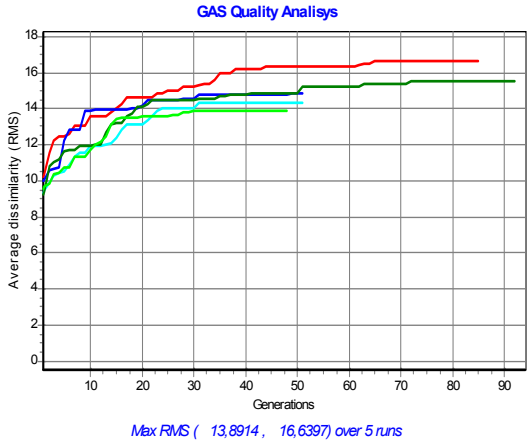
VIII



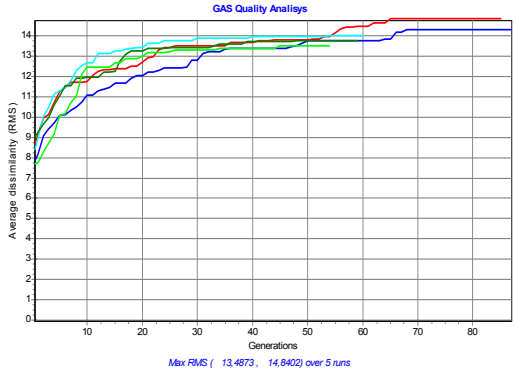
IX



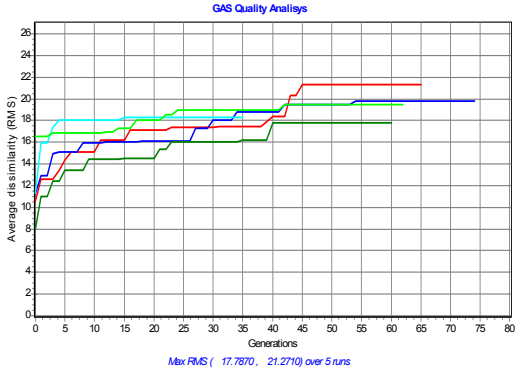
X



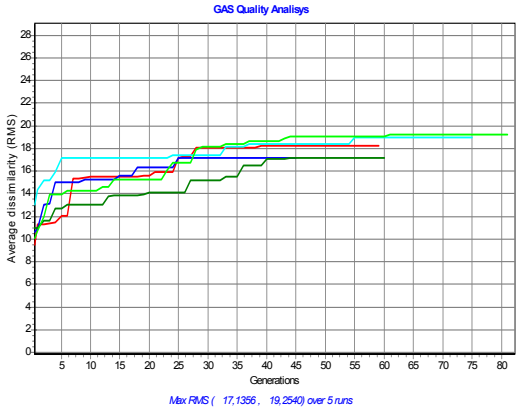
XI



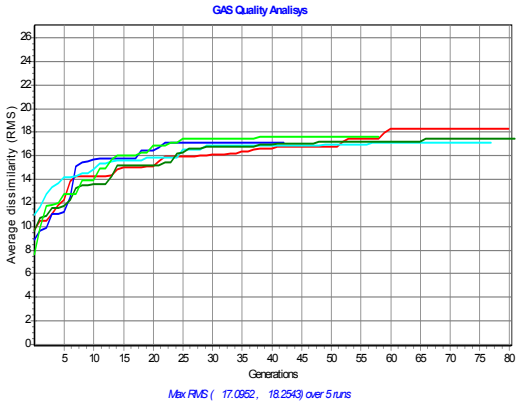
XII



XIII



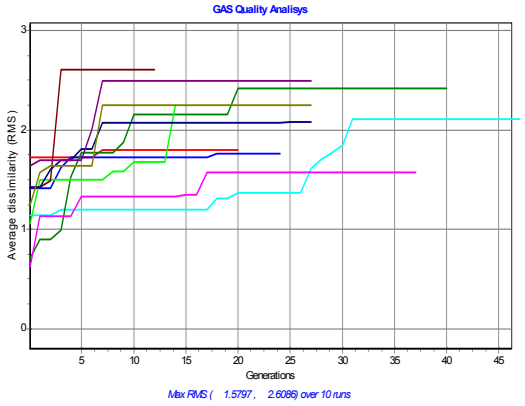
XIV



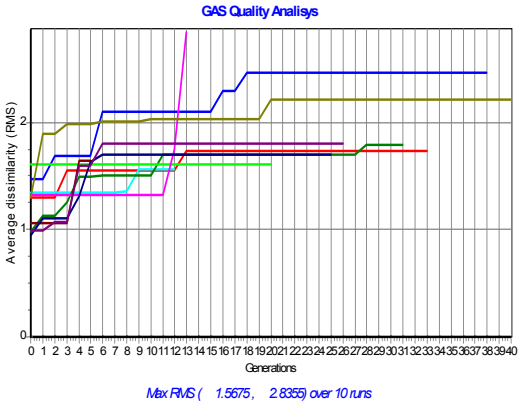
XV

XVI

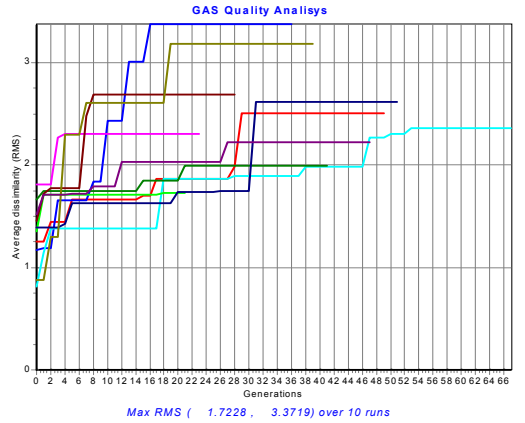
Фигура 3.7



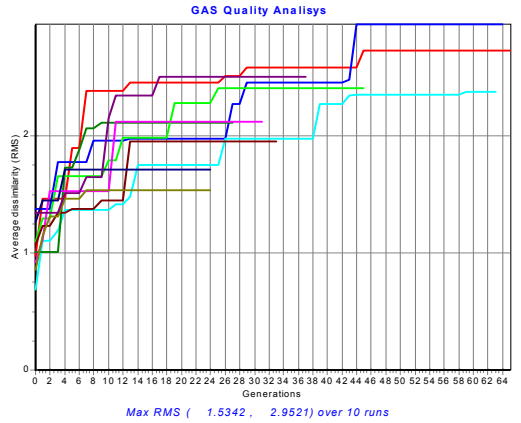
I



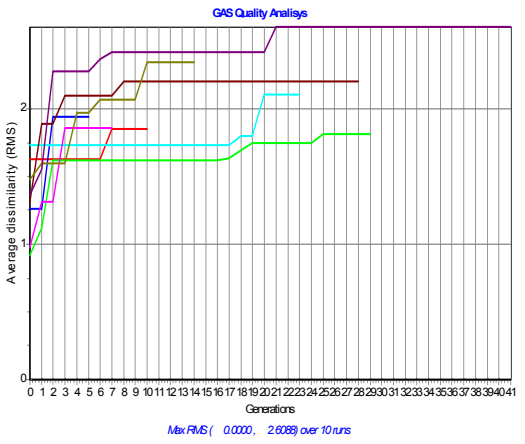
II



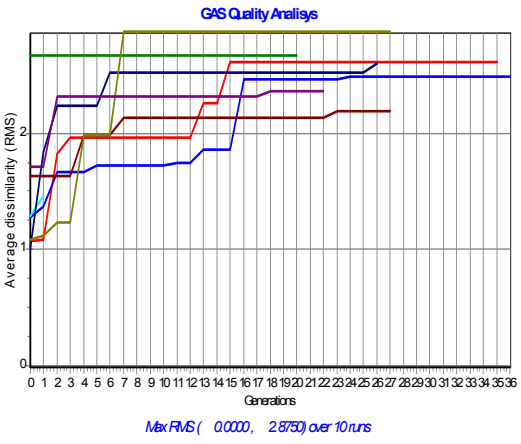
III



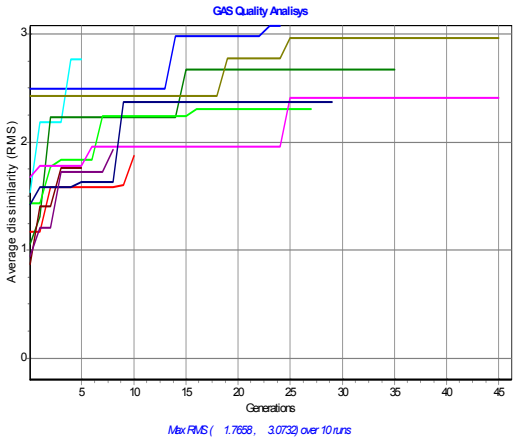
IV



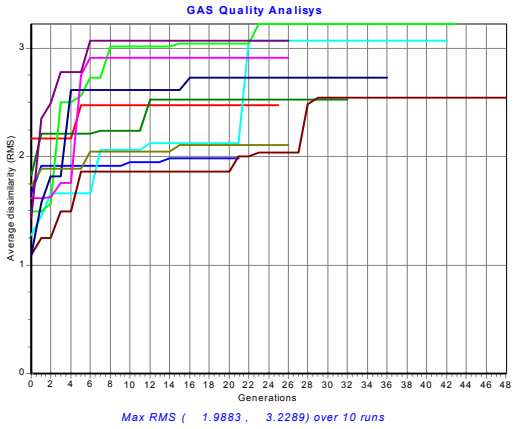
V



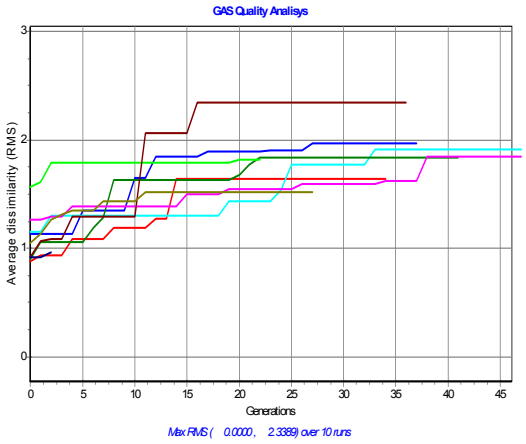
VI


GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (1.7658 , 3.0732) over 10 runs

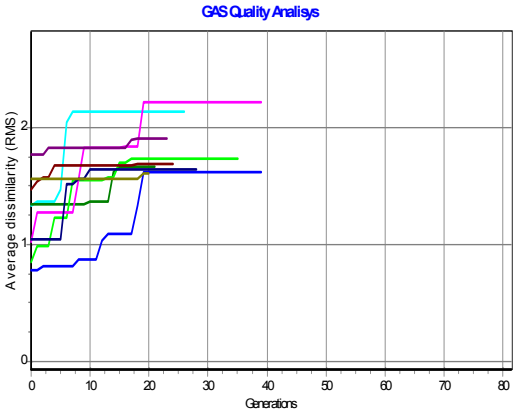
VII


GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (1.9883 , 3.2289) over 10 runs

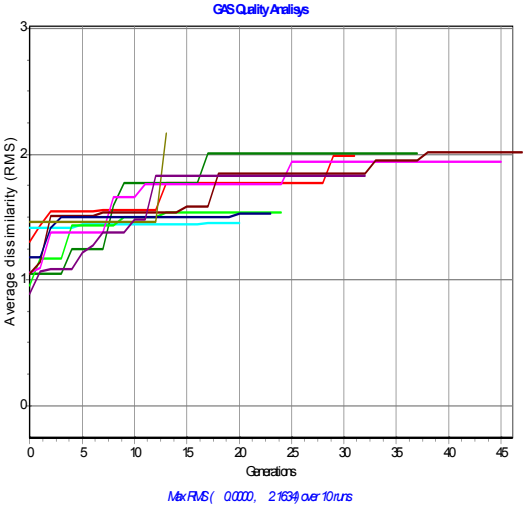
VIII

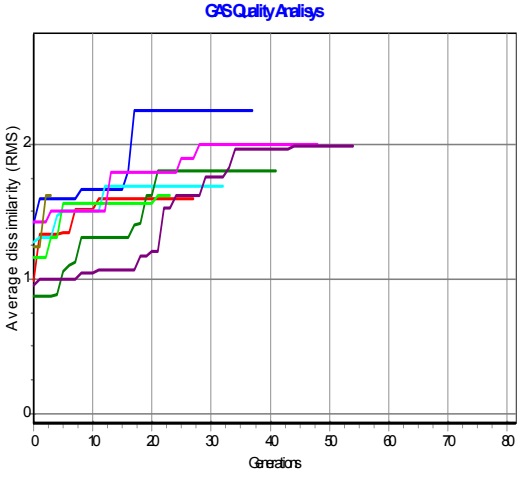

GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (0.0000 , 2.3389) over 10 runs

IX

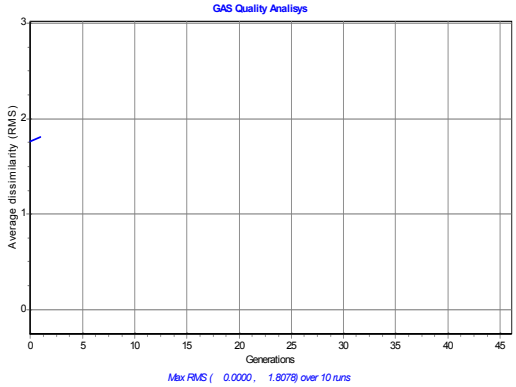

GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (0.0000 , 2.2175) over 10 runs

X

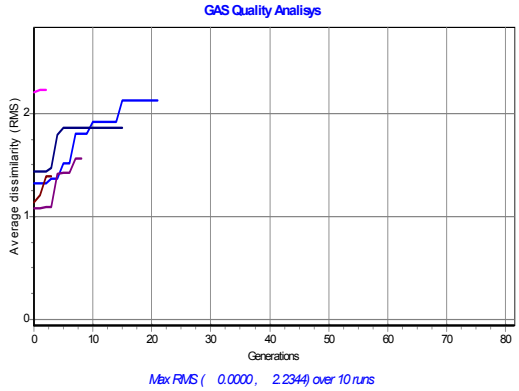

GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (0.0000 , 2.1634) over 10 runs


GAS Quality Analisys
Average dissimilarity (RMS)
Generations
Max RMS (0.0000 , 2.2592) over 10 runs

XI



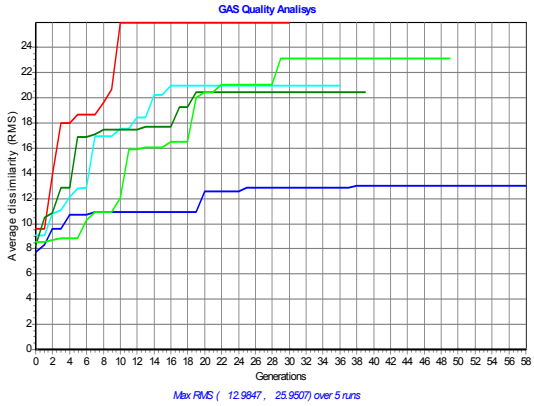
XII



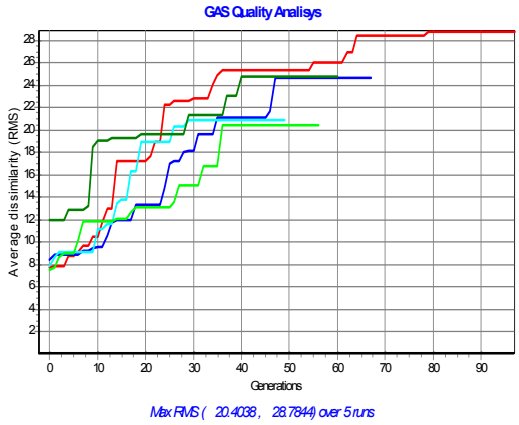
XIII

XIV

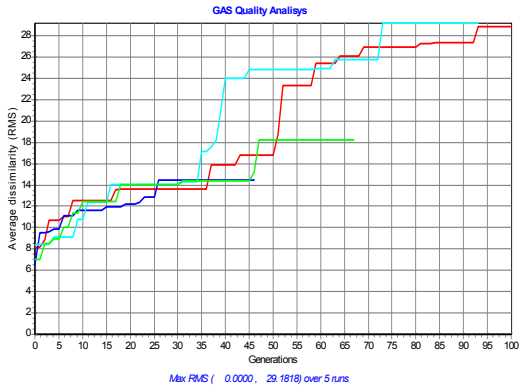
Фигура 3.8



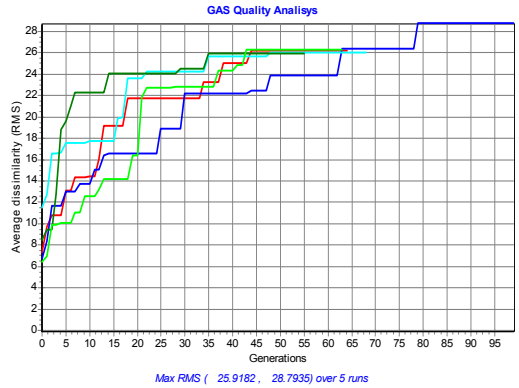
I



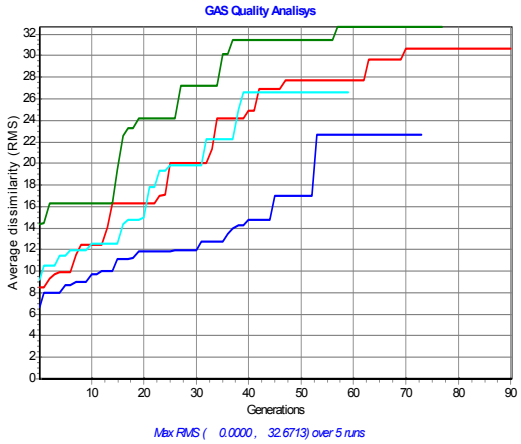
II



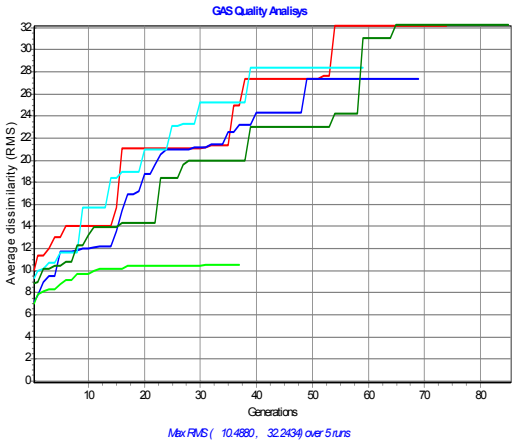
III



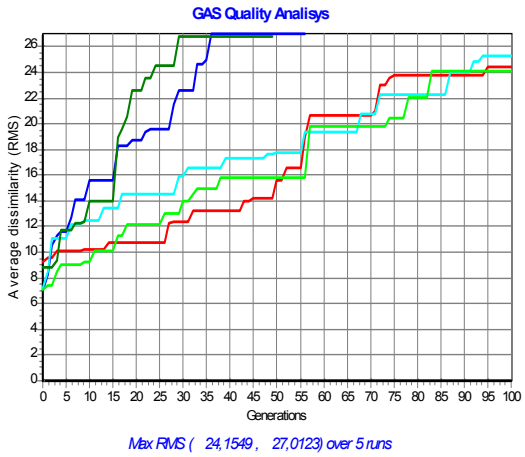
IV



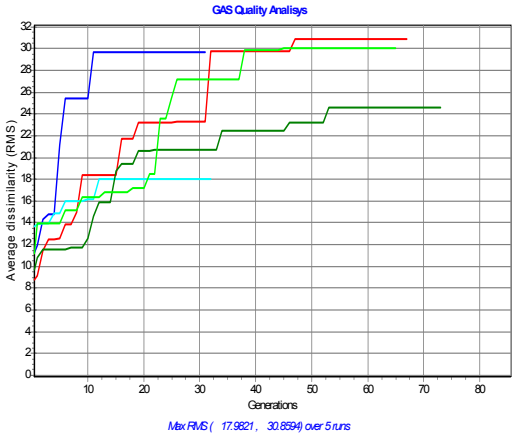
V



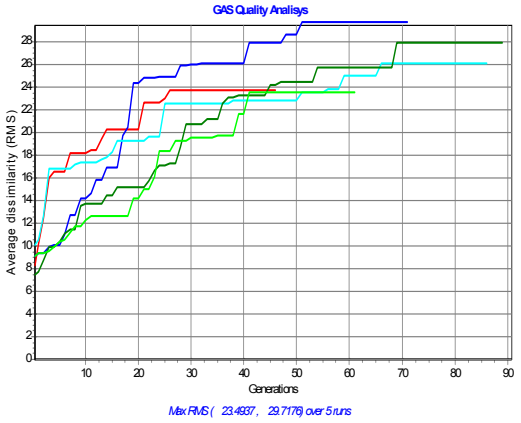
VI



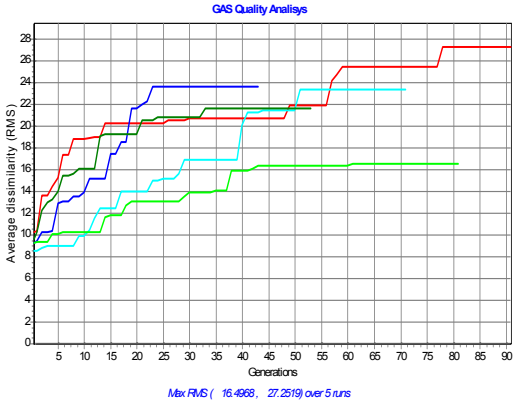
VII



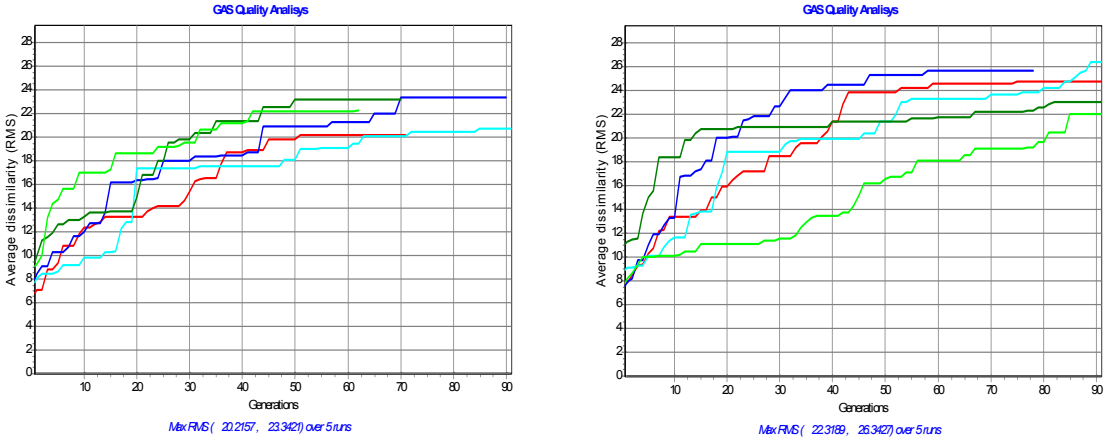
VIII



IX



X



XI XII

Фигура 3.9

Анализ на резултатите

Устойчивостта на метода е оценена в т.нар. “скорост на сходимост”, която е пропорционална на броя на генерациите, необходими за достигане на сходимост на средното различие между конформерите. От Фигура 3.8 и Фигура 3.9 се вижда, че скоростта на сходимост на негъвкавите молекули 3 и 4, значително се различава при различните изпълнения. По тази причина, влиянието на параметрите върху сходимостта е изследвана специално за структури 1 и 2.

Най-общо, тенденцията е за повишаване на устойчивостта с намаляване на отношението N_p/N_c. Поради това, най-ниската скорост на сходимост се наблюдава при експерименти IX и X, където N_p/N_c е съответно равно на 4 и 6, за структури 1 и 2. Скоростта на сходимост е по-добра при $N_p/N_c = 1$. Очевидно, по-голямо N_c увеличава вероятността родителите да бъдат заместени с дъщерни индивиди. По този начин еволюцията е по-динамична и шансовете за бързо нарастване на средното различие са по-големи.

За разлика от други генетични алгортими за генерация на конформери, предлаганият подход използува и максимизира оценъчна функция, която директно влияе върху степента на покритие на конформационното пространство с ограничен брой точки. Интервалът, в който варират стойностите на ценовата функция са

дадени под всяка от Фигура 3.6 до Фигура 3.9. Тесен интервал на окончателните стойности на средното различие, съответствува на по-добра възпроизводимост на генетичната еволюция. Очевидно, интервалът е по-малък при големи N_p /N_c. Например, при експерименти IX и X за структура 1, средното различие има стойности съответно в интервалите [21.8 , 23.7] и [19.2 , 22.7] Å. В действителност, тесни интервали се получават и в други експерименти, но съответните графики на средната разлика по генерации не се получават близки. За структура 2, най-тесният интервал е получен при високо N_p /N_c , равно на 4 и 6 при експерименти IX и X , с интервали на средната разлика съответно [13.6, 15] и [12.4, 13.6]. Аналогично, най-високата възпроизводимост за структури 3 и 4 се получава при високо отношение N_p /N_c (N_p /N_c =4 в експеримент XI и N_p /N_c =3.3 при експеримент X).

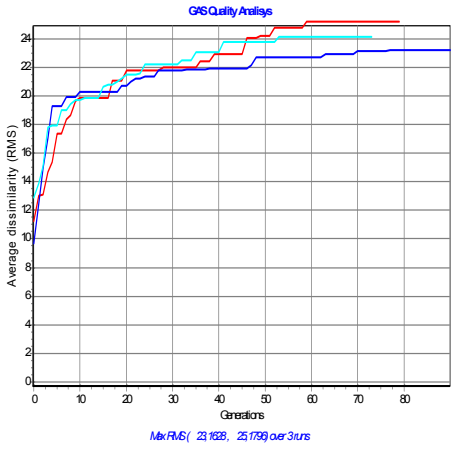
Подобрената възпроизводимост при висока стойност на отношението N_p /N_c може да се обясни с т.нар. по-добър контрол на родителите над децата. По-ниска степен на динамичност на еволюцията води до по-добра възпроизводимост при повтарящите се изпълнения.

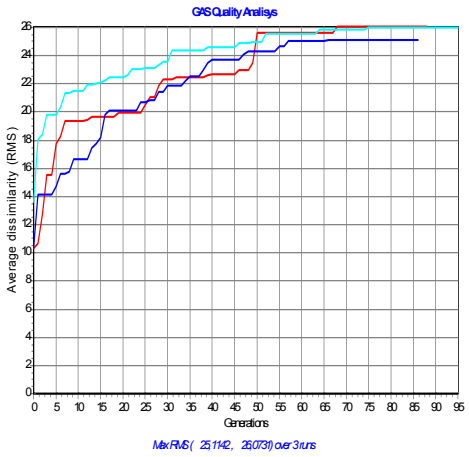
При някои експерименти, еволюцията приключва предварително без достигане на сходимост. Това се дължи на невъзможността да се генерират неизродени структури за първоначалната или последващи популации, при зададени ограничения за потенциалната енергия. Тези ситуации са типични за относително твърдите структури 3 (експерименти V, VI, IX-XII) и 4 (експерименти III и V). Наблюдават се обаче и при гъвкави структури като 2 (експерименти I и III).

Получените резултати показват, че степента на покритие на конформационното пространство намалява с увеличаване на отношението N_p /N_c. Така, най-ниските стойности на средното различие за структури 1-4 се получават при най-високите стойности на отношението N_p /N_c (експерименти IX и X за структура 1, интервали [21.8, 23.7] и [19.2, 22.7] Å, както и експерименти IX и X за структура 2, интервали [13.6, 15.0] и [12.3, 13.6] Å, експерименти IX за структура 3, интервали [0, 2.2] и експерименти XI за структура 4, интервали [20.2, 22.3] Å.

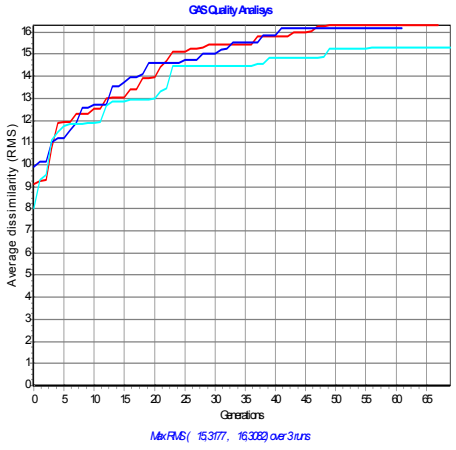
Тези резултати са получени с ограничение за потенциалната енергия на неоптимизираните 3D структури. Обикновено конформерите се подлагат на

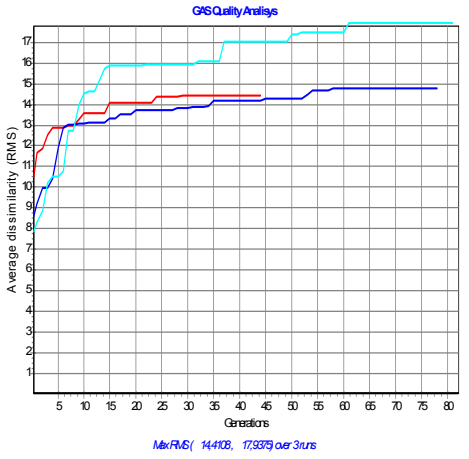
процедура за минимизация на потенциалната енергия на всяка стъпка, преди проверката за надвишаване на прага за енергията. Резултатите, получени при оптимизация със силово поле са илюстрирани на Фигура 3.10a-d, където са показани параметрите са провеждане на експерименти I за структура 1 (Фигура 3.10a), структура 2 (Фигура 3.10b), структура 4 (Фигура 3.10d) и експеримент X за структура 3 (Фигура 3.10c). Първата и втората графика на всяка фигура са от изпълнение на генетичния алгоритъм без и с 3D проверка за израждане. Както е показано на Фигура 3.10a-d, оптимизацията със силово поле значително подобрява възпроизводимостта, но намалява окончателните стойности на средната разлика. Предварителното отхвърляне на дегенерираните конформери води до увеличаване на окончателните стойности на средното различие.



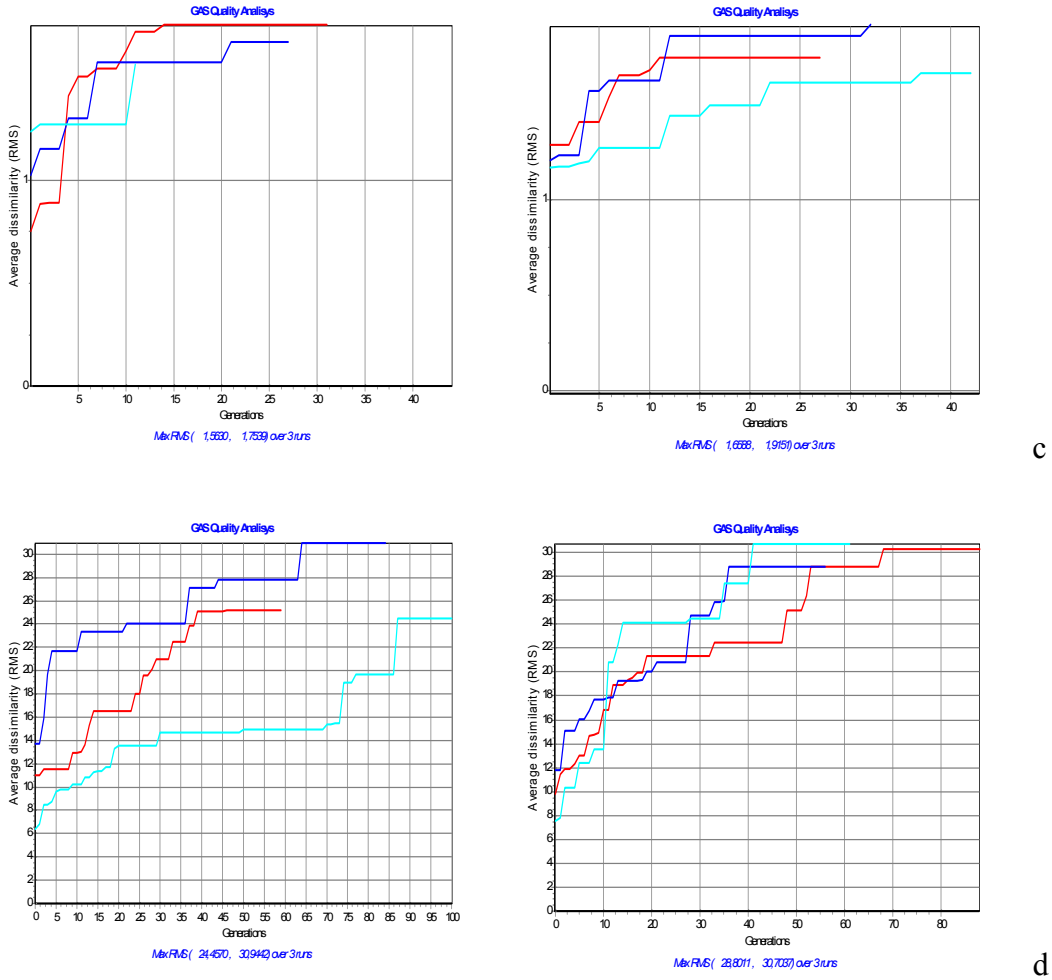


a



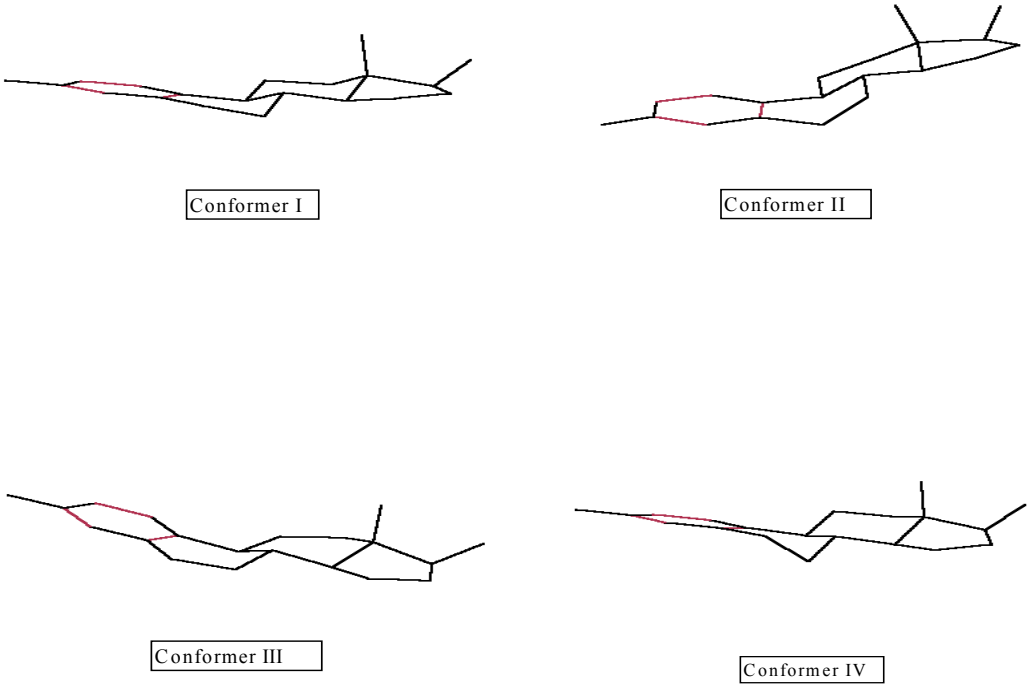


b



Фигура 3.10

При изпълнението на генетичния алгоритъм върху структура 3, в експеримент X с оптимизация със силово поле се получават 4 различни конформера (Фигура 3.11). Сравнението между тези структури и конформерите, получени с изчерпателно търсене [38,62], води до извода, че и при двата подхода се получават близки резултати за покритие на конформационното пространство.



Фигура 3.11

Изводи

Получените резултати показват, че приложението на генетичен алгоритъм е подходящо за решаване на задачата за конформационно покритие. Вместо да се търси множество от (мета)стабилни конформери, се генерират малък брой конформери, които оптимално покриват конформационното пространство и удовлетворяват зададените енергийни ограничения. Използването на генетичен алгоритъм позволява да не се прилага изчерпателно търсене на всички конформери, което е трудоемка задача, особено при големи и гъвкави молекули. Недостатък е недетерминистичният характер на алгоритъма, което води до получаването на различно множество конформери при всяко изпълнение.

3.2. 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 стъпкова (постепенна) модификация 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. За всяка новосъздадена структура се прилага алгоритъма от т. 3.1 за генериране множеството от максимално различни конформери. Критерият за оптимизация при използуване на генетичен алгоритъм може да бъде дефиниран специфично за всяка оптимизационна задача. В предложения метод се използува множество от правила, записани в синтаксиса на разработения формален език (т. 2.2) за описание на структурни, електронни и други ограничения за химични структури.

Основни принципи

Предложеният метод за генериране на лидери има за цел да създава нови структури при зададени структурни, пространствени и електронни ограничения. За целта се използува генетичен алгоритъм за създаване на нови структури чрез сглобяване на фрагменти, и последващо приложение на друг генетичен алгоритъм за намиране на максимално най-различните конформери на новата структура.

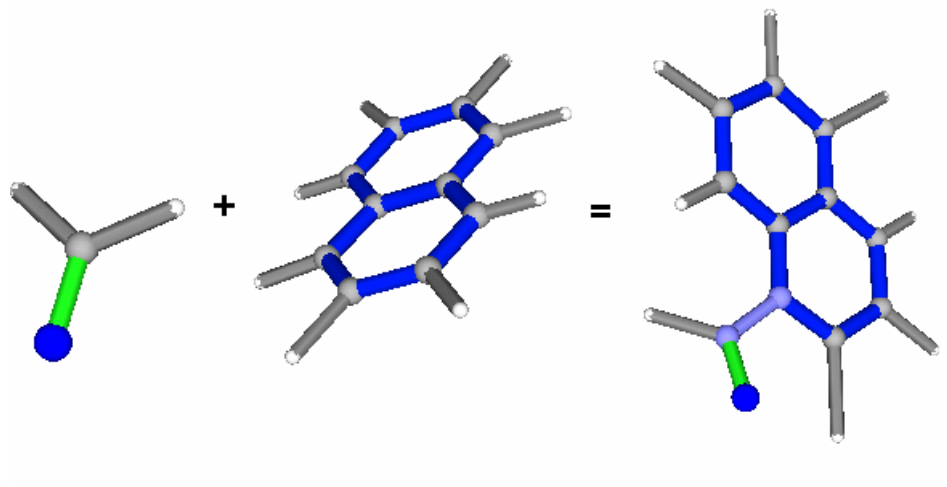
Фрагменти

Необходимо условие за работата на алгоритъма е наличието на набор от фрагменти (*речник*), които се съединяват, за да се получи нова структура. Фрагментите представляват тримерни структури и/или части от структури. Фрагментите трябва да бъдат записани в файл (в специфичен за софтуера формат). Като фрагменти може да се използуват произволен набор структури, освен това има възможност за създаване на файл с фрагменти, получени от дадено множество структури чрез разкъсване на ацикличните им връзки.

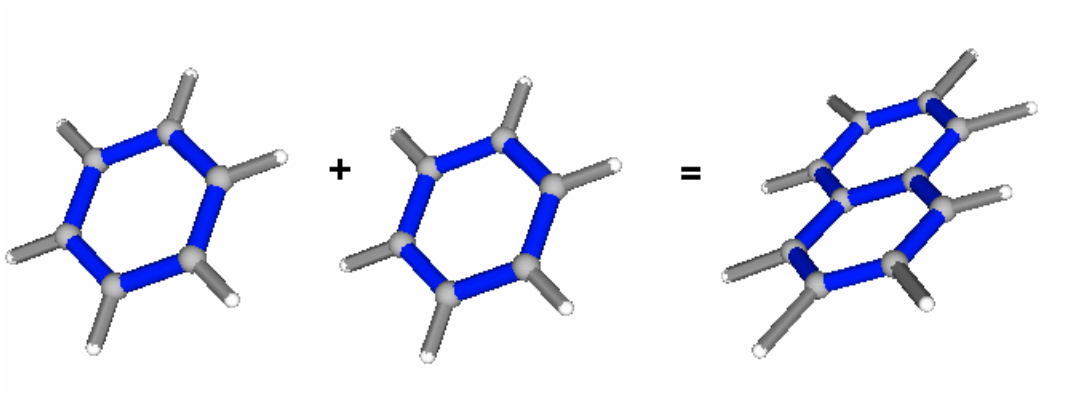
Представяне на индивидите, генетични операции

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). Ще бъдат разгледани генетичните операции, отнасящи се до първия етап на приложение на генетичен алгоритъм, а именно, генерирането на нови молекули чрез комбиниране на фрагменти. Вторият етап е приложение на генетичен алгоритъм за намиране на множество конформери на новата структура е описан в точка 3.1 на дисертационната работа.

Операцията *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 (Фигура 3.12.), 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 между-молекулните 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 (Фигура 3.13.), така че да се получат полициклични съединения. В този случай са въведени правила за спазване на валентностите.



Фигура 3.12.



Фигура 3.13.

Операцията *мутация* представлява замяна на един атом с друг, като предварително е зададена таблица с възможните замени (Таблица 3.5). Мутация се извършва с ниска вероятност (0.05), която може да се променя от потребителя.

Атоми, подлежащи на взаимна замяна	
Csp^3, Si	Въглерод, Силиций
Nsp^3, P(III)	Азот, Фосфор от 3 валентност
-S-, -O-	Сяра, Кислород, свързани с единични връзки
O=, S=	Сяра, Кислород, свързани с двойни връзки
N#, P#	Азот, Фосфор , свързани с тройни връзки
Csp^2, S(IV)	Въглерод, Сяра от четвърта валентност
F, Cl, Br, I	Флуор, Хлор, Бром, Йод

Таблица 3.5

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, които ще бъдат описани по-долу.

Описание на алгоритъма

На Фигура 3.14. е показана обобщена блок-схема на предложения алгоритъм за генериране на нови съединения при зададени ограничения. В съответствие с принципите на генетичните алгоритми, последователно се генерира множество от популации, всяка от които трябва да подобрява предишната.

Генериране на първоначалната популация

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. Свързването става по същия начин, както е описано за втория етап на операцията кръстосване. Новата структура може да бъде добавена в популацията или отхвърлена, в зависимост от нейната оценка. Първоначалната популация се счита запълнена, ако са генерирани предварително зададен брой молекули (размер на популацията N_p).

Генериране на следваща популация

Новите структури се генерират с описаните генетични операции *кръстосване* и *мутация*. Всяка нова структура се подлага на:

- Проверка за минимална и максимална молекулна маса;
- Оптимизация със силово поле [38] за осигуряване стабилност на молекулата. Ако след оптимизацията структурата има енергия над зададен праг, то тя се отхвърля;
- Генериране на конформери на новата структура. Едно от нововъведенията в метода за генериране на лидери е приложението на ГА на две нива. Докато на първо ниво се генерират структури, които удовлетворяват химическите

изисквания, то на второ ниво друг генетичен алгоритъм (т.3.1) се използува за търсене на конформери на тези структури.

- Оценка на конформерите на новата структура, при което на всяка структура се присвоява ранг. Ако рангът е нулев, структурата се отхвърля.

При генериране на зададения брой нови структури N_c всички родителски и дъщерни обекти се подреждат по намаляващ ранг, и следващата популация се образува от първите N_p структури, т.е. тези с най-добър ранг. Поради това, рангът на структурите във всяка популация не намалява и с всяка еволюция се намират по-добри структури.

Оценка на новите структури

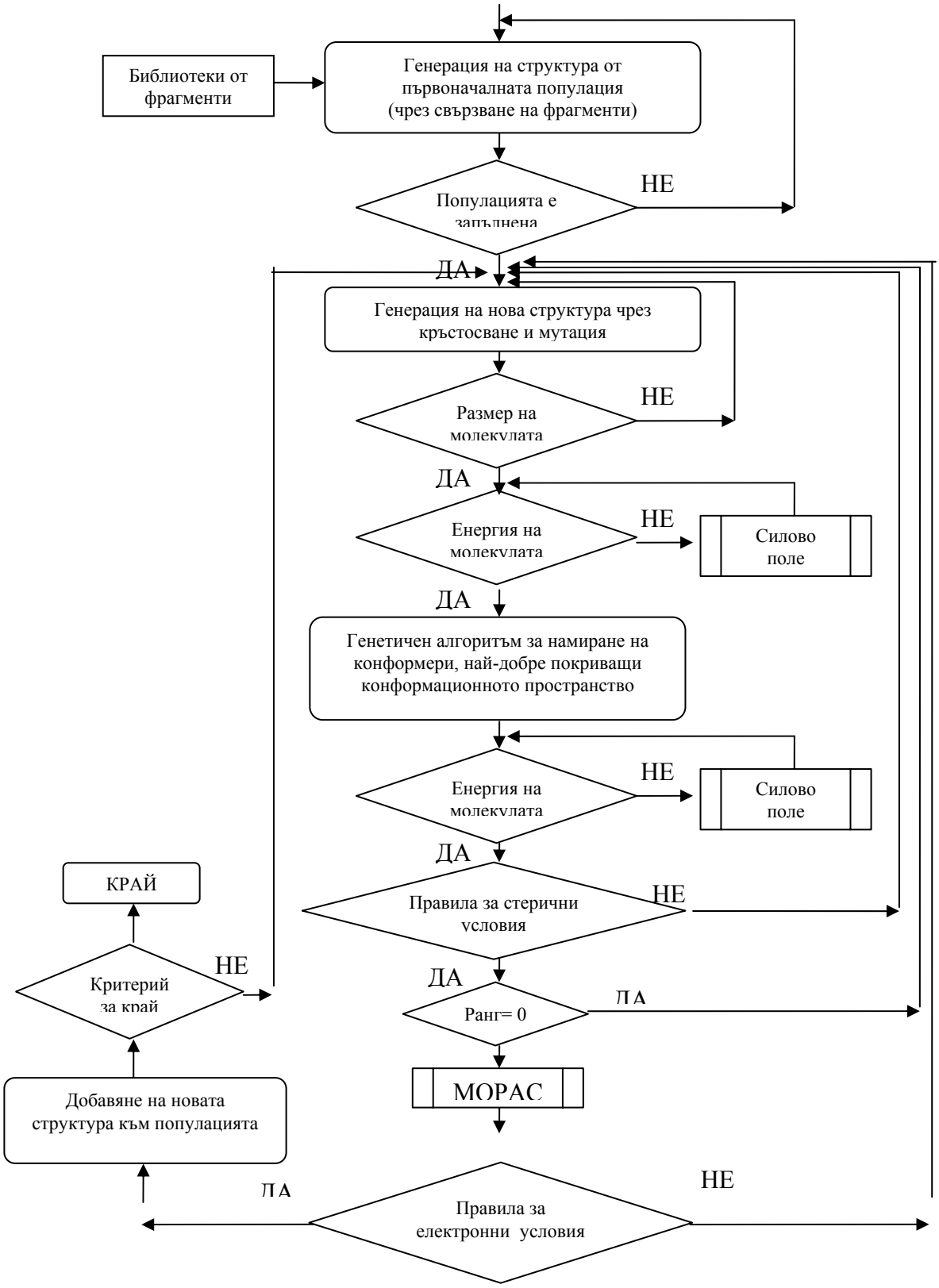
Към всеки конформер на новата структура се прилага оценъчна функция, която се формира като претеглена сума на геометрични, енергетични и електронни компоненти. Геометричният компонент оценява доколко структурата съвпада с фармакофора. Енергетичният компонент осигурява стабилност на генерираните структури. Обикновено той се основава на Ван дер Ваалсова енергия, която позволява конформерите със стерични конфликти да бъдат елиминирани. Енергията може да се пресмята при минимизация с използуване на силово поле. Електронният компонент отразява електронното сходство на структурата и фармакофора. Структурите се подлагат на квантово-химични изчисления, ако има зададени електронни условия. Всеки компонент се задава чрез разработения формален език (т.2.2) като набор от условия, и в зависимост от изпълнението им, на конформера се присвоява определен ранг. Ако рангът е нулев (или по-малък от зададена от потребителя стойност), то конформерът се отхвърля. Оценъчната функция се получава като сума от ранговете на всички компоненти. Ако оценката на конформера е над зададен от потребителя праг, то той се добавя към новата популация. Възможно е да се използува и модификация на алгоритъма, при който се запазва само конформера с най-висок ранг.

Предложеният метод позволява да се наложат множество различни ограничения върху търсените структури. Тъй като рангът на структурите във всяка следваща

популация не намалява, то алгоритъмът представлява метод за многоцелева оптимизация.

Критерий за край

- невъзможност за намиране на нови структури с по-добър ранг от родителските
- надвишаване на зададения брой итерации


Библиотеки от фрагменти
Генерация на структура от първоначалната популация (чрез свързване на фрагменти)
Популацията е запълнена
НЕ
ДА
Генерация на нова структура чрез кръстосване и мутация
Размер на молекулата
НЕ
ДА
Енергия на молекулата
НЕ
Силово поле
ДА
Генетичен алгоритъм за намиране на конформери, най-добре покриващи конформационното пространство
Енергия на молекулата
НЕ
Силово поле
ДА
Правила за стерични условия
НЕ
ДА
Ранг= 0
ДА
МОРАС
Правила за електронни условия
НЕ
ДА
Добавяне на новата структура към популацията
Критерий за край
НЕ
КРАЙ

Фигура 3.14.

3.3. Изводи

При анализ на гъвкави молекули химичните съединения се представят с множество от техни конформери. Разработен е генетичен алгоритъм за намиране на множество от максимално различни конформери на дадено съединение. 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. В допълнение към разработения алгоритъм се използува и насочен *tweak* метод [37], чрез който се търсят конформери, удовлетворяващи зададени ограничения за разстояния между атоми и/или фрагменти.

За оценка на различието между конформерите се използува RMS разстояние, което е строго дефинирано чрез декартовите координати на две структури. Дефиницията за RMS разстояние е разширена и е въведена мярка за средно различие, за да може да се оцени различието между повече от две структури.

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.

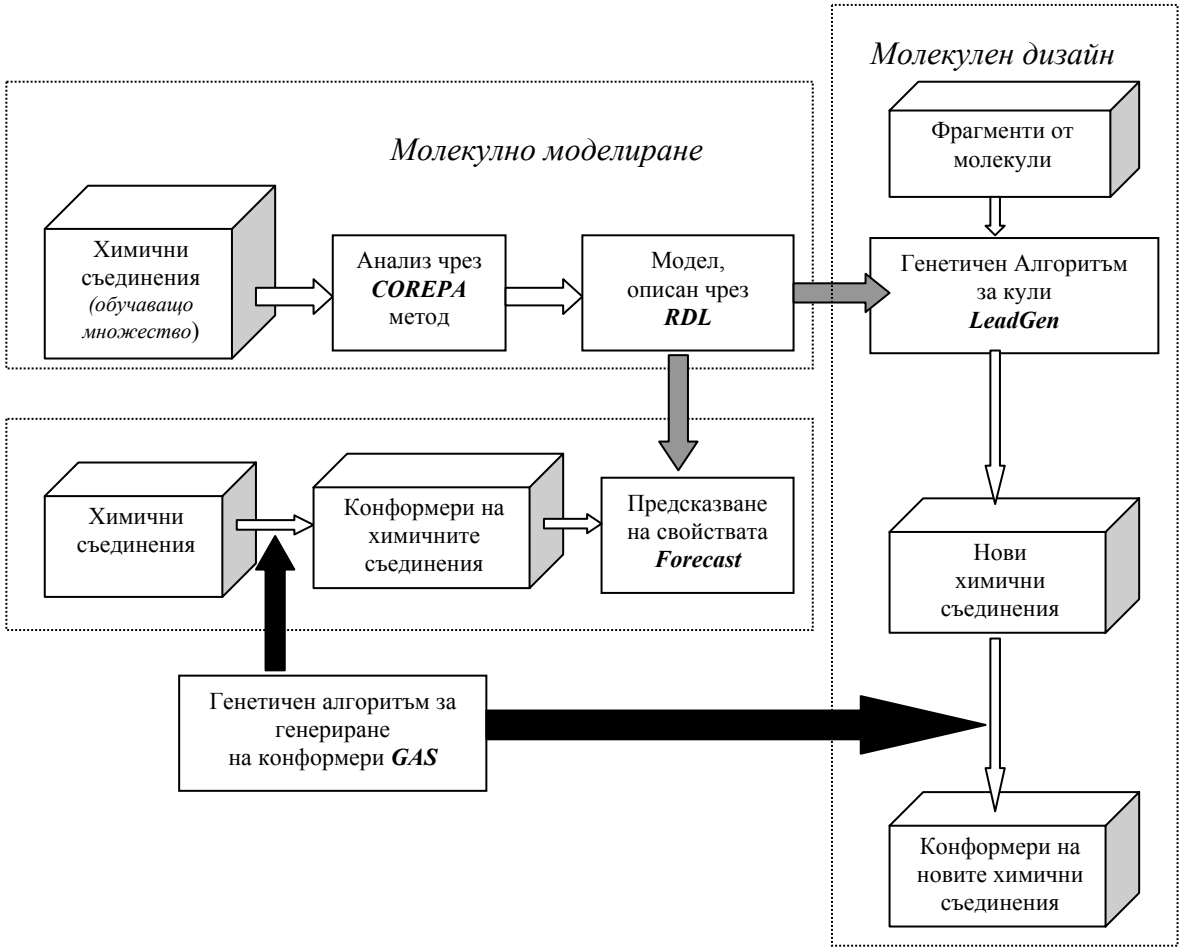
Генерирането на конформери се осъществява чрез ротация около ациклични връзки, обръщане на свободни ъгли и пирамиди. Последните две операции позволяват да се модифицират и цикличните фрагменти, които в голямата част от известните методи и софтуер се разглеждат като твърди и не се модифицират. Методът позволява също така запазване на стереохимичните ограничения върху структурата, или игнорирането им по желание на потребителя.

Направено е изследване на работата на алгоритъма, чрез прилагането му върху 4 различни по гъвкавост структури, като е оценено влиянието на управляващите параметри върху устойчивостта и възпроизводимостта на генетичния алгоритъм и върху получените множества конформери. Устойчивостта се оценява чрез скоростта на сходимост, при която се достига максимално различие между множеството конформери. Възпроизводимостта е оценена чрез границите на изменение на средното различие при няколко повторения на изпълнението на

алгоритъма. При увеличаване на отношението N_p/N_c, възпроизводимостта нараства, докато устойчивостта и степента на покритие, измерена чрез средното различие намалява. Тази тенденция се наблюдава при всички тестови структури.

Прилагането на оптимизация със силово поле води до подобряване на възпроизводимостта, при което се получава множество конформери с по-малко средно различие.

Предложеният метод за генериране на конформери понастоящем е вграден във всички програмни приложения, разработени в Лабораторията по математична химия при Университет "Проф. А.Златаров" - Бургас.



Фигура 3.15.

За получаване на нови съединения със зададени свойства е разработен генетичен алгоритъм на две нива. Желаните свойства се задават чрез правила, записани посредством езика RDL. Това позволява задаването на разнообразни свойства и експертни условия. Предимство на алгоритъма е че дава възможност и за отчитане гъвкавостта на съединенията в хода на генерацията, което го отличава от известните алгоритми.

Разработените методи са взаимно свързани и предлагат цялостен подход към решаването на задачи в молекулното моделиране (построяване на модел на изследваното свойство) и последващото му използуване за предсказване и генериране на нови обекти Фигура 3.15.)

4. Приложение на разработените методи за моделиране и изследване на химични структури

4.1. Програмни системи

Описаните в глави **Error! Reference source not found.** 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. Програмните системи позволяват:

COREPA - реализация на COREPA метода за построяване на модел на биологичната активност чрез дърво на решението, описан в т. **Error! Reference source not found.**, с многобройни възможности за визуализация на резултатите;
FORECAST - програмна система, позволяваща количествено и качествено прогнозиране на биологични свойства при готови файлове с правила на езика RDL (извлечени чрез COREPA);
LEADGEN - програмна система, реализираща описания в т. 3.2 алгоритъм за генериране на нови съединения с предварително зададени свойства
GAS32 - програмна система, реализираща описания в т.3.1 алгоритъм за намиране на множеството от максимално различни конформери. Този алгоритъм е интегриран и във всички останали приложни програми, с цел да може прозрачно да се генерират конформери при решение на съответната задача. Освен това GAS32 дава възможност и за изследване на възпроизводимостта и стабилността на алгоритъма при многократно изпълнение върху едни и същи структури и при различни настройки.
Всички програмни системи използуват общи модули за визуализация на химичните бази данни, химичните структури (двумерни и тримерни), изчисляване на параметри, въвеждане на нови структури.

4.1.1. Файлове, символен и графичен вход, изчисляване на свойства

Химичните съединения в разработените програмни системи се съхраняват в специфичен файлов формат (*.cmp*). В настоящите версии не се използува система за 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.

За всяка структура се съхранява следната информация:

SMILES
Химично наименование
CAS No.
Свързаност (брой атоми, типове атоми, брой връзки, типове връзки между атомите)
Тримерни координати на атомите
Числови и стрингови параметри, отнасящи се за цялата молекула (глобални параметри)
Числови параметри, отнасящи до всеки атом или връзка (локални параметри)

Имената на глобалните и локални параметри се поддържат в отделен общ текстов файл. Всяка структура може да съдържа стойности за произволно подмножество от изброените параметри.

CMP файловете се създават от приложния софтуер, като химичната информация може да се въведе по следния начин:

- Свързаност, зададена в текстов файл, съдържащ SMILES. На Фигура 4.1 е показан пример за много структури, зададени с SMILES означения.
- Свързаност, 3D информация и глобални параметри, зададен в текстов файл в SDF/MOL/MOL2/XYZ формат. На Фигура 4.2 е показан пример за една структура, зададена в SDF формат.

```
O=P(OCC(CCCC)CC)(OCC(CCCC)CC)OCC(CCCC)CC
O=C(N)CCl
O=C(OO)C
Oc(c(cc(c1)C(c(cc(c(O)c2Br)Br)c2)(C)C)Br)c1Br
O=S(=O)(c(ccc(c1)Cl)c1)c(ccc(c2)Cl)c2
Cl[Si](c1ccccc1)(c2ccccc2)Cl
```

Фигура 4.1

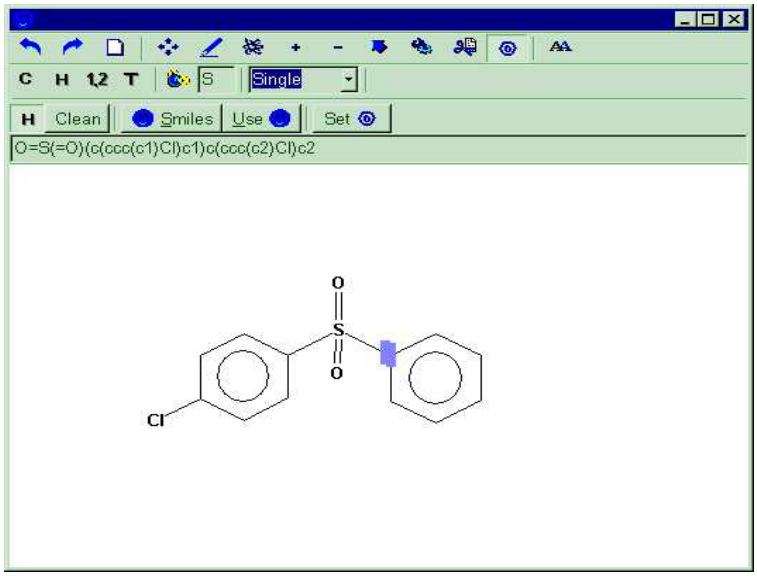
- Свързаност (SMILES) и параметри, зададени в EXCEL таблица. Тази възможност се използува обикновено при наличие на данни за голям брой съединения в текстов формат.
- Свързаност, зададена чрез двумерния редактор, реализиран в програмната система.

```
0174177
 -ISIS-  04180016452D

 25 27  0  0  0  0  0  0  0  0  0999 V2000
    2.9700   -1.3830    0.0000 C   0  0  0  0  0
    2.9700   -2.0760    0.0000 O   0  0  0  0  0
    0.9870    0.4380    0.0000 O   0  0  0  0  0
    2.3580   -1.0440    0.0000 N   0  0  0  0  0
    0.0180    0.3960    0.0000 C   0  0  0  0  0
    1.8150   -1.3650    0.0000 C   0  0  0  0  0
-2.5050    0.6900    0.0000 C   0  0  0  0  0
   -2.3250    1.2660    0.0000 C   0  0  0  0  0
   -3.2760    1.7490    0.0000 N   0  0  0  0  0
   -2.6790    1.7490    0.0000 C   0  0  0  0  0

 10 13  1  0  0  0
 16 10  1  0  0  0
 15 16  1  0  0  0
 14 15  1  0  0  0
 13 14  1  0  0  0
19 21  4  0  0  0
 23 22  4  0  0  0
 23 21  4  0  0  0
 23 25  1  0  0  0
 25 24  3  0  0  0
M  END
>  <UN>
0174177
>  <BIOSPV4.MIC>
1
>  <BIOSPV15.MIC>
.125
$$$$
```

Фигура 4.2

На Фигура 4.3 е показан интерфейсът на двумерен редактор за структурни диаграми на химични съединения. Двумерният редактор позволява химичното съединение да се чертае по познатия на химика начин за рисуване на структурна диаграма. Може да се работи с няколко типа връзки (прости, двойни, тройни, ароматни), да се сменят типовете атоми с произволен атом от периодичната таблица, както и да се използуват предефинирани фрагменти за чертаене на сложни структури. Осигурява се контрол на валентностите, като с червено се показват атомите, при които има нарушаване на валентността.Има възможност за въвеждане

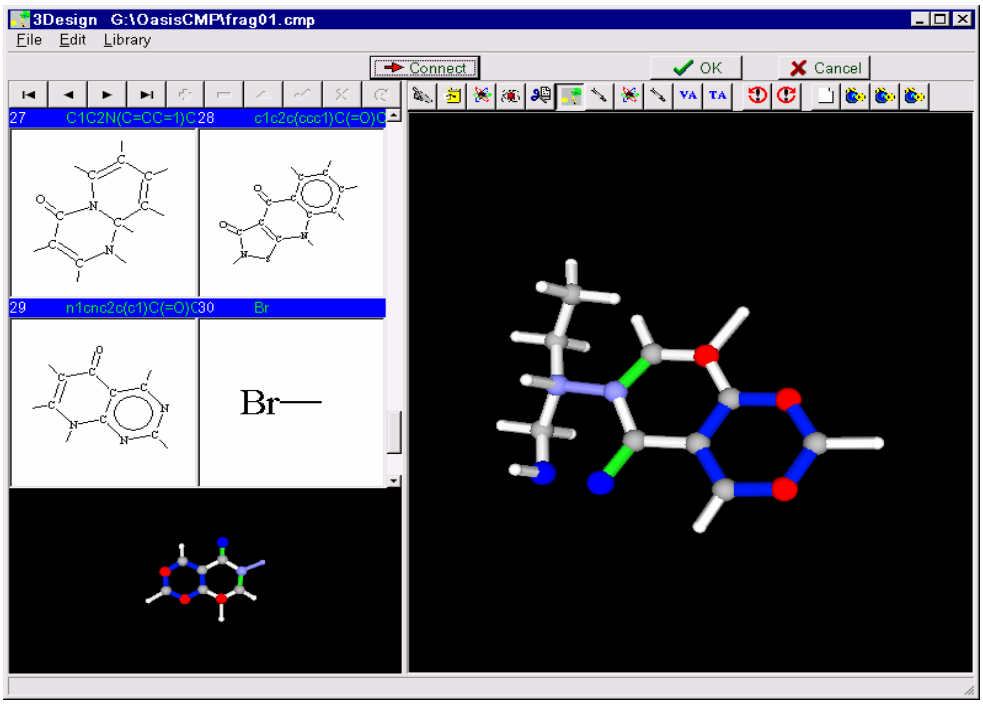
на стереохимични спецификации. От чертежа може да се генерира стандартно SMILES означение, както и от SMILES означение може да се генерира чертеж.



Фигура 4.3

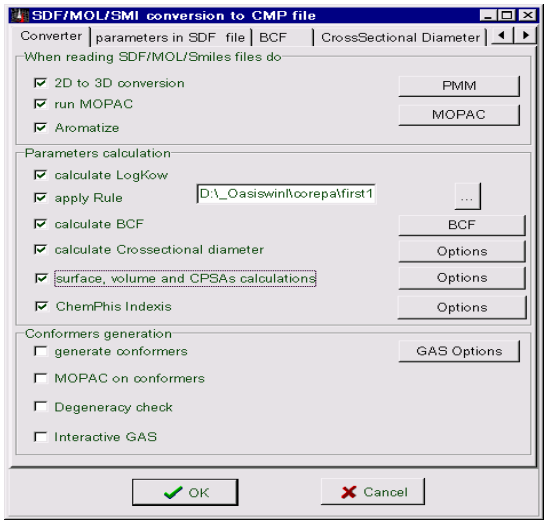
- Свързаност и тримерни координати, зададени чрез тримерния редактор, реализиран в софтуера. На Фигура 4.4. е показан интерфейсът на тримерен редактор за конструиране на химични съединения.

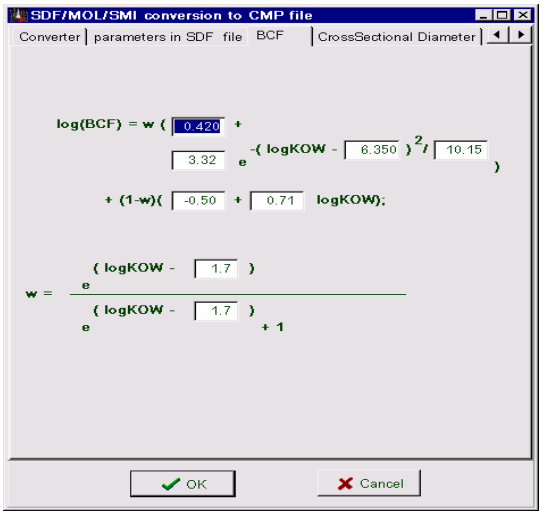
За визуализация на тримерните модели се използува стандартната OpenGL библиотека. Тримерният редактор позволява произволен *.cmp* файл да се зареди като библиотека с фрагменти и да се използуват тези фрагменти за конструиране на нова структура. Фрагментите могат да се свързват по ациклични връзки (connect button) или да се слепват по цикличните връзки (glue button). Дава се възможност за промяна на тип на атом, тип на връзка, валентни и торзионни ъгли, изтриване на атом и връзка, оптимизация на структурата, както и undo/redo операции.



Фигура 4.4.

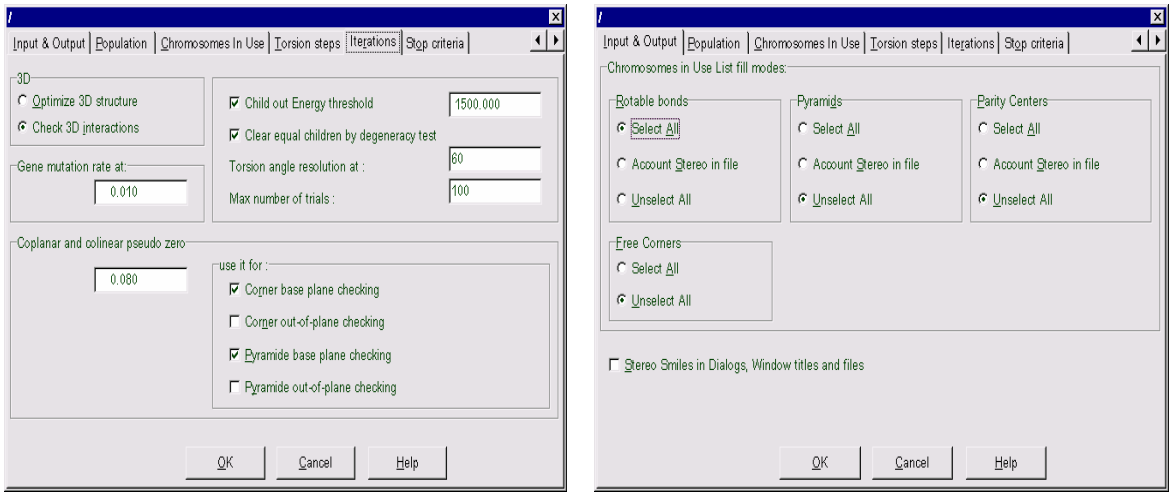
- Изчисляване на различни параметри на вече въведени химични съединения (физикохимични свойства, геометрични и електронни индекси, енергия, ефективен диаметър, обем и повърхност на молекулата и други). На Фигура 4.5. са показани част от настройките за изчисляване на различни параметри на химични съединения.





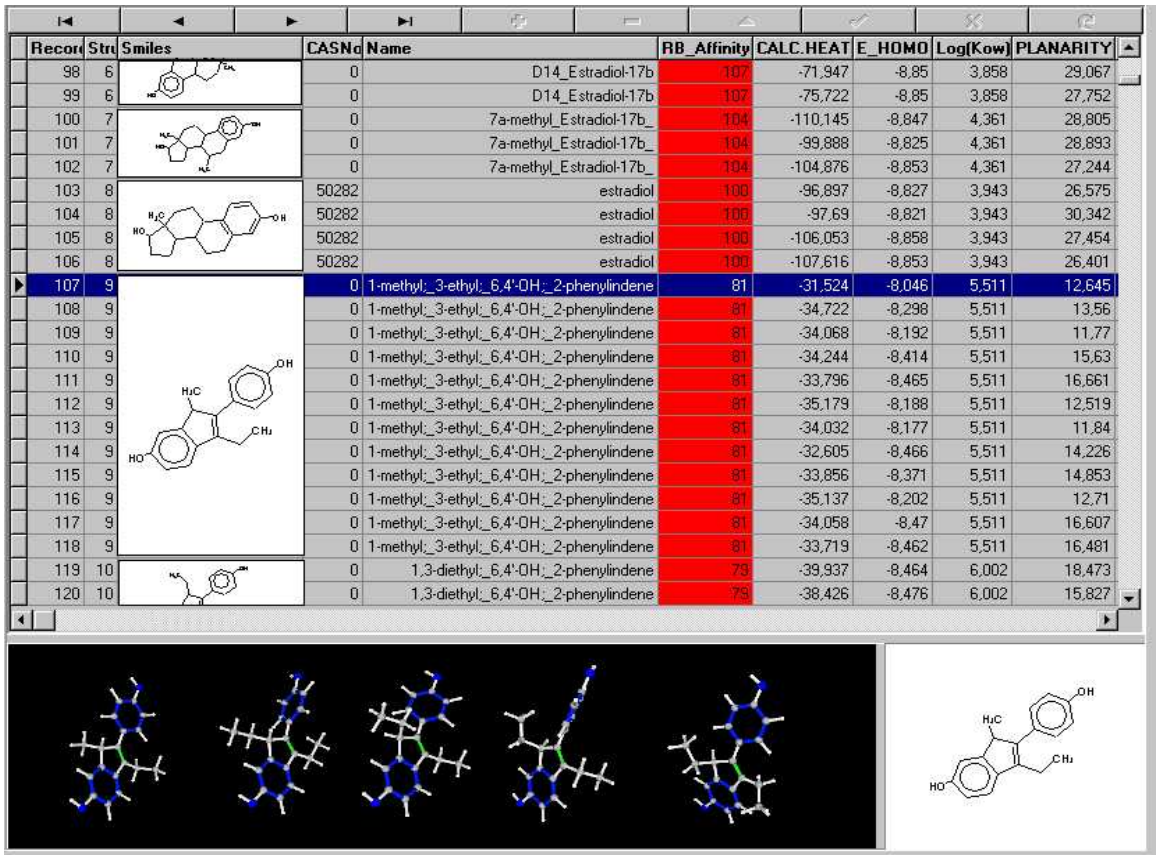
Фигура 4.5.

- Генериране на конформери на въведените структури с генетичния алгоритъм, описан в т.3.1. На Фигура 4.6. е показан интерфейсът за част от настройките на генетичния алгоритъм за намиране на множеството от максимално различни конформери.



Фигура 4.6.

Тъй като *.cmp* файловете са в специфичен формат и не се използува стандартна база данни, то не могат директно да се използуват достъпните в Borland Delphi компоненти за визуализиране на данните в табличен вид. С тази цел е разработен компонент (наследник на TdataSet), който осигурява подходящ интерфейс между данните от *.cmp* файла и стандартните Database визуални компоненти. За улесняване на четенето на произволен запис от *.cmp* файла (тези файлове са съставени от записи с различна дължина, по един запис за всяко химично съединение) е разработено индексиране на файловете. На Фигура 4.7. е показан интерфейсът за визуализиране на база данни от химични съединения.



Record	Str	Smiles	CASNo	Name	RB_Affinity	CALC.HEAT	E_HOMO	Log(Kow)	PLANARITY
98	6		0	D14_Estradiol-17b	107	-71,947	-8,85	3,858	29,067
99	6		0	D14_Estradiol-17b	107	-75,722	-8,85	3,858	27,752
100	7		0	7a-methyl_Estradiol-17b_	104	-110,145	-8,847	4,361	28,805
101	7		0	7a-methyl_Estradiol-17b_	104	-99,888	-8,825	4,361	28,893
102	7		0	7a-methyl_Estradiol-17b_	104	-104,876	-8,853	4,361	27,244
103	8		50282	estradiol	100	-96,897	-8,827	3,943	26,575
104	8		50282	estradiol	100	-97,69	-8,821	3,943	30,342
105	8		50282	estradiol	100	-106,053	-8,858	3,943	27,454
106	8		50282	estradiol	100	-107,616	-8,853	3,943	26,401
107	9		0	1-methyl;_3-ethyl;_6,4'-OH;_2-phenylindene	81	-31,524	-8,046	5,511	12,645
108	9		0	1-methyl;_3-ethyl;_6,4'-OH;_2-phenylindene	81	-34,722	-8,298	5,511	13,56
109	9		0	1-methyl;_3-ethyl;_6,4'-OH;_2-phenylindene	81	-34,068	-8,192	5,511	11,77
110	9		0	1-methyl;_3-ethyl;_6,4'-OH;_2-phenylindene	81	-34,244	-8,414	5,511	15,63
111	9		0	1-methyl;_3-ethyl;_6,4'-OH;_2-phenylindene	81	-33,796	-8,465	5,511	16,661
112	9		0	1-methyl;_3-ethyl;_6,4'-OH;_2-phenylindene	81	-35,179	-8,188	5,511	12,519
113	9		0	1-methyl;_3-ethyl;_6,4'-OH;_2-phenylindene	81	-34,032	-8,177	5,511	11,84
114	9		0	1-methyl;_3-ethyl;_6,4'-OH;_2-phenylindene	81	-32,605	-8,466	5,511	14,226
115	9		0	1-methyl;_3-ethyl;_6,4'-OH;_2-phenylindene	81	-33,856	-8,371	5,511	14,853
116	9		0	1-methyl;_3-ethyl;_6,4'-OH;_2-phenylindene	81	-35,137	-8,202	5,511	12,71
117	9		0	1-methyl;_3-ethyl;_6,4'-OH;_2-phenylindene	81	-34,058	-8,47	5,511	16,607
118	9		0	1-methyl;_3-ethyl;_6,4'-OH;_2-phenylindene	81	-33,719	-8,462	5,511	16,481
119	10		0	1,3-diethyl;_6,4'-OH;_2-phenylindene	79	-39,937	-8,464	6,002	18,473
120	10		0	1,3-diethyl;_6,4'-OH;_2-phenylindene	79	-38,426	-8,476	6,002	15,827

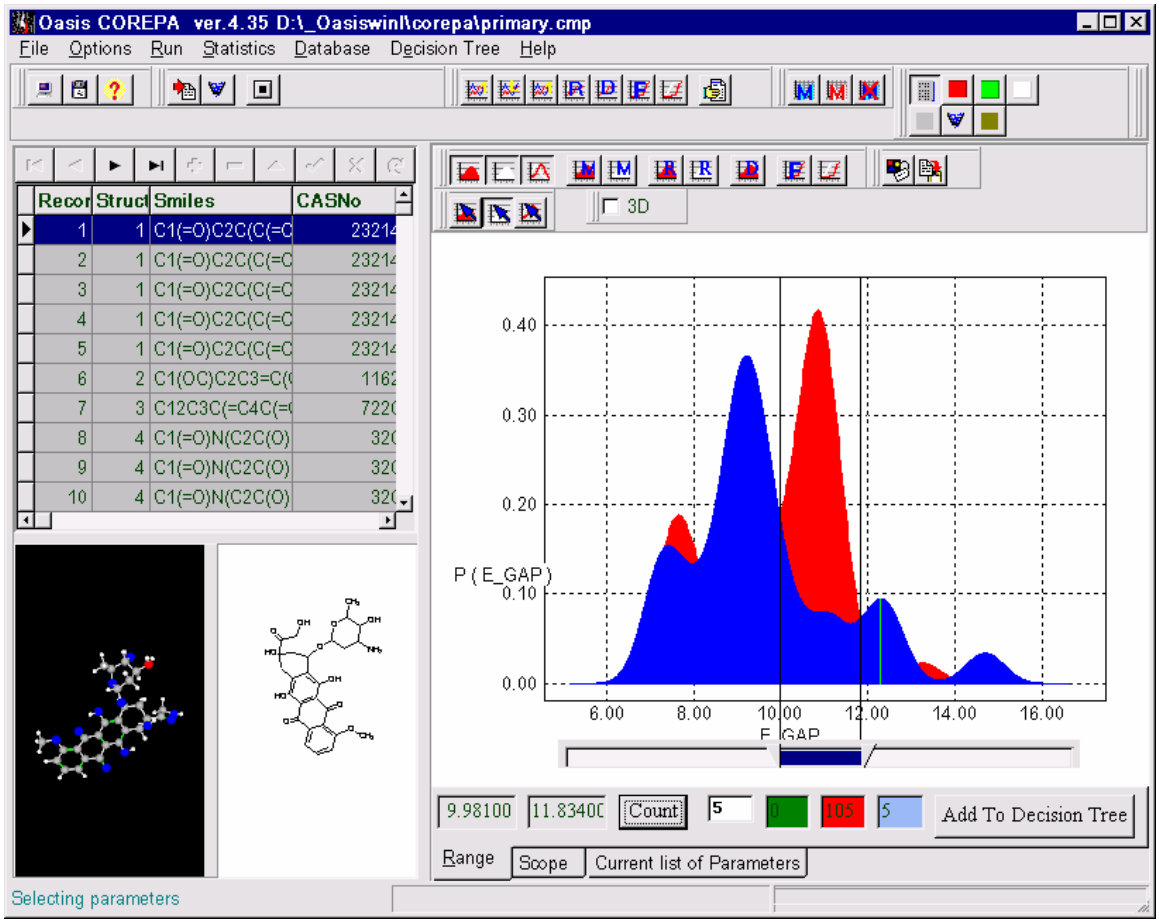
Фигура 4.7.

4.1.2. Програмна система Common Reactivity PAttern

COREPA е приложна програма, която реализира метода COREPA, описан в глава **Error! Reference source not found.**. Освен това системата позволява:

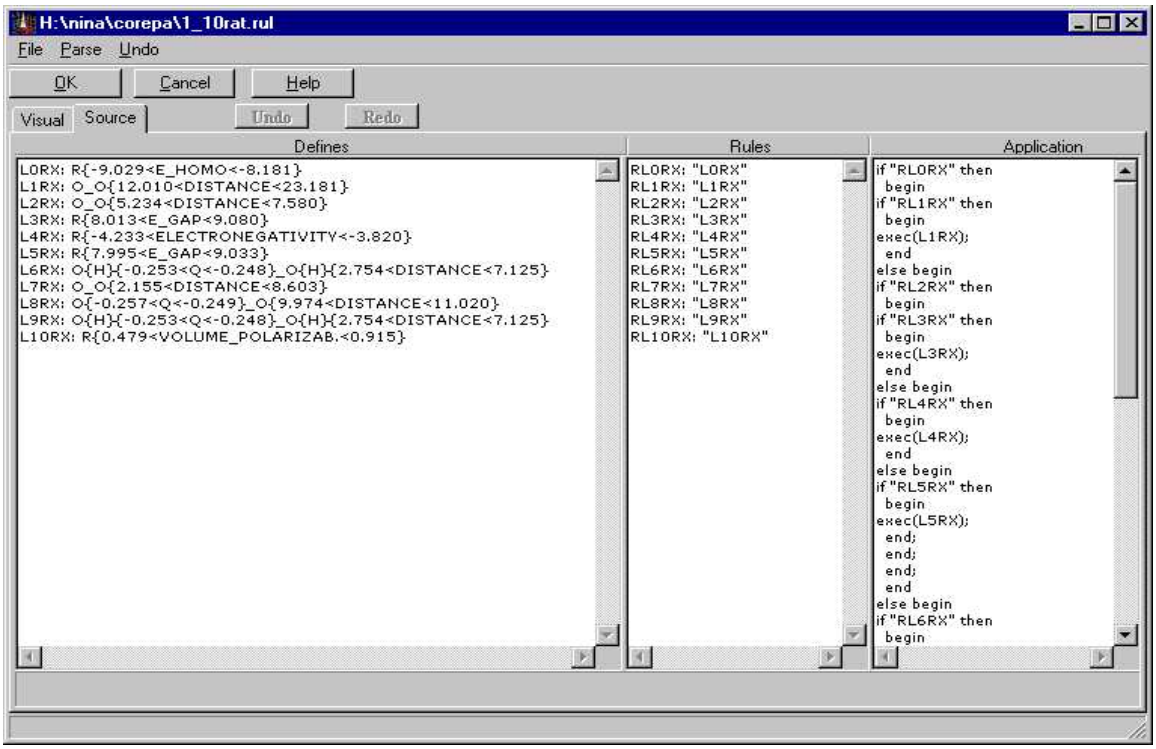
- Използуване на предварително подготвени данни за химични структури, като обучаващи данни за извличане на общия реакционен образ, използувайки метода, описан в т. 2.1;
- Визуализиране на молекулните и атомни дескриптори, на които се дължи реакционния образ;
- Оценка на изведения модел според различни критерии;
- Определяне на общите интервали за конформерите на всички съединения;
- Визуализиране на двумерни и тримерни модели на структурите;

- 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.



Фигура 4.8.

На Фигура 4.8. е показан главният екран на програмата COREPA. Химичните съединения в базата данни с експериментални резултати за биологична активност (в случая естрогенна активност, RBA) се разделят на две групи - активни и неактивни. Избира се множество от молекулни параметри (дескриптори), които ще бъдат оценявани по COREPA метода. Дървото на решението се построява, както е описано в т. 2.1. и може да се запише в текстов файл с правила , както е показано на Фигура 4.9.



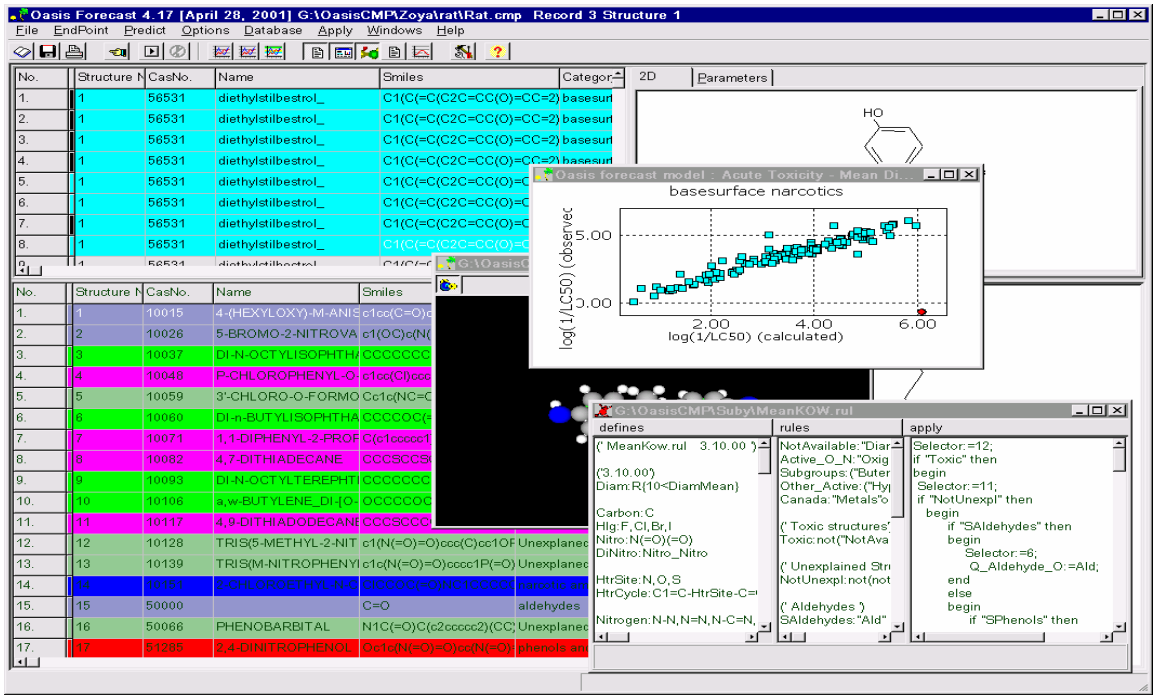
Фигура 4.9.

4.1.3. Програмна система FORECAST

Извеждането на модел на дадено (биологично) свойство на химичните съединения има за цел последващото му прилагане върху други химични съединения, за предсказване на техните свойства. Моделът, получен чрез COREPA представлява дърво на решението, като във всеки възел на дървото е поставено условие, задаващо критерий, свързан със структурата на молекулата или други изчислени или експериментално получени параметри. Правилата се записват чрез средствата на разширено SMILES означение и на разработения формален език (т. 2.2) и могат да се прилагат многократно върху различни бази данни.

Дървото на решението представлява класификатор, който класифицира химичното съединение в една от няколко групи (например на активни и нактивни съединения), въз основа на изведения модел за механизма на действие на биологичното свойство.

След като моделът е построен, и съединенията класифицирани, то се прилага метод за количествена оценка на активността (например на токсичността) в групата съединения с общ механизъм на действие. За количествена оценка също се използуват различни типове молекулни параметри, като физикохимични свойства, квантово химични електронни или стереохимични параметри.



Фигура 4.10.

С тази цел е разработена програмна система FORECAST, която позволява:

- 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);

- Вход от двумерни и тримерни модели на химичните съединения, както и автоматично изчисляване на тримерните координати при зададена свързаност;
- Автоматично изчисляване на необходимите молекулни дескриптори (геометрични, електронни, физикохимични и други);
- Конформационно размножаване на съединенията при желание от потребителя, чрез описания в т. 3.1 генетичен алгоритъм;
- Визуализация на разпределението на конформерите по енергия;
- Двумерна и тримерна визуализация на химичните структури.
- 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. Прилага се класификация чрез дърво на решенията, като по-нататък за всяка група се извежда регресионен модел, който дава количествена оценка на активността (токсичността) на съединенията с общ механизъм на действие.

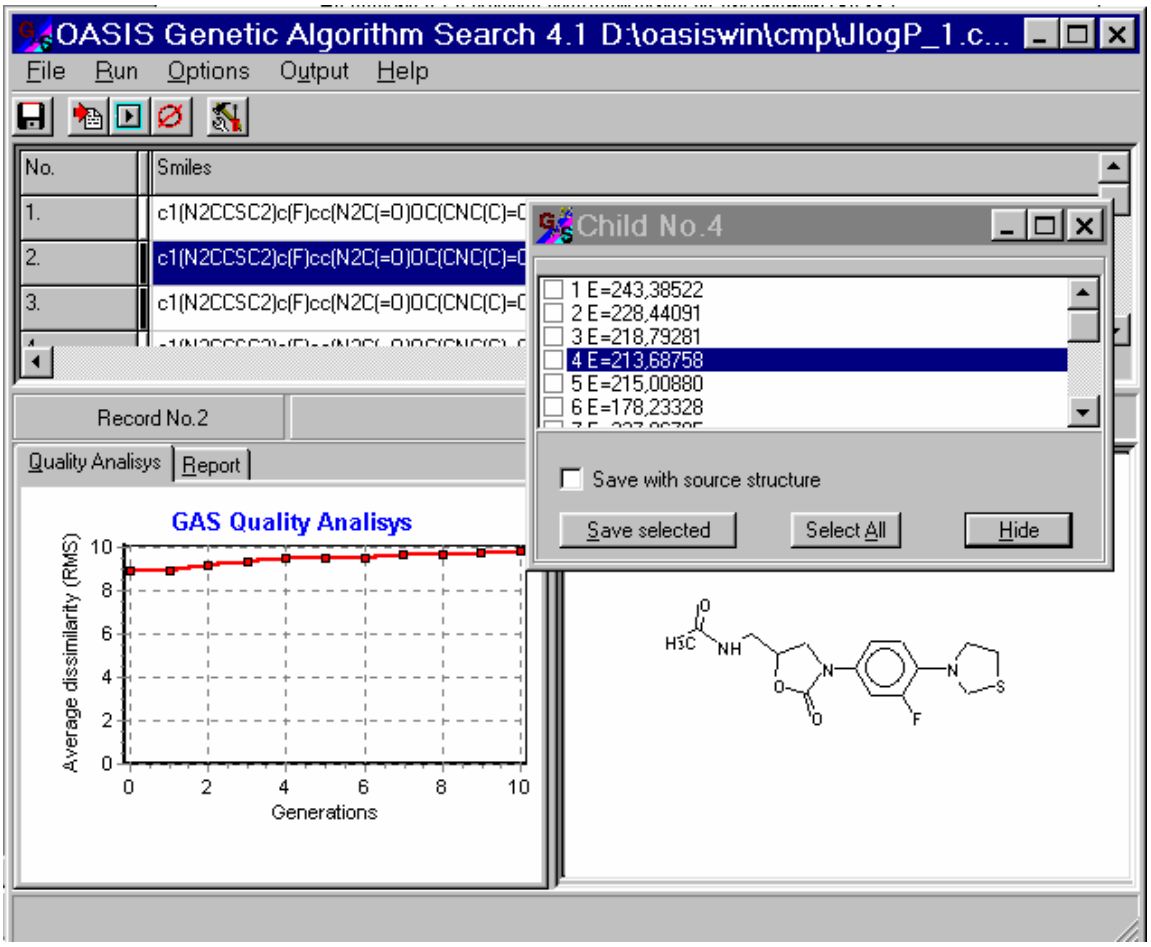
На **Error! Reference source not found.** е показан главният екран на програмата FORECAST.

4.1.4. Програмна система Genetic Algorithm Search (GAS32)

Функционалност

В програмната система GAS32 е реализиран генетичният алгоритъм за генериране на множеството от максимално различни конформери, описан в точка 3.1.

На Фигура 4.11 е показан главният екран на програмата GAS32.



Фигура 4.11

Системата позволява:

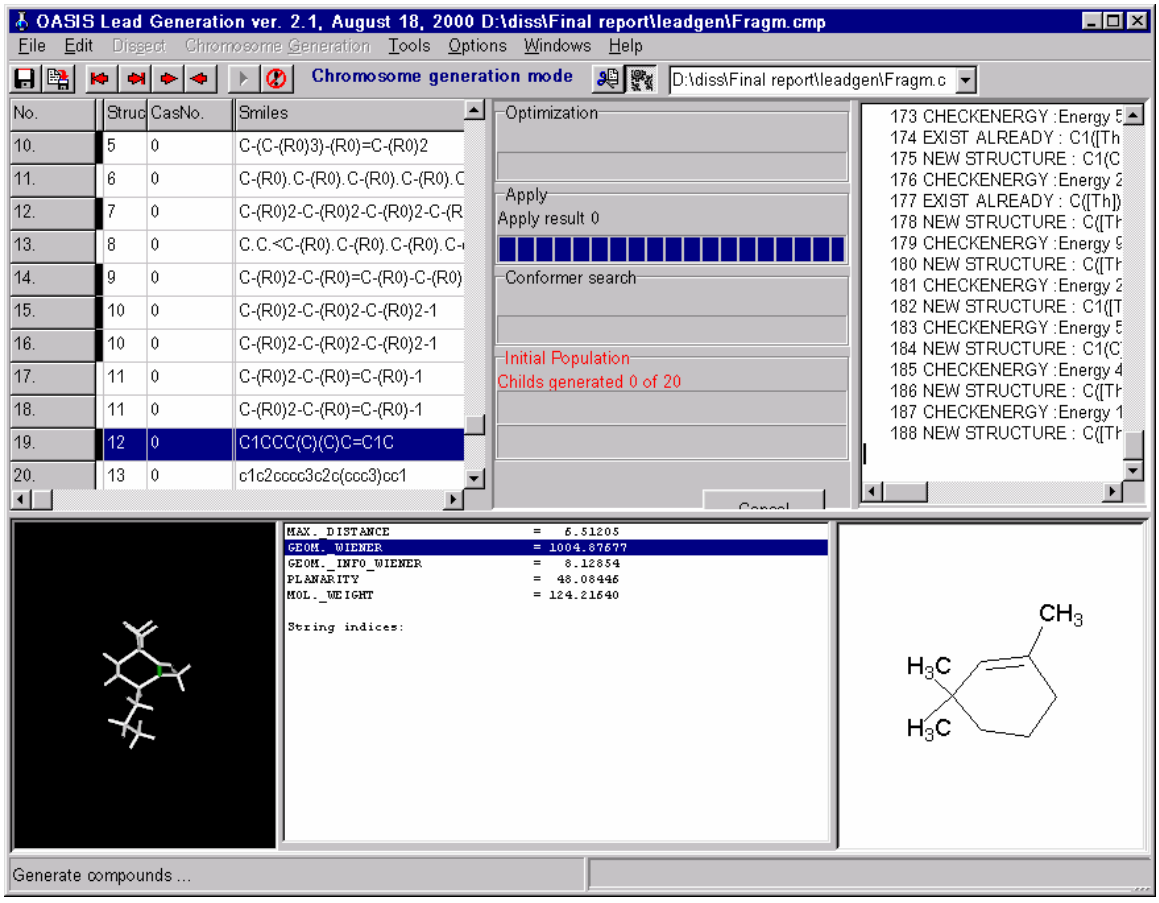
- Отваряне на *.cmp* файл
- Задаване на настройки на генетичния алгоритъм (размер на популацията, критерий за край и други)
- Генериране на конформери на едно или повече избрани съединения и визуализирането им

Изследване влиянието на различните настройки върху алгоритъма, чрез многократното му изпълнение при фиксиране на част от настройките.

4.1.5. Програмна система LEAD GENerator (LEADGEN)

В програмната система LEADGEN е реализиран описаният в т.3.2 генетичен алгоритъм за генериране на нови химични съединения при зададени ограничения.

На Фигура 4.12. е показан главния екран на програмната система LEADGEN



Фигура 4.12.

Системата *LEADGEN* 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, чрез разкъсване на структурите от произволен *.cmp* файл по прости връзки. Получените прости фрагменти се записват в *.cmp* файл и могат да се използуват за конструиране на новите съединения

4.2. Приложение

4.2.1. Биологична активност на химичните съединения

Биологичната активност на химичните съединения обикновено експериментално се измерва с концентрацията на изследваното вещество, необходима, за да се прояви стандартен биологичен ефект. Числената оценка за биологичната активност може да бъде:

- Относителната концентрация $(c_I / c_A) * 100\%$, където c_A е концентрацията, необходима за проява на биологичен ефект при съединение *A*, а c_I е концентрацията, необходима за проява на биологичния ефект при съединение *I*.
- Афинитет за свързване към известен рецептор на изследваното вещество. Оценката за афинитета обикновено е относителна спрямо еталонно химическо вещество, и се измерва като количеството еталонно вещество, изместено от рецептора, в резултат на взаимодействието на изследваното вещество с рецептора.

Биологична активност е обобщено понятие за ефекта на химично съединение в биологични системи. В токсикологията представляват интерес следните свойства:

Остра токсичност (*acute toxicity*)
Фототоксичност (*phototoxicity*)
Метаболизъм (*metabolism*)
Способност за разрушаване на ендокринната система (*endocrine disruption ability*) , афинитет за свързване към естрогенния/андрогенен рецептор (*estrogenic /androgenic binding affinity*)
Афинитет за свързване към ретиноидния рецептор (*retinoid activity, Vitamin A = retinoic acid=retinol*)
Чувствителност на кожата (*Skin Sensitization Potential*)
Канцерогенност (*cancerogenicity*)

Експерименталната оценка *in vitro* (тестове с клетки) или *in vivo* (тестове с организми) на тези свойства за всички съединения от съществуващите бази данни е трудоемка и практически неосъществима. Част от известните бази данни с химични съединения са тези на Химическото бюро на Европейската общност:

European **IN**ventory of **E**xisting **C**ommercial **S**ubstances (EINECS) - 100 106 вещества
High **P**roduction **V**olume **C**hemicals – **HVPC data set –** около 2500 вещества
Low **P**roduction **V**olume **C**hemicals – **LVPC** data set **-** 66675 вещества

В САЩ е известна базата данни на Американската агенция по опазване на околната среда:

TSCA (US Environmental Protection Agency data set**)**

В редица европейски и национални документи е предвидено извършването на систематична оценка на химичните съединения от тези бази. Поради невъзможността за експериментална оценка, се търсят надеждни компютърни методи за предсказване на биологичните свойства.

4.2.2. 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 и Американската Агенция за опазване на околната среда 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 се построява дърво на решението, което по-нататък позволява предсказването на естрогенната активност на съединенията. Експерименталните данни за естрогенната активност на дадено съединение се получават по следния начин: радиоактивно белязаният 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.

Използувани бази данни от химични съединения с експериментална оценка на биологичната им активност (обучаващо множество)

За извеждане на модела се използува множество съединения с известна експериментална оценка на естрогенната активност, малка част от които са показани в Таблица 4.1. За всяко съединение от таблицата са дадени неговото химично наименование, начин на действие (*агонист* - молекула, която се свързва към рецептора и инициира действието му; *антагонист* - молекула, която се свързва към рецептора и блокира действието му), експериментална оценка на относителния афинитет на свързване с естрогенния рецептор, брой генерирани конформери, които се използуват при извеждането на модела.

Таблица 4.1

	Химично Съединение	Действие	Относителен афинитет на свързване с естрогенния рецептор	Брой генерирани конформери
	Hexestrol	агонист	302	6
	17β-estradiol	агонист	100	4
	Coumestrol	агонист	94	11
	β-zearalanol	агонист	16	48
	5-Androstenediol	агонист	6	4
	4-OH Tamoxifen	антагонист	178	63
	LY117018	антагонист	100	44
	ICI - 182,780	антагонист	89	46
	ICI- 164,384	антагонист	85	102
	Tamoxifen	антагонист	7	149

Бази данни от химични съединения , за които е прогнозирана биологичната активност чрез дърво на решението

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, **TSCA**, **EINECS**. 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 / въвеждането на множество физикохимични параметри, които ще се използуват за моделиране на биологичната активност. Тези дейности са извършени с помощта на разработените програмни системи.

4.2.3. Получаване на модела за предсказване на естрогенната активност чрез програмна система COREPA

Според описания в точка 2.1.1 алгоритъм, приложението на COREPA изисква две предварителни стъпки:

- Разделяне на обучаващото множество на групи от активни и неактивни съединения според стойностите на активността
- Избор на множество от параметри (молекулни дескриптори), които потребителят предполага, че могат да имат отношение към активността.

Разделяне на групи в зависимост от активността

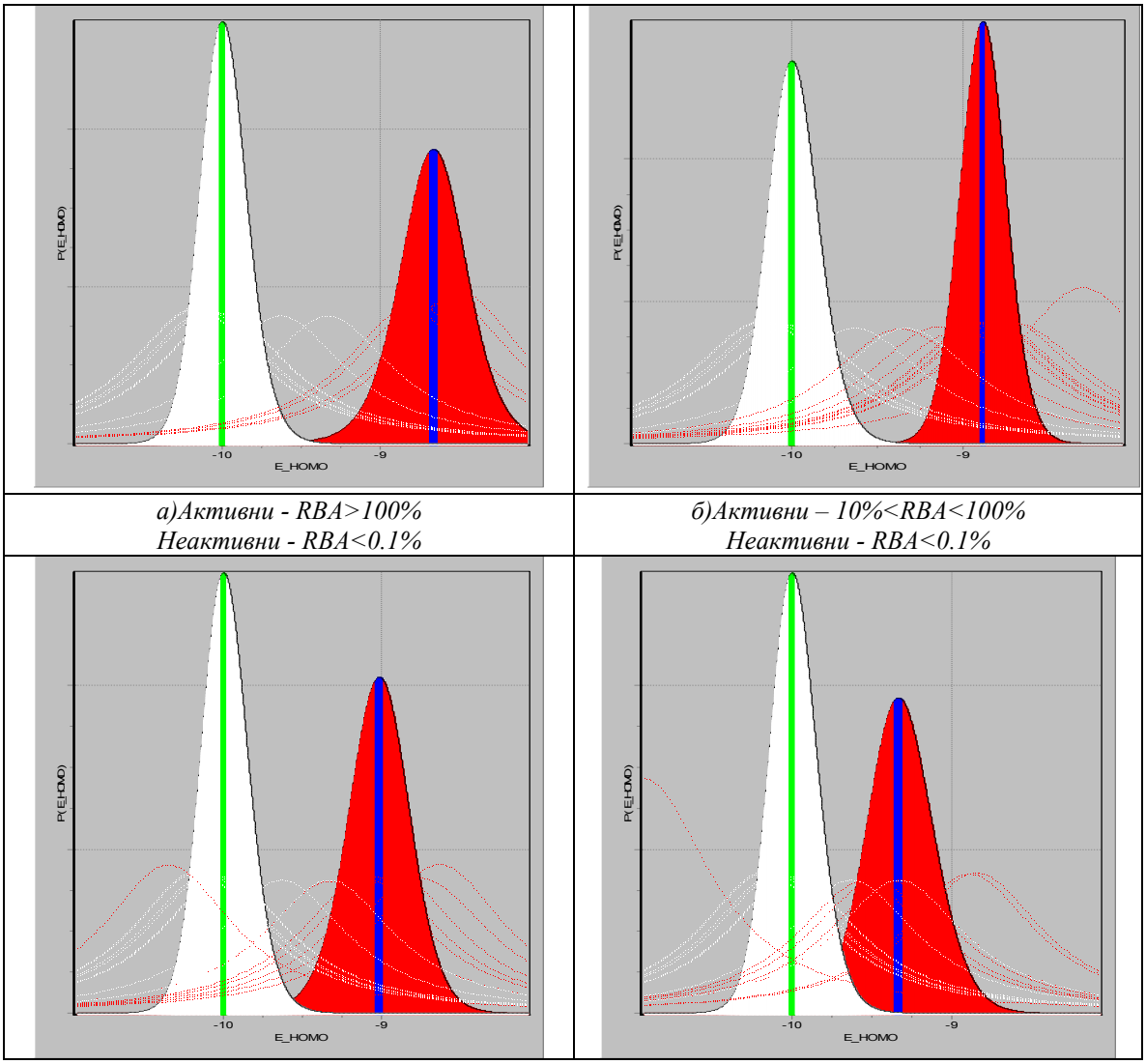
Съединията с експериментално определена естрогенна активност са разделени в следните групи:

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%

Избор на молекулни дескриптори

Множеството молекулни дескриптори се избират според известните от литературата хипотези за това кои глобални и локални дескриптори са свързани с афинитета към естрогения рецептор. [94, 95, 93, 90, 98, 99, 7, 62, 64]. Стереоелектронните параметри се изчисляват с MOPAC 93, разширен с модул, който позволява изчисляването на допълнителни дескриптори [66]. Като глобални параметри (оценяващи цялата молекула) се използуват електронегативност (EN; изчислена като $0.5(E_{HOMO} + E_{LUMO})$), диполен момент (μ), енергия на граничните орбитали (E_{HOMO} и E_{LUMO}), разликата между граничните енергетични нива ($E_{HOMO-LUMO}$, или E-gap), а като параметри за индивидуалните атоми се използуват атомните заряди, (q_i), граничните атомни заряди (f_i^{HOMO} и f_i^{LUMO}), донорните и акцепторни суперделокализационни индекси (S_i^E и S_i^N) и атомните поляризуемости (π_i). (i обозначава i-ия атом в молекулата).

Използуват се също дескриптори, оценяващи геометричната конфигурация на молекулата (*стерични дескриптори*) - *GW* (сума от геометричните разстояния [65], L_{max} (най-голямото междуатомно разстояние), $d\{ij\}$ (разстояния между атоми i и j), планарност (нормализирана сума на торзионните ъгли в молекулата [66]). Тези параметри са избрани поради това, че хидрофобността, пространственото разположение и размера на молекулата са посочени като важни параметри за предсказване и интерпретация на свързването на химичните съединения към естрогенния рецептор. [94, 95, 7, 99, 51, 30]. VolP е избран като физикохимичен дескриптор, дефиниран като сума от атомните поляризуемости, и представлява усреднената оценка на съединението да променя електронната си плътност по време на химичните взаимодействия [83, 51]. Използуването на този параметър се основава на предишни наблюдения, че по-поляризуемите химични съединения обикновено имат по-голям афинитет за свързване с естрогенния рецептор [7]. Ниски стойности на VolP (VolP > 0) отразяват по-високата концентрация на заряд в молекулата и съответно по-висока поляризуемост (по-малка липофилност) [51].



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 4.13.

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 4.13. 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

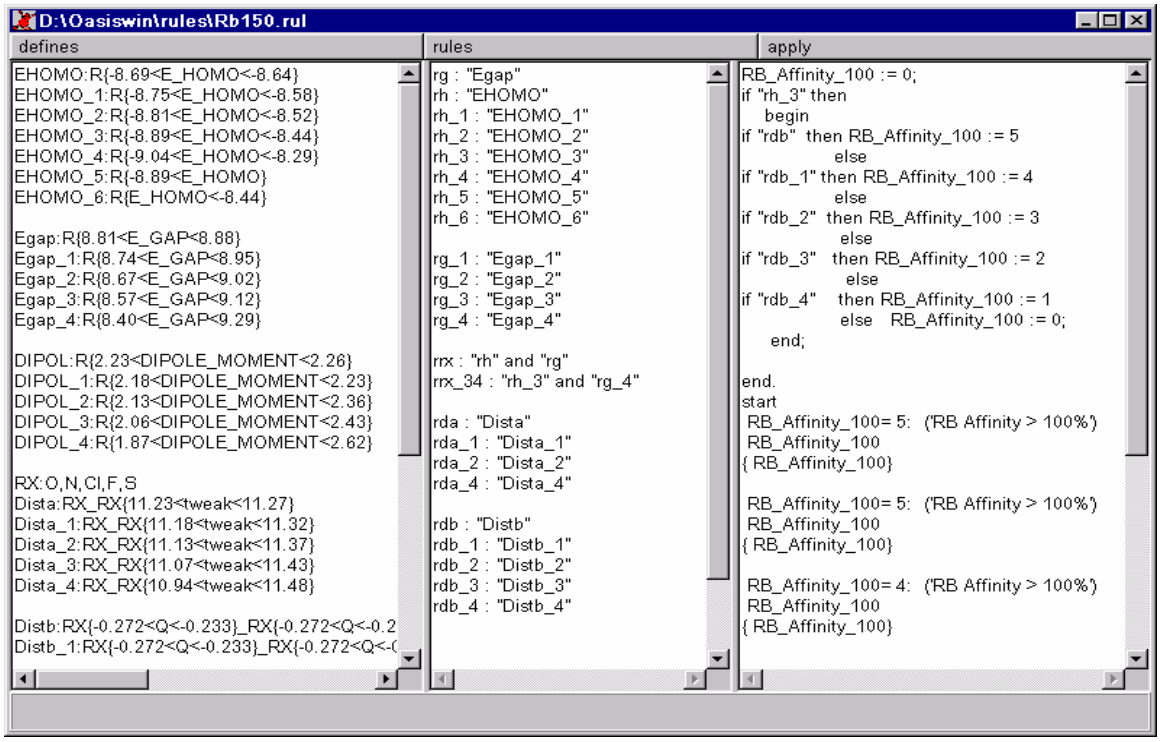
неактивните съединения (*RBA<0.1%*). От фигурите се вижда, че с нарастване на биологичното сходство между тестовите подмножества от Фигура 4.13.а до Фигура 4.13..г, нараства и сходството между вероятностните плътности за активните и неактивни съединения – те се припокриват по-значително и параметърът става по-малко дискриминиращ. Корелацията между биологичното сходство на съединенията от тестовите подгрупи и сходството между съответните COREPA образи е доказателство за връзката между глобалната нуклеофилност на съединенията и тяхната биологична активност.

Дърво на решението

Съгласно описания в точка 2.1.1 алгоритъм е изведено по едно дърво на решението за всяка от разглежданите 5 групи съединения с близка естрогенна активност.

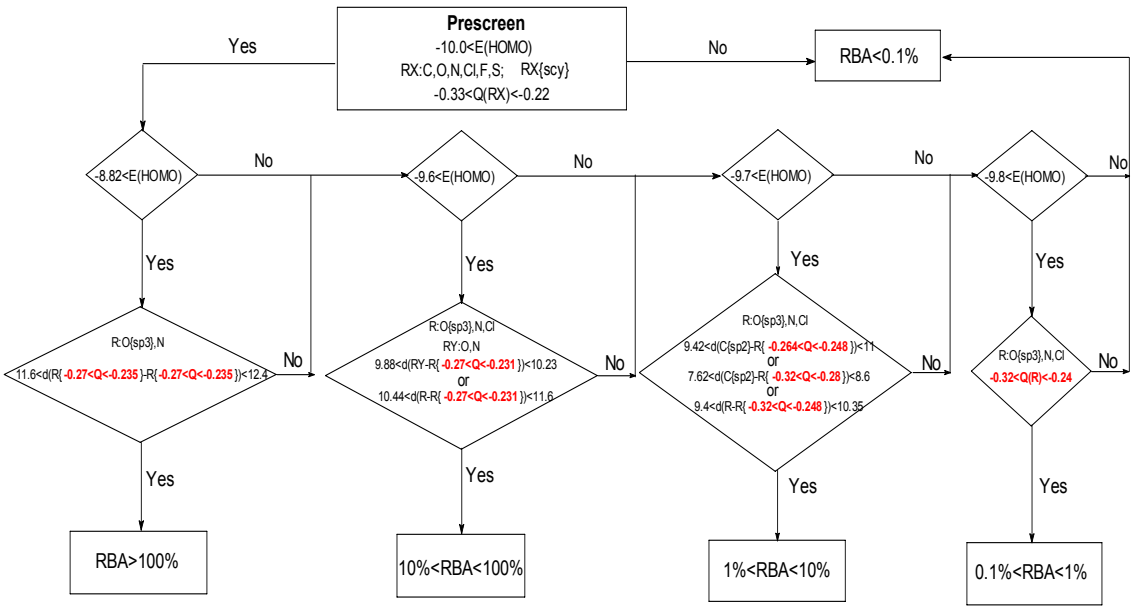
Група 1 (висока активност)	RBA>100%
Група 2	10% < RBA<100%
Група 3	1% < RBA<10%
Група 4 (ниска активност)	0.1% < RBA<1%
Група 5 (неактивни)	RBA<0.1%

На Фигура 4.14. е показано дървото на решението за групата на съединения с висока активност (RBA > 100%) , записано като RDL правила.



Фигура 4.14.

На Фигура 4.15. е показано опростен вариант на дърво на решението за всички групи на активност.



Фигура 4.15.

Валидиране на модела

Моделът за биологичната активност може да бъде изведен, като се използват 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 използваните. Във втория случай проблемът се състои в наличието на известни различия в механизма на рецепторно взаимодействие при различните организми. Построяването на QSAR модела е извършено на три етапа - модел I, модел II, модел III.

Таблица 4.2

Relative Binding Affinity	RBA>100%	10%<RBA<100%	1% < RBA < 10%	0.1% < RBA < 10%
Data set 1	100% predicted	90% Predicted False positives No False negatives	70% predicted False positives No False negatives	70% predicted False positives No False negatives
Data set 2	100% predicted	87% Predicted False positives No False negatives	81% predicted False positives No False negatives	51% predicted False positives No False negatives

За построяването на модел I е използувана изследваната от Waller *et al*. (1996) 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 (използването 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 (Таблица 4.3)

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 4.3 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 4.3

	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'-	+	n	n	neg	neg

		methylenebis(cyclohexylamine)					
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

a Chemicals with non-zero pKi rank, according to Model I

b Predicted RBA<1%

neg- естрогенна активност не е наблюдавана

e No tested chemilals

c Relative potency to Estrodiol

d Relative potency to Estrodiol < 0.0000

f Reference negatives

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 4.4 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 4.4

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

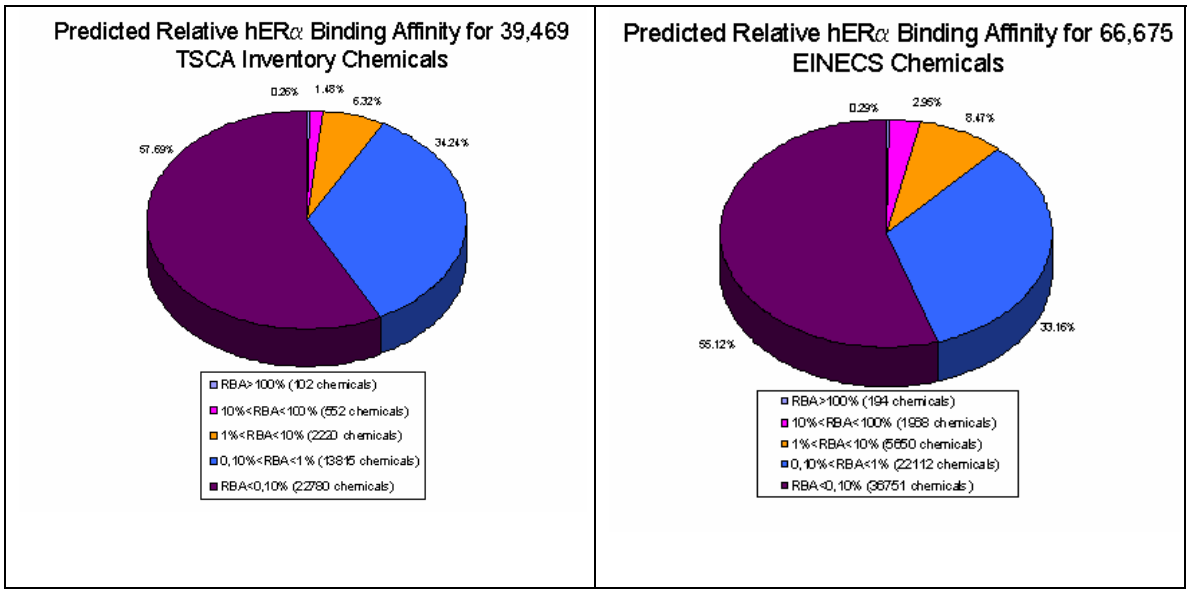
c lowest effect levels

d potent antiestrogen

Въпреки количествените QSAR оценки, направени в проекта EDAEP, тези прогнози трябва да бъдат възприемани като качествени. Така например, съединенията с прогнози RBA>100% трябва да бъдат считани за такива с висока естрогенна активност, а не точно RBA>100%. От 10 LPV съединения, за три е потвърдена активност RBA > 100% (според ERE-CALUX test). Останалите активни съединения обаче имат активност много по-голяма от тази на предсказаните като слаби естрогени. Експерименталните резултати за слабите естрогени, от своя страна показват стойности по-малки от стойностите получени от QSAR модела 1<RBA<10%. Тези съединения трябва да бъдат считани за слаби естрогени, а не за имащи активност точно в интервала 1<RBA<10%. Като извод може да се посочи, че QSAR моделът коректно идентифицира съединенията с естрогенна активност, но завишава числената оценка на активността. Този факт се дължи вероятно на принципа на извеждане на модела, при който целта е била да се минимизира броя на реално активните съединения, прогнозирани като неактивни, за сметка на неактивните съединения, прогнозирани като активни. От друга страна, трябва да се вземе предвид, че прогнозирането е извършено върху голяма база данни, съдържаща само по един конформер от всяко съединение и не е известно дали това е активния изомер. За разлика от това, в базата данни, използувана за построяване на модела се съдържат множество конформери. Проблемът за отчитане на конформационната гъвкавост при скрининга е решен с използуването на 'tweak' алгоритъм. Този алгоритъм е ефективен за скрининг на големи бази данни, но не осигурява достатъчно добро покритие на конформационното пространство, особено при циклични структури (tweak алгоритъмът изследва гъвкавостта само на цикличните фрагменти).

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 (HVPC), 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 4.16. са показани резултатите от

предсказване на афинитета за свързване с естрогенния рецептор за базите данни *TSCA* и *EINECS*.



Фигура 4.16.

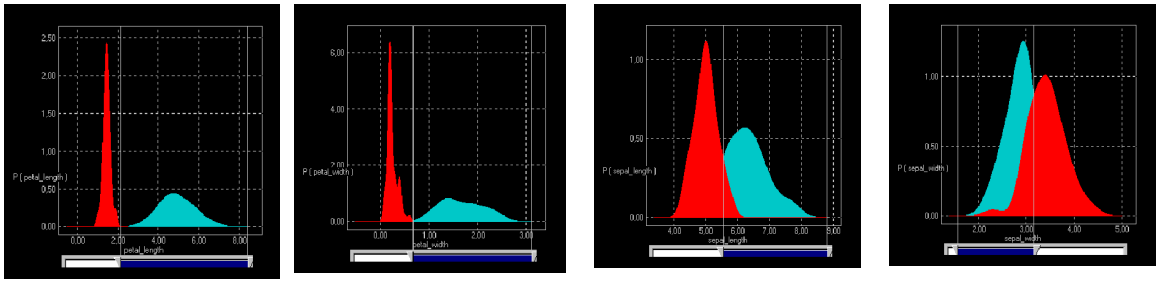
При анализ на получените прогнози трябва да се има предвид, че моделите са основани на експерименталните резултати от *in-vitro* тестове (експерименти с клетки), които се обуславят от степента на свързване с рецептор. Следователно, въз основа на тези данни могат да бъдат правени прогнози само за степента на свързване с рецептора.

In-vivo поведението на химичните съединения е интегрален резултат от голям брой взаимодействия, които могат да "затъмнят" ефекта от свързването с естрогенния рецептор. Тази сложност е причина за несъответствието между някои *in vitro* и *in vivo* (експерименти с живи организми) резултатите за някои химически съединения и не може да се отчете като недостатък на моделирането.

4.2.4. IRIS data set

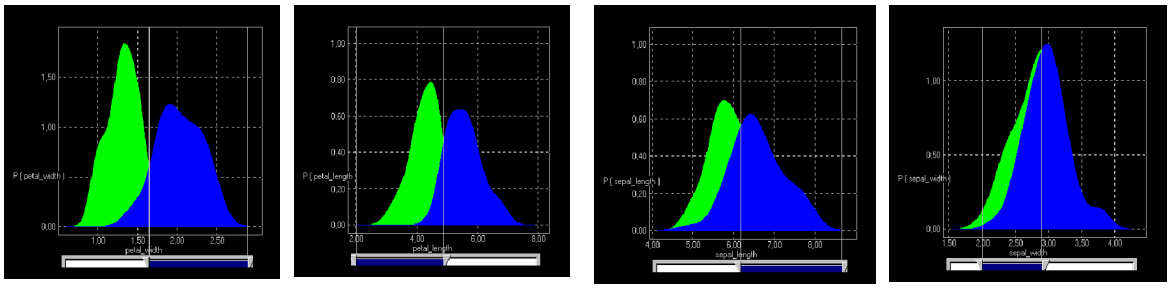
Проведен е експеримент, при който COREPA методът е приложен към известното в литературата за разпознаване на образи множество данни *IRIS data set* . То съдържа данни за 3 разновидности на ириси (*setosa, virginica, versicolor* - по 50 точки за

всяка), като за всяка точка (цвете) са измерени 4 параметъра : *petal length, petal width, sepal length, sepal width*. Първоначално данните са разделени на две групи - 50 точки от класа *setosa* и 100 точки от класовете *virginica* и *versicolor*. На Фигура 4.17.са показани оценените вероятностни плътности спрямо всеки параметър, за двете групи. Параметрите *petal length* и *petal width* дават най-добро разделяне за тези групи. За оценка на разделянето е използувано разстояние на Helinger, което има максимална стойност 2 при различни вероятностни разпределения и 0 при еднакви. Полученото условие за *petal length* напълно отделя класа *setosa* от другите два (първото условие в дървото на решението наФигура 4.19.). По-нататък се разглеждат двете групи *virginica* и *versicolor*, за които по аналогичен начин се получават разделящите условия (Фигура 4.18.).



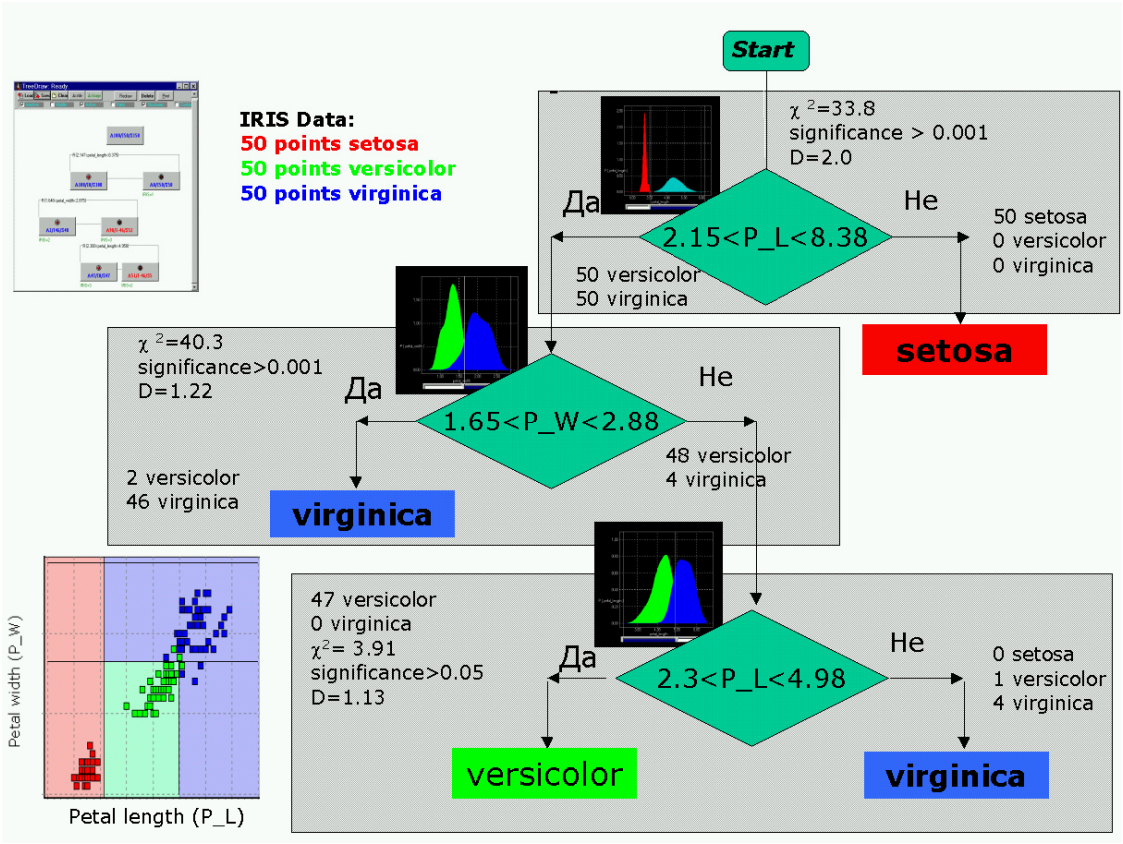
petal length	petal width	Sepal length	sepal width
D=2.0	D=1.97	D=1.02	D=0.49

Фигура 4.17.



petal length	petal width	Sepal length	sepal width
D=1.22	D=1.12	D=0.29	D=0.10

Фигура 4.18.



Фигура 4.19.

Резултантното дърво на решението е показано на Фигура 4.19. Класифицират се вярно 147 от всичко 150 точки.

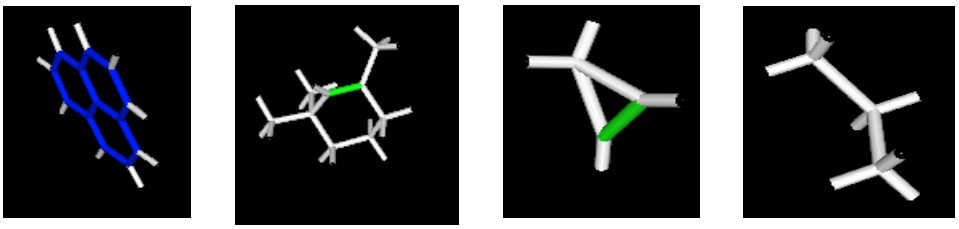
4.2.5. Приложение на генетичния алгоритъм за откриване на максимално различни конформери

Генетичният алгоритъм за намиране на множество от максимално различни конформери се използува като

- Самостоятелна програмна система GAS32
- Модул във програмните системи FORECAST, COREPA, LEADGEN, което позволява генерирането на конформери при необходимост за съответната задача.

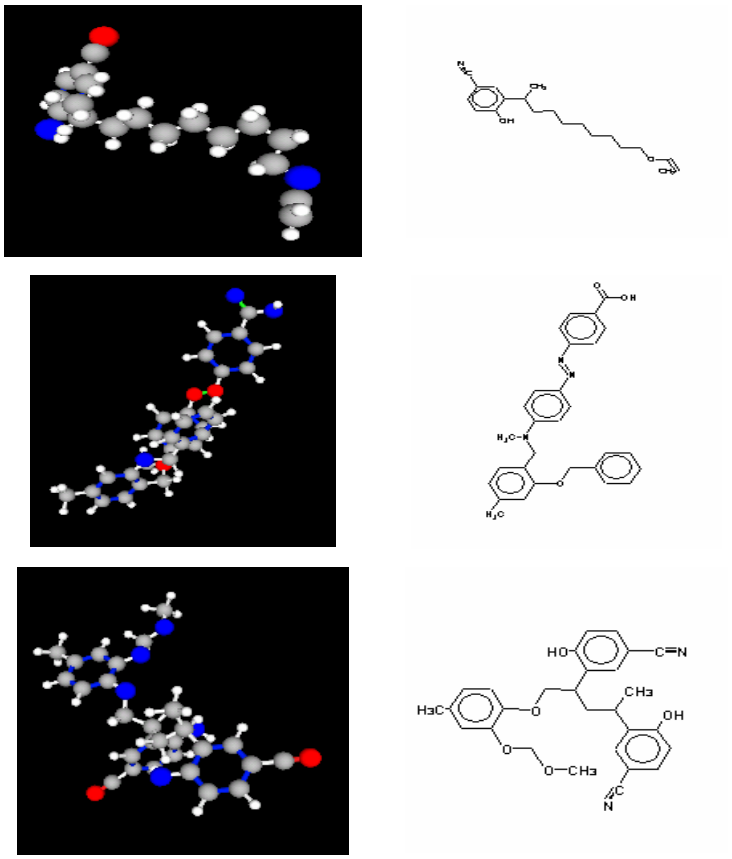
4.2.6. Използуване на моделите за генериране на нови обекти (LEADGEN)

На Фигура 4.20. са показани част от фрагментите, използувани в програмната система Leadgen за генериране на нови съединения с висока естрогенна активност.



Фигура 4.20.

На Фигура 4.21.са показани част от новите съединения, генерирани от програмната система Leadgen.



Фигура 4.21.

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 субмолекулните свойства на дадена биоактивност. 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.

Приложение I

SMILES

SMILES е синтаксис за запис на молекулната свързаност в символен низ. Молекулната свързаност е зададена като молекулен граф, където възлите на графа съответстват на атомите в молекулата, а ребрата на графа съответстват на (ковалентните) връзки. Следват дефиниции на атоми, връзки, вериги, разклонения и циклични структури.

Атоми

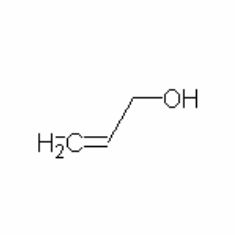
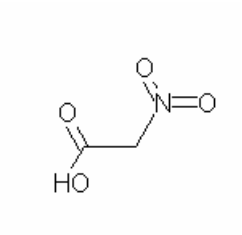
Атомите се представят чрез техните атомни символи. Втората буква на двубуквен атомен символ трябва да бъде малка. Водородните атоми, свързани към основните атоми, от които са изградени органичните съединения ((C, N, B, O, P, S, Si, F, Cl, Br, I) могат да бъдат пропуснати. Техният брой се извежда от най-малката валентност на атома и зададените му връзки. Символите на атоми извън това множество трябва да бъдат заградени в квадратни скоби. Атомите C и N , принадлежащи на ароматни цикли трябва да бъдат зададени с малки букви, съответно c и n.

Връзки

Единичните, двойни, тройни, ароматни и йонни връзки се обозначават съответно със символите '-', '=', '#', '*', '.' . При несвързани структури се използува символът '_'. Символите за единични и ароматни връзки могат да бъдат пропускани.

Вериги

Линейните структури се представят със поредицата от атомите, образуващи веригата. Например структурата на фигура П.1 може да бъде представена с C=C-C-O или C=CCO или OCC=C. (подразбира се, че атомите, които нямат буквено означение на рисунката на фрагмента са въглеродни атоми - C).

Структура		
	C=C-C-O	N(=O)(=O)-CC(O)=O
SMILES	C=CCO	C(N=(O)=O)C=(O)-O
Низ	OCC=C	O=C(O)CN=(O)=O
	Фигура П.1	Фигура П.2

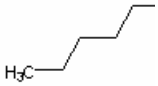
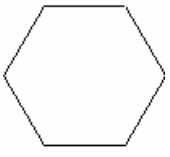
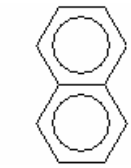
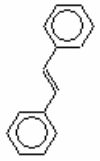
Разклонения

Разклоненията се отбелязват, като се обграждат със скоби Например, структурата на фигура П.2 може да бъде представена със всеки от трите SMILES стринга N(=O)(=O)-CC(O)=O или C(N=(O)=O)C=(O)-O или O=C(O)CN=(O)=O. Вложените скоби обозначават разклонения на разклоненията

Циклични структури

Представянето на цикличните структури е основано на т.нар. обхващашо дърво. Обхващащото дърво може да се получи чрез итеративно разкъсване на циклите, докато структурата не съдържа повече цикли. Разкъсаните връзки наричаме връзки, които затварят циклите. Обозначениет на цикличната структура се получава от SMILES обозначението на обхващашото дърво и индикатори на затварящите връзки. За индикаторите се използуват числата от 1 до 9, които се поставят след атомните символи на атомите, свързани със затварящите връзки.

Циклите се маркират с цели числа от 1 до 9, които се поставят по двойки след символите на свързаните в цикъл атоми. Примери:

Структура				
SMILES стринг	CCCCCC	C1CCCCC1	c12ccccc1cccc2	c1ccccc1C=Cc1ccccc1

Фигура П.3

Приложение II

Разширение на SMILES

В описаната по-горе SMILES спецификация <num> обозначава затваряне на цикъл.

<atomlabel>::= C | c | N | n | O | o | S | s | B | P | F | Cl | Br | I | R | <otheratom>

<otheratom> е символ на химичен елемен, ограден с квадратни скоби [ ]

R обозначава произволен тежък атом при подфрагментно търсене.

Връзката по подразбиране между атоми, записани с малки букви е ароматна, между атоми, записани с големи букви е единична.

<atomqualifier>::= {<adlabel>} | {‘<realnum><’<atomdescriptor>‘<<realnum>’} | {‘<realnum><’<moldescriptor>‘<<realnum>’}

<adlabel>*::= acy | scy | dcy | sp1 | sp2 | sp3 | ar | 2+ | 2- | p+ | p- | h | h2 | h3 | h4 | + | - | .

Означения: acy – не принадлежи на цикъл; scy – принадлежи на цикъл; dcy – принадлежи на повече от един цикъл; ar – участвува в π-кондензирани циклир;p+ - positive chiral parity.

<bondentry>::= <bondlabel>[<bondqualifier>]

<bondlabel>::= - | = | # | * | . | _

Типовете химични връзки в зададения ред са : проста, двойна, тройна, ароматна, йонна, без връзка.

<bondqualifier>::= {<bdlabel>} | {‘<realnum>< ’<bondescriptor>‘<<realnum>’}

<bdlabel>* ::= t | c | g | g+ | g-

Означения: t - trans; c - cis; g - gauche; g+ - gauche clockwise; g- - gauche counter clockwise.

<atomdescriptor>* и <bondescriptor>* са имена от предварително зададен списък на дескриптори..

<alphanum>::= <letter> | <num>

<letter>::= a | b | … | z | A | B |…| Z

<num>::= 0 | 1 | 2 | .. | 9

<realnum>::= ‘-’<num>[<num>]‘.[num]’

Литература

1. Basak S., Grunwald G., Niemi G., Use of Graph-Theoretic and Geometric Molecular Descriptors in Structure-Activity Relationships, in From Chemical Topology to Three-Dimensional Geometry, edited by Balaban A., Plenum Press N.Y., 1997
2. Baxter M.J., Beardah C.C.,Beyond the histogram − improved approaches to simple data display in archaeology using kernel density estimates, Department of Mathematics, Statistics and Operational Research, The Nottingham Trent University, http://science.ntu.ac.uk/msor/ccb/romenew.ps
3. Baxter M.J., Beardah C.C.,MATLAB Routines for Kernel Density Estimation and the Graphical Representation of Archaeological Data Department of Mathematics, Statistics and Operational Research, The Nottingham Trent University, http://science.ntu.ac.uk/msor/ccb/caarev.ps
4. Boething R.S., Mackay D. (editors), Handbook of Property Estimation Methods for Chemicals. Environmental and Health Sciences, Lews Publishers, 2000
5. Bohacek R.S., McMartin C., Multiple Highly Diverse Structures Complementary to Enzyme Binding Sites: Results of Extensive Application of a de Novo Design Method Incorporating Combinatorial Growth
6. Bonchev, D., 1983, Information-theoretic Indices for Characterization of Chemical Structures, Research Studies Press, Chichester
7. Bradbury S.P., Mekenyan O.G., Ankley G.T. 1996. Quantitative structure-activity relationships for polychlorinated hydroxybiphenyl estrogen receptor binding affinity: an assessment of conformational flexibility. *Environ Chem Toxicol* 15:1945-1954.
8. Breiman, L., Friedman, J., Olshen, R., and Stone, C. Classification and Regression Trees, Wadsworth International Group, Belmont, CA, 1984
9. Brian M., Barnard E., Phone Clustering Using The Bhattacharyya Distance, Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, http://citeseer.nj.nec.com/2625.html

10. Byholt M., Hjort N.L., Sometimes nonparametric beat parametric even when the model is right, Statistical Research Report, Dep. Of Mathematics, University of Oslo, No. 18, Oct.1996
11. Chang, G.; Guida, W. C.; Still, W. C. An Internal Cordinate Monte Carlo Method for Searching Conformational Space. *J. Am. Chem. Soc.* 1989, 111, 4379-4386.
12. Charifson P. S. (ed), Practical Application of Computer-Aided Drug Design, Vertex Pharmaceutical Incorporated, Cambridge, Massachusetts
13. Charifson P. S., Leach A. R., Rusinko A., The Generation and Use of Large 3D Databases in Drug Discovery, Glaxo Wellcome Inc. , http://www.netsci.org/Science/Cheminform/feature03.html
14. Chem-X is supplied by Chemical Design Ltd, Roundway House, Cromwell Park, Chipping Norton, Oxfordshire OX7 5SR UK.
15. Clark D.E., Firth M.A., Murray C.W., MOLMAKER: De Novo Generation of 3D Databases for Use in Drug Design, J.Chem.Inf.Comput.Sci, 1996, 36, 137-145
16. Clark, D.E., Jones, G., Willet, P., Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Compariso of Conformational – Searching Algorithms for Flexible Searching. J. Chem. Inf. Comput. Sci. 1994, 34, 197-206.
17. Combes R., The Use of Artificial Intelligence Systems for Predicting Toxicity, Pestic. Sci. 1995, 45, 179-194
18. CONCORD written by R.Pearlman et al. And distributed by TRIPOS Associates, St.Louis, MO, USA
19. Dalby A., Nourse J., Hounshell D., Gushurst A., Grier D., Leland B., Laufer J., Description of Several Structure File Formats Used By Computer Programs Developed at Molecular Design Limited, J. Chem. Inf. Comput. Sci, 1992, 32, 244-255
20. Davies K.,Upton R., 3D Pharmacophore Searching, Chemical Design Ltd., http://www.netsci.org/Science/Cheminform/feature02.html
21. Devillers J., Genetic Algorithms in Mollecular Modelling, Academic Press, London, 1996

22. Devroye L., Gyorfi L., Lugosi G., A probabilistic Theory of Pattern Recognition, Springer, 1996
23. Duda R., Hart P., Stork D., Pattern Classification, 2nd ed., John Wiley & Sons, 2000
24. Fayad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. (editors), Advances in Knowledge Discovery and Data Mining, The MIT Press, 1996
25. ftp://ftp.wiley.com/public/sci_tech_med/pattern
26. Gehlhaar D.K. , Moerder K.E., Zichi D., Sherman C.J., Ogden R.C., Freer S.T., De Novo Design of Enzyme Inhibitors by Monte Carlo Ligand Generation, J.Med.Chem. 1995, 38, 466-472
27. Gillet V.J., Newel W., Mata P., Myatt G., Sike S., Zsoldos Z., Johnson A.P., SPROUT: Recent Developments in the de Novo Design of Molecules, J.Chem. Inf. Comput. Sci., 1994, 34, 207-217
28. Glen R., Payne A., A Genetic Algorithm for the Automated Generation of Molecules Within Constraints, J. of Computer-Aided Molecular Design, 9 (1995) 181-202
29. Goldstein D.L.,Katzenellenbogen J.A., Luthey-Shulten Z.A., Seielstad D.A., Wolynes P.G., Three-Dimensional Model for Hormone Binding Domains of Steroid Receptors, Proc Natl Acad.Sci. USA (Biochemistry), 1993, 90, 9949-9953
30. Goldstein RA, Katzenellenbogen JA, Luthey-Schulten ZA, Seielstad DA, Wolynes PG. 1993. Three-dimensional model for the hormone binding domains of steroid receptors. *Proc Natl Acad Sci USA* (Biochemistry) 90:9949-9953.
31. Haykin, S. Neural Networks: A Comprehensive Foundation , 1998 , Prentice Hall
32. http://www.cse.unsw.edu.au/~quinlan/
33. http://www.recursive-partitioning.com/Classification/
34. http://www.statsoft.com/textbook/stclatre.html
35. http://www.uni-karlsruhe.de/~ed01/Hyle/Hyle5/mainzer.htm
36. http://yoda.cis.temple.edu:8080/UGAIWWW/lectures/C45/
37. Hurst T., Flexible 3D Searching: The Directed Tweak Technique. J.Chem.Inf. Comput. Sci., 1994, 34, 190-196
38. Ivanov, J. M., Karabunarliev, S.H., Mekenyan O.G., 3DGEN: A System for an Exhaustive 3D Molecular Design. J. Chem. Inf. Comput. Sci. 1994, 34, 234-243

39. Jones G., Wilett P., Glen R., Molecular Recognition of Receptor Sites using a Genetic Algorithm with a Description of Desolvation, J.Mol.Biol. (1995) 245, 43-53
40. Jones, G., Willet, P., Glen R.C., Genetic Algorithms for Chemical Structure Handling and Molecular Recognition. In Genetic Algorithms in Molecular Modeling; Devillers J., Ed., Academic Press: London, 1996; p 211.
41. Judson R. et al., Docking Flexible Molecules: A Case Study of Three Proteins., J. of Computational Chemistry, vol. 16, No. 11, 1405-1419
42. Karabunarliev S., Dimitrov S., ***Nikolova N.***, Mekenyan O., Prediction of Acute Aquatic Toxicity of Noncongeneric Chemicals: Rule-based and Quantitative Structure-Activity Relationships, In Proc. 8th workshop on QSAR in the Environmental Sciences, Baltimore'98
43. Karabunarliev, S., Dimitrov, S., ***Nikolova, N.,*** Mekenyan O. Prediction of Acute Aquatic Toxicity of Noncongeneric Chemicals: Rule-based and Quantitative Structure-Activity Relationships. In: *Quantitative Structure-Activity Relationships and Endocrine Disruption*, Walker, J. D. (ed.); SETAC Press: Pensacola, FL, USA (in press 2000)
44. Karabunarliev, S., Mekenyan O.G., Karcher W., Russom C.L., Bradbury S.P., Quantum-Chemical Descriptors for Estimating the Acute Toxicity of Substituted Benzenes to the Guppy and Fathead Minnow, Quant.-Struc.-Act.Relat., 1996, 15, 311-320
45. Karabunarliev, S.H., ***Nikolova N.***, Nikolov N. and O.G. Mekenyan. . Rule interpreter: A chemical language that implements decision rules based on molecular structure. *J. Chem. Inf. Comput. Sci.* (submitted 2000).
46. Karcher W., Karabunarliev S., The Use of Computer-based SAR in the Risk Assessment of Industrial Chemicals, J. Chem. Inf. Comput. Sci., 1996, 36, 672-677.
47. Kohonen T., "Learning Vector Quantization", Helsinki University of Technology,Lab of Comp. And Information Science, Report TKK-F-A-601, 1986
48. Kubinyi H., ,Comparative Molecular Field Analysis (CoMFA), Computational Chemistry Encyclopedia

49. Labute P., QuaSAR Cluster: A Different view of Molecular Clustering, Journal of the Chemical ComputingGroup, http://www.chemcomp.com/article/cluster.htm
50. Langowski, J. Computer prediction of toxicity. Pharmaceutical Manifacturing International (International Review of Pharmaceutical Technology, Research and Development) , 1993, 77-80
51. Lewis DFV, Parker MG, King RJB. 1995. Molecular modeling of the human estrogen receptor ligand interactions based on site-directed mutagenesis and amino acid sequence homology. *J Steroid Biochem Mol Biol* 52:55-65.
52. Lim H., Bioinformatics and Cheminformatics in the Drug Discovery Cycle, Lecture Notes In Computer Science 1278, BioInformatics, German Conference on Bioinformatics, GCB'96, Leipzig, Germany, September/October 1996
53. Lipton M., Still, W.C. The Multiple Minimum Problem in Molecular Modeling. Tree Searching Internal Coordinate Conformational Space. J. Comput. Chem. 1988, 3, 3-21
54. MACCS is supplied by MDL Information Systems Inc. 14600 Catalina Street, San Leandro, CA 94577, USA
55. Manganaris S., A Bayesian Density Estimation Algorithm , Dept. of Computer Science , Vanderbilt University ,Box 1679, Station B ,Nashville, TN 37235, U.S.A. 1996
56. Marshall G.R. Binding Site Modelling of Unknown Receptors. In 3D QSAR in Drug Design: Theory, Methods and Applications: Kubinyi, H. Ed., Escom: Leiden, 1993, pp80-116
57. McLachlan G., Discriminant Analisys and Statistical Pattern Recognition, John Wiley and Sons, 1992
58. Mekenyan O., Karabunarliev S., ***Nikolova N.***, "The New OASIS tools for Fuzzy Modelling of Chemical Structures" , Годишник на Софийския Университет, т.91
59. 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.

60. Mekenyan O., Karabunarliev S., ***Nikolova N.***, New Developments in a Hazard Identification Algorithm For Hormone Receptor Ligands: COREPA-C, 8-th International Workshop on Quantitative Structure-Activity Relationship (QSAR) in the Environmental Sciences, May, 16-20, 1998, Baltimore, Marylend, USA.
61. Mekenyan O.G., Ivanov J., Karabunarliev S., Bradbury S.P., Ankley G., Karcher W., A Computationally-Based Hazard Identification Algorithm That Incorporates Ligand Flexibility. 1. Identification of Potential Androgen Receptor Ligands., Environ. Sci. Technol., 1997, 31, 3702-3711
62. Mekenyan O.G., Ivanov, J. M., Karabunarliev, S.H., Bradbury, S.P., Ankley G.D., Karcher, W., COREPA: A New Approach for the Elucidation of Common Reactivity Patterns of Chemicals. 1. Stereoelectronic requirements for androgen receptor binding. Environ. Sci. Technol. 1997, 31, 3702-3711
63. Mekenyan O.G., Ivanov, J. M., Veith, G.D., Bradbury, S.P., DYNAMIC QSAR: A New Search For Active Conformations and Significant Stereoelectronic Indices. Quant. Struct. - Act. Relat. 1994, 13, 302-307
64. 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
65. Mekenyan OG, Peitchev D, Bonchev D, Trinajstic N, and Bangov I. 1986. Modeling the interaction of small molecules with biomacromolecules. I. interaction of substituted pyridines with anti-3-azopyridine antibody, *Arzneim Forsch* 36:176-183.
66. Mekenyan, O. G., G.D. Veith. 1994. The Electronic Factor in QSAR: MO-Parameters, Competing Interactions, Reactivity and Toxicity, *SAR QSAR Environ. Res.* 2:129-143.
67. Mekenyan, O.G., Karabunarliev, S.H., Ivanov, J.M., Dimitrov, D.N., A New Development of the OASIS Computer System. Comput. Chem., 1994, 18, 173-187.
68. MOSES : A Multiobjective Optimization Tool for Engineering Design, Carlos A Coello & Alan D. Christiansen, Dep. Of Comp. Science, Tulane University, New Orleans, LA 70118, USA

69. Muller K.R., Mika S., Ratch G., Tsuda K., Scholkopf B., An Introduction to Kernel Based Learning Algorithms, IEEE Transaction on Neural Networks, vol.12, No.2, March 2001, pp 181-198
70. Myashita Y., Li Z., Sasaki S., Chemical Pattern Recognition and multivariate analisys for QSAR studies, Trends in Analitycal Chemistry, vol. 12, no.2, 1993
71. Nishibata Y., Itai A., Confirmation of Usefulness of a Structure Construction Program Based on Three-Dimensional Receptor Structure for Rational Lead Generation, J. Med.Chem. 1993, 36, 2921-2928
72. Oncologic is supplied by LogiChem. http://www.logichem.com/
73. Orr M.J.L., Introduction to radial basis networks, TR Centre for Cognitive Science, Univ. Of Edinburgh, Scotland, http://anc.ed.ac.uk/~mjo/, 1996
74. Pao, Y. H. (1989). Adaptive Pattern Recognition and Neural Networks, Addison-Wesley Publishing, Reading, Massachusetts, 1990.
75. Payne, A. W. R., Glen, R. C. Molecular recognition using a binary genetic search algorithm. J. Mol. Graphics. 1993, 11, 74-91
76. Pearlman D., Murcko M., CONCERTS: Dynamic Connection of Fragments as an Approach to de Novo Ligand Design, J.Med.Chem., 1996, 39, 1651-1663
77. Picket S.D., Mason J.S., McLay I.M, Diversity Profiling and Design Using 3D Pharmacophories: Pharmacophore-Derived Queries (PDQ), J.Chem.Inf.Comput.Sci, 1996, 36, 1214-1223
78. Richard A.M., Application of SAR Methods to Noncongeneric Data Bases Associated with Carcinogenicity and Mutagenicity: Issues and Approaches, Mutat. Res. 305, 73-97
79. Richon A. Stanley S.,An Introduction to QSAR Methodology, *Glaxo Wellcome Research,* http://www.netsci.org/Science/Compchem/feature19.html
80. Ruiz A., Lopez-de-Teruel P.E., Nonlinear Kernel-Based Statistical Pattern Analisys, IEEE Transaction on Neural Networks, vol.12, No.1, January 2001, pp. 16-32
81. Rumelhart, D.E., McClelland, J.E. (eds.): Parallel Distributed Processing, MIT Press, Cambridge, Massachusetts, 1986

82. Sammon W.,"A nonlinear mapping for data structure analisys", IEEE Trans. On Computers, vol.C-18,no.5,pp.401-409,May 1969
83. Schuurmann, G. 1990. Quantitative structure-property relationships for the polarization, solvatochromic parameters and lipophilicity. *Quant Struct -Act Relat* 59:326-333.
84. Scott, D.W., Density Estimation, Rice University, Houston, TX (http://rice.edu)
85. Silverman, B. W., Density Estimation for Statistics and Data Analysis, Chapman and Hall, 1986
86. Sippl, M.J., Stegbuchner H., Superposition of Three-Dimensional Objects: A Fast and Numerically Stable Algorithm for the Calculation of Matrix for Optimal Rotation. Comput. Chem. 1991, 15, 73-78
87. Smelie, A., Kahn, S.D., Teig. L., Analisys of Conformational Coverage, 1. Validation and Estimation of Coverage. J. Chem. Inf. Comput. Sci. 1995, 35, 285-294
88. Smelie, A., Kahn, S.D., Teig. L., Analisys of Conformational Coverage, 2. Application of Conformational Models. J. Chem. Inf. Comput. Sci. 1995, 35, 295-304
89. SYBYL. Tripos Associates Inc., St. Louis, MO.
90. VanderKuur JA, Wiese T, and Brooks SC. 1993. Influence of estrogen structure on nuclear binding and progesterone receptor induction by the receptor compex. *Biochemistry* 32:7002-7008.
91. VanDrie J.,3D Database Searching in Drug Discovery, Pharmacia & Upjohn Laboratories, http://www.netsci.org/Science/Cheminform/feature06.html
92. Veith G.D., Mekenyan O.G., A QSAR Approach for Estimating the Aquatic Toxicity of Soft Electrophiles, 1993, Quant. Struct.-Act.Relat., 1983, 12, 349-356
93. Waller CL, Juma BW, Earl Gray L, Jr., Kelce WR. 1996b. Three-dimensional quantitative relationships for androgen receptor ligands. *Toxicol Appl Pharmacol* 137:219-227.
94. Waller CL, Minor DL, McKinney JD. 1995. Examination of the estrogen receptor binding affinities of polychlorinated hydroxybiphenyls using three-dimensional quantitative structure-activity relationships. *Environ Health Perspect* 103:702-707.

95. Waller CL, Oprea TI, Chae K, Park H-K, Korach KS, Laws SC, Wiese TE, Kelce WR, and Gray LE, Jr. 1996a. Ligand-based identification of environmental estrogens. *Chem Res Toxicol* 9:1240-1248.

96. Wehrens, R., Pretsch, E., Buydens L., Quality Criteria for Genetic Algorithms for Structure Optimization, J.Chem.Inf.Comput.Sci, 1998, 38, 151-157

97. Weininger, D., SMILES, a Chemical Language and Information System. Part 1, J.Chem. Inf. Comput. Sci. 1988, 28, 31-36

98. Wiese T, Brooks SC. 1994. Molecular modeling of steroidal estrogens: novel conformations and their role in biological activity. *J Steroid Biochem Molec Biol* 50:61-72.

99. Wurtz JM, Bourguet W, Renaud JP, Vivat V, Chambon P, Moras D, Gronemeyer H. 1996. A canonical structure for the ligand-binding domain of nuclear receptors. *Nature Struct Biol* 3:87-94.

100. Бончев Д., Мекенян О., Физикохимия – строеж и свойства на молекулите, Наука и изкуство, София , 1994.

101. Дашевский В.Г. Конформационный анализ органических молекул, М.Химия, 1982

102. Дуда, Р. и П.Харт. Распознавание образов и анализ сцен. Москва. Мир, 1976.

103. Коваленко И.Н., Филиппова А.А., Теория вероятностей и математическая статистика, М., "Высшая школа", 1973

104. Розенблит В.Е., Голендер, Логико-комбинаторные методы конструирования лекарств, 1983

105. Ту, Дж., Гонсалес Р., Принципы распознавания образов. М.Мир, 1978.

106. Тютюлков Н., Квантова Химия, Наука и изкуство, 1978

107. Шараф, М.А., Илмен Д.Л., Ковалски, Б.Р. Хемометрика, Ленинград, “Химия”, 1989

Обозначения и термини :

R^n - n размерно Евклидово пространство

x, A - черен шрифт се използува за обозначаване на вектори и матрици

$\boldsymbol{x}^t$ **-** транспониран вектор **x**

ω - състояние на природата (клас)

p(.) - плътност на вероятността

p(a,b) - съвместна плътност на вероятността, т.е. плътността на вероятността за едновременно появяване на *a* и *b*

*p(***x**|θ*)* - условната плътност на вероятността за **x** при условие θ

w - тегловен вектор

антагонист (antagonist)

Молекула, която се свързва към рецептора и блокира действието му.

agонист (agonist)

Молекула, която се свързва към рецептора и инициира действието му, т.е. предизвиква конформационни промени, позволяващи предаването на електрически (чрез йони) или механичен сигнал.

Фармакофор

тримерната структура на фрагмент, състоящ се от центрове и разстоянията между тях, които определят биологичните свойства на съединението.

SMILES

Simplified Molecular Input Line Specification - спецификация за обозначение на състава и свързаността на химично съединение със символен стринг.

Компютърни методи за моделиране на химични съединения.