ПЛОВДИВСКИ УНИВЕРСИТЕТ "ПАИСИЙ ХИЛЕНДАРСКИ"
КАТЕДРА "АНАЛИТИЧНА ХИМИЯ И КОМПЮТЪРНА ХИМИЯ"

ПЛАМЕН НИКОЛОВ ПЕНЧЕВ

КОМПЮТЪРНА ИНТЕРПРЕТАЦИЯ НА МОЛЕКУЛНИ СПЕКТРИ С ЦЕЛ РАЗКРИВАНЕ НА СТРУКТУРАТА НА ОРГАНИЧНИ СЪЕДИНЕНИЯ

А В Т О Р Е Ф Е Р А Т

ЗА ПРИСЪЖДАНЕ НА НАУЧНА СТЕПЕН
ДОКТОР НА НАУКИТЕ

ПЛОВДИВ
2016

Дисертантът работи като доцент в катедра „Аналитична химия и компютърна химия" при Химическия факултет на Пловдивски университет "П. Хилендарски", където е проведена основната част от изследванията. Част от работата е извършена в Техническия университет във Виена (1 г. и 3 м.) и Държавния университет на щата Аризона (4 г. и 6 м.).

Някои от изследванията включени в дисертацията, са финансирани от Фонд "Научни изследвания" – проекти Х-124/91, Х-447/94, ДДВУ02/37 (2011 г.), по Grant GM62457 by National Institutes of Health, USA, (2001 – 2003) и проект НИ13 ХФ 006 с финансиране на ПУ "П. Хилендарски". По темата на дисертацията са защитени 7 дипломни работи, със (съ)ръководител П. Пенчев.

Дисертацията съдържа 313 страници, 68 таблици, 53 фигури и 45 уравнения. Библиографската справка обхваща 815 статии, 60 монографии и книги, както и 19 програмни продукта и ръководства или Интернет страници. По дисертацията са отпечатани 24 публикации, от които 12 в списания с импакт фактор, 4 в международни специализирани научни списания и 3 в български списания, цитирани в списания с импакт фактор. Общият импакт фактор на статиите е 15.991. Върху тях са забелязани 103 цитата от чужди автори. Номерата на фигурите, таблиците и уравненията в автореферата съвпадат с тези от дисертацията. Оставена е номерацията на частите в дисертацията.

Защитата на дисертационния труд ще се състои на 19.9. 2016 г. от 13 ч. в XX аудиторията на ПУ, ул. "Цар Асен" № 24. Материалите по защитата са на разположение на интересуващите се в отдел „Развитие на академичния състав и докторантури" към ПУ „Паисиий Хилендарски" и Националния център за информация и документация към Министерството на образованието, младежта и науката.

ИЗПОЛЗВАНИ СЪКРАЩЕНИЯ И ОЗНАЧЕНИЯ В АВТОРЕФЕРАТА

- ИБТ – интерпретационно библиотечно търсене
- ИНМ - изкуствена невронна мрежа;
- ИЧ - инфрачервен;
- ЛДА - линеен дискриминантен анализ;
- МОП - максимална обща подструктура/и;
- 1D, 2D – кратки означения за едномерен и двумерен ЯМР;
- BADLIST - списък от подструктури, които не присъстват в генерираната структура;
- DEPT - Distortionless Enhancement by Polarization Transfer spectrum; ^{1}H-^{1}H COSY - proton-proton homonuclear correlation spectrum; HMBC - Long range ^{1}H-^{13}C Heteronuclear Multiple Bond Correlation experiment.
- FT – трансформация на Фурие;
- FT-IR – ИЧ (апарат) с трансформация на Фурие;
- GOODLIST - списък от подструктури, които присъстват в генерираната структура;
- HQI - хит качествен индекс; нормирана мярка за подобие между два спектъра;
- I_k или A_k - интензитет на k-тата ивица в спектъра (в a.u.);
- MACROATOM - списък от подструктури, които присъстват в генерираната структура;
- $N_{k,m}$ - брой на обектите от клас *k* класифицирани към клас *m*;
- P_k - точност на класификацията за обектите от клас *к*;
- R_k - степен на класификацията за обектите от клас *к*;

У В О Д

Разкриване на структурата на органичните съединения или тяхната идентификация е изключително сложна задача, която се решава в наши дни с използване на разнообразни инструментални методи за анализ: мас-спектрометрия, ЯМР, ИЧ и Раман спектроскопия. Надеждността и ефективността на интерпретация/идентификация зависят силно от квалификацията на колектива от химици-органици и спектроскописти в областите на различните методи. Като се отчете, че теоретичната основа на всеки един от различните спектрални методи е обширна и непрекъснато развиващата се област, става очевидно, че на ръководителя на изследването е трудно да овладее, прилага и контролира работата с всички методи, използвани при анализ на дадена проба. Много често, колективът от учени е от различни академични институции, а това затруднява ефективната комуникация между тях, при което субективният характер на вземаните от човека решения се мултиплицира и може да доведе до неотчитане на всички възможни структурни решения. Всичко това оправдава въвеждането на (полу)автоматизирани системи за поддържане, търсене и интерпретация на спектрална информация.

Наред с тези причини, възникналата потребност от експресни химически анализи в индустрията, свързването на спектралните апарати към мощни компютърни системи, както и способността за анализ на сложни смеси с помощта на методите на GC/IR и GC/MS доведоха до получаването на огромни по обем и сложни по структура аналитични данни, чиято обработка е немислима без използването на съвременни компютри. Електронно-изчислителната техника значително превъзхожда човешкия мозък в способността си да съхранява огромни масиви информация. Машините са способни да обработват данните по-бързо, по-точно, по-надеждно и по-евтино, нямат предубеждения и предпочитания и не се изморяват да извършват еднообразна дейност.

Тема и **основна цел** на настоящата дисертация е *разработването, прилагането, програмирането и едновременното използване на редица математически метода за интерпретацията на УВ-Вид, ИЧ, Раман, ^{13}С ЯМР и мас-спектри*. Разработената от нас комплексна система от програмни продукти включва методи за търсене в спектрални библиотеки, регресионен анализ на спектри, определяне на максимална обща структура и класификация на молекулни спектри с използването на няколко класификационни алгоритъма: линеен дискриминантен анализ, изкуствени невронни мрежи и методите на максимална обща подструктура и на най-близките съседи. Една част от методите са разработени, оптимизирани и тествани в докторската дисертация на автора, но по-голямата част от изследванията са нови.

Като резултат от работата на автора по дисертацията са създадени няколко спектрални библиотеки и три програми, работещи в среда на Windows. Всички разработени програмни продукти и спектрални библиотеки са със свободен достъп по Интернет и се използват, освен в научната работа, и за обучение на студенти бакалври и магистри по дисциплините, преподавани от автора.

РЕЗУЛТАТИ И ОБСЪЖДАНЕ

Резултатите от настоящите изследвания могат методически да се разделят на две части: (1) създаване на нови методи и (2) прилагане и оптимизиране на описани в литературата подходи. Междинно положение между тях заема програмирането на описани в литературата методи, с цел тяхното прилагане в аналитичната практика. По същество, програмната реализация е научно-приложно изследване за създаване на конкретни компютърни методи и затова то е включено в тази част. Научно-приложно изследване е и създаването на спектрални библиотеки от ИЧ, ATR, Раман и напълно отнесени ^{13}C ЯМР спектри, което е описано в глава от дисертацията СПЕКТРОСКОПСКИ ИЗМЕРВАНИЯ И СОФТУЕР.

В лабораторията по молекулна спектроскопия са измерени на апаратите **FT-IR Perkin-Elmer 1750** и **VERTEX 70 FT-IR** (Bruker Optics) стотици ИЧ, ATR и Раман спектри на органични съединения от автора, негови дипломанти и колеги. Голяма част от тези спектри беше използвана за създаване на спектрални библиотеки, а част от тях - за целите на експериментална проверка на програмираните алгоритми. Допълнително бяха преобразувани в спектрални библиотеки ИЧ, ATP, Раман и УВ-Вид спектри, измерени в други лаборатории.

Над 60 хиляди напълно отнесени ^{13}C-ЯМР спектри от спектралната колекция на Sadtler бяха обработени и редица от тях бяха премахнати – в резултат бе създадена библиотеката **LAST**. Други 1000 ^{13}C-ЯМР спектъра бяха набрани ръчно от статии на списанието Phytochemistry (2002-2006 г., т. 61-67). От тези спектри е съставена спектралната библиотека **PhyChem**. Допълнително, 500 ^{13}C-ЯМР спектъра бяха набрани ръчно от статии на списанието Phytochemistry (2001-2002 г., т. 57-60). Тези спектри се използват за тестови цели при интерпретационно търсене в библиотеките **LAST** и **PhyChem**.

В резултат на тази работа бяха създадени редица спектрални библиотеки, които са дадени в таблици 2.3.1 – 2.

Таблица 2.3.1. Брой на съединенията в ИЧ спектралните библиотеки, чиито спектри са измерени в нашата лаборатория.

Библиотека	Брой спектри	Брой нови спектри[1)]	Новите спектри са измерени на
IR01	105	0	-
IR02	181	20	VERTEX 70
IR03	197	53	Perkin-Elmer 1750
IR04	179	179	Perkin-Elmer 1750
IR05	52	52	VERTEX 70
IR06	197	197	Perkin-Elmer 1750
IRSubst	55	18	VERTEX 70
Общо	**966**	**519**	

[1)] Нови спектри означава спектри, измерени от автора след дисертацията за степен „Доктор".

За поддържане на тези спектрални библиотеки, както и за редица изчисления в дисертацията от автора бяха създадени няколко програми, които са описани подробно в глава II на дисертацията и кратко в следващия текст. Изключение прави програмната система **UVLib**, чийто код е написан от Димитър Христозов по дизайн на автора.

Таблица 2.3.2. Брой на съединенията, чиито спектри са включени в останалите вибрационни спектрални библиотеки.

Библиотека	Брой спектри	Вид спектри	Описание на спектрите
IR13484	13 484	ИЧ	Измерени на Bruker IFS 85
IR1000	1 000	ИЧ	Измерени на Bruker IFS 85
SR	200	ИЧ	Демо библиотека на Sadtler
Bruker	350	ИЧ	Демо библиотека на Bruker
Raman1	100	Раман	Измерени на RAM II
RaOpus1	100	Раман	Измерени на RAM II
Raman	200	Раман	Измерени на RAM II
RaPlants	116	Раман	Измерени на RFS-100
RaR	330	Раман	Измерени на RAM II и RFS-100
ATR	100	ATR	Измерени на VERTEX 70 с MIRacle

2.5.1. Програма за търсене в библиотеки от вибрационни спектри IRSS.

Новата версия на тази програмна система е създадена от автора на настоящата дисертация. Кодът е написан и компилиран на **Delphi 1**. **IRSS** е потребителски ориентирана програма (user friendly), която работи в среда на Windows. Функционалните възможности на програмата са описани в подробно в дисертацията и са като на всяка една програма за библиотечно търсене.

2.5.2. Програма за интерпретационно библиотечно търсене InferCNMR.

Тази програмна система е създадена от автора на настоящата дисертация. Кодът е написан и компилиран на **Delphi 2**. Тя е потребителски ориентирана програма, която работи в среда на Windows. Потребителят може да разглежда библиотеките от напълно отнесени спектри, както и резултатите от интерпретационното търсене. Програмата позволява експортиране на получените подструктури в текстов MOL файл, както и вмъкването на информацията за тях във файл, който се използва от структурния генератор **Houdini** на системата за разкриване на структурата на неизвестни съединения **Sesami**.

2.5.3. Програмата NeuNet.

Новата версия е програмирана на **Delphi 2** и освен обучение на ИНМ с право разпространение на сигналите и обратно разпространение на грешките, може да извършва анализ на главните компоненти. В настоящите изследвания тази програма е използвана за получаване на функциите на надеждност на подструктурите, получени при интерпретационното библиотечно търсене в библиотеки от напълно отнесени ^{13}C-ЯМР спектри – вижте т. 3.8.3.2.

2.5.4. Програмата UVLib.

Това е програма, работеща в среда на Windows, за търсене на УВ-Вид електронни спектри в спектрални библиотеки, кодът на която е написан на **Delphi 5**. Програмата поддържа спектрите, структурата на съединенията като 2D таблица на свързаност и химична информация за тях. Търсенето се извършва по спектрална крива и се използват четири алгоритъма, които се дават с уравнения (3.1.2) – (3.1.5).

3.1. Алгоритми за библиотечно търсене в ИЧ и Раман спектрални библиотеки.

В програмата **IRIS** бяха реализирани описани в литературата алгоритми за изчисляване на спектрално подобието между библиотечните спектри и спектъра на съединението, което се идентифицира.

3.1.2. Алгоритми за търсене по пикове в ИЧ и Раман спектри.

Алгоритмите за търсене по пикове, описани в литературата, се разделят на (a) *прави* (forward) - за идентификация на чисти вещества, и (b) *обратни* (reverse) - за разкриване на съставките на анализираната смес. Реализациите на тези алгоритми сме взаимствали от системата Sadtler, но сме извършили редица промени.

Ако непознатият спектър се разглежда като множество **U**, съдържащо **M** пика, а всеки библиотечен - като множество **R**, съдържащо **N** пика сечението между **U** и **R** е множество **I**, съдържащо **K** пика. Идеалното съвпадение между непознатия спектър и референтния ще е изпълнено при

$$\mathbf{U \equiv R \equiv I,\ т.е.\ M = N = K.}$$

В приложната спектроскопия точното съвпадане на пиковете на два спектъра е по-скоро изключение, отколкото правило, ето защо алгоритмите трябва да боравят с една неопределеност (tolerance) в положението на ивиците на референтния спектър, както по абсцисата (вълновото число), $\Delta\nu$, така и по ординатата (абсорбцията), $\Delta\mathbf{A}$.

$$\nu^{R} - \Delta\nu \le \nu^{U} \le \nu^{R} + \Delta\nu$$
$$\mathbf{A}^{R} - \Delta\mathbf{A} \le \mathbf{A}^{U} \le \mathbf{A}^{R} + \Delta\mathbf{A}$$

Мярката за спектрално подобие, реализирана в нашата система представлява трицифрено число **ABC**, всяка цифра от което се изчислява по независим начин. Първите две числа определят доколко сечението между двете пикови таблици е еднакво с пиковата таблица на непознатия спектър, цифрата **A**, или с тази на библиотечния, цифрата **B** (и двете закръглени до цели числа):

$$\mathbf{A = 9\ K\ /\ M} \quad \text{и} \quad \mathbf{B = 9\ K\ /\ N,}$$

Третото число определя доколко добре пиковете съвпадат по вълново число:

$$\mathbf{C = 9[1 - \Sigma|\nu_U^K - \nu_R^K|/(K\ \Delta\nu)]}$$

където ν_U^K и ν_R^K са положенията по абсцисата (вълновите числа) на k-тите съвпаднали пикове в непознатия и библиотечния спектър, а сумата се взима от 1 до **K**. Очевидно, максималната стойност на тази сума е $\mathbf{K} \times \Delta\nu$ и в този случай **C** е нула. При пълно съвпадение на пиковете по вълнови числа сумата е нула и тогава **C = 9**. (В системата Sadtler изчисляването на цифрата **C** не е описано.)

В програмната система **IRSS** по тази схема се изчисляват две мерки на подобие, наречени *хит качествен индекс* (hit-list quality index, HQI):

- при правия алгоритъм $\mathbf{HQI_F = ABC}$
- при обратния $\mathbf{HQI_R = BAC}$

Така изчислявани тези мерки притежават редица недостатъци. Един от тях е еднаквата неопределеност за всички пикове (особено по ординатата), която е

критикувана в литературата. Също така числото **B** силно зависи от броя ивици в библиотечния спектър. Това води до появата на спектри с по-малък брой пикове **N** в началото на хит-списъка. За избягване на тези недостатъци беше дефиниран и приложен **HQI**, изчисляван по уравнение (3.1.1), наречен *скаларно произведение на пикови таблици*:

$$HQI_P = 999 \frac{\sum_k A_k^U A_k^R}{\left\|A^U\right\| \cdot \left\|A^R\right\|} \qquad (3.1.1)$$

където скаларното произведение в числителя се извършва само за съвпадащите пикове, а в знаменателя на HQI_P стоят съответните норми на пиковите таблици.

3.1.3. Алгоритми за търсене по спектрална крива.

В програмата са заложени четири различни алгоритми за търсене по спектрална крива в зависимост от използваните мерки за спектрално подобие между спектрите: (1) *средно квадратично отклонение*, (2) *средно абсолютно отклонение*, (3) *скаларно произведение*, и (4) *коефициент на корелация*. Ако с A^U_K означим абсорбцията при *k*-тото вълново число в непознатия спектър, а с A^R_K - тази в референтния, то съответните хит-качествени индекси са пропорционални на:

- средното квадратичното отклонение между спектрите:

$$S_1 = \sqrt{\sum_k (A_k^U - A_k^R)^2 / N} \qquad (3.1.2)$$

- средното абсолютно отклонение между спектрите:

$$S_2 = (1/N) \sum_k \left| A_k^U - A_k^R \right| \qquad (3.1.3)$$

- скаларното произведение между спектрите:

$$S_3 = \frac{\sum_k A_k^U A_k^R}{\left|A^U\right| \cdot \left|A^R\right|} \qquad (3.1.4)$$

- коефициента на корелация между спектрите:

$$S_4 = \frac{\sum_k (A_k^U - \overline{A^U})(A_k^R - \overline{A^R})}{\sqrt{\sum_k (A_k^U - \overline{A^U})^2 * \sum_k (A_k^R - \overline{A^R})^2}}, \qquad (3.1.5)$$

където $|A_U|$ и $|A_R|$ са големините на векторите (спектрите), а **N** е броят на разглежданите абсорбционни стойности.

Мерките S_1 и S_2 показват различието между спектрите. Тяхната максимална теоретична стойност е 1.0 при условие, че спектрите са нормирани в интервала 0.0 - 1.0 a.u. Те заемат минималната си стойност 0.0 при напълно идентични спектри. В програмата те са преобразувани в **HQI** по формулата:

$$HQI_k = 999\ (1 - S_k);\ k = 1, 2$$

S_3 и S_4 показват подобието между спектрите, като тяхната максимална стойност е 1.0 за напълно идентични спектри. Минималната стойност на S_3 е 0.0 (за напълно ортогонални спектри) поради това, че стойностите на абсорбцията са по-големи или равни на нула. Минималната стойност на S_4 е -1.0 за отрицателно

корелирани спектри. Ето защо съответните **HQI** се изчисляват от s_3 и s_4 чрез нормирането им в интервала 0.0 - 1.0 и последващо умножение с 999. Всички тези алгоритми са препрограмирани от автора в софтуера **IRSS** за поддържане на, и търсене в, библиотеки от ИЧ, ATR и Раман спектри.

3.2. Експериментална проверка на алгоритмите за търсене в библиотеки от ИЧ спектри.

3.2.1. Търсене по ивици (пикове) в ИЧ спектрите.

Предложените в т. 3.1.2 мерки за подобие при търсенето по пикове зависят от много параметри, най-важните от които са неопределеностите по вълновото число $\Delta\nu$ и по ординатата ΔA. (Точно те определят кои пикове ще съвпаднат.) Стойностите на праговете (*thresholds*) - t_R и t_U, използвани при алгоритъма за намиране на пиковете в библиотечния и изследвания спектър също влияят, макар и неявно, на изчисляваните **HQI**. Други параметри от по-общ характер, които също определят неявно стойностите на **HQI**, са начинът на регистриране на спектъра, методът, използван при изглаждане на спектрите, както и параметрите на базовата линия при корекция на последната.

Таблица 3.2.1. Позиция на идентифицираното съединение в хит-списъка за спектралните мерки: средно квадратичното отклонение (с.к.о.); средно абсолютно отклонение (с.а.о.), скаларно произведение (с.п.) и коефициент на корелация (к.к.). Хит-списъкът е от 200 спектъра (съкратен вариант – премахнати са 9 съединения с 1 1 1 1).

химично име	с.к.о.	с.а.о.	с.п.	к.к.
2-аминобензофенон	-*	-	1	1
2-аминоизобутанова киселина	1	7	1	1
2-нитробензалдехид	-	-	2	2
(3,4-диметоксифенил)оцетна киселина	1	2	1	1
4-бромбензалдехид	-	-	1	1
4-хлорбензалдехид	-	-	1	1
4'-метоксиацетофенон	4	92	1	1
4-нитробензалдехид	-	-	1	1
4-ацетилбифенил	1	2	1	1
валерофенон	1	3	1	1
2-метилциклохексанон	1	2	1	1
1-бензил-4-пиперидон	-	-	1	1
алфа-йонон	-	-	1	2
циклохептанон	57	-	3	3
3-аминобензоена киселина	-	-	1	1
кумарин	-	-	1	6
антрацен	2	8	12	1
2-фенил-1,3-индандион	2	21	1	1
цис-5-норборнен-ендо-2,3-дикарбокси анхидрид	1	2	1	1
диметилизофталат	4	6	1	1
4-хидроксибензалдехид	15	-	1	1

* не е сред първите 200 хита.

При спазване на стандартна методика за измерване и обработване на ИЧ спектри последните параметри почти не оказват влияние на изчисляваните мерки за подобие. Ето защо идентификацията на непознато вещество по пикова таблица

беше оптимизирана по стойностите на ΔA и $\Delta\nu$, а $t_R = t_U = 0.03$ a.u. Надеждността на спектралните мерки се определяше от броя на правилно идентифицираните съединения, т.е. тези които се появяват на първо място в хит-списъка.

За провеждане на тестовете бяха избрани тридесет спектъра от нашите библиотеки **IR01**, **IR02** и **IR03**: те са избирани по десет от библиотека. Тези спектри бяха разглеждани като спектри на непознати вещества, снимани в друга лаборатория, които се идентифицират с помощта на библиотеката **IR13484**, съставена от 13 484 спектъра, т.е. изследвана е задачата за идентификация на вещества по техните ИЧ спектри. Имената на тези съединения са дадени в таблица 3.2.1, заедно с позициите в списъка с резултати (хит-списъка) при търсене по спектрална крива – това са резултати за следващата т. 3.2.3.

За определяне на оптималните стойности на неопределеностите ΔA и $\Delta\nu$, те бяха променяни в интервалите

$$\Delta\nu = 3, 4, 5, \ldots 40 \text{ cm}^{-1} \text{ и } \Delta A = 0.1, 0.2, \ldots 1.0 \text{ a.u.}$$

Бяха използвани и трите метода за пиково търсене, описани в т. 3.1.2. – **прав, обратен и скаларно произведение.**

За получаване на количествени изводи, позицията на идентифицираното съединение в съответния хит е осреднена по 30-те съединения и е представена част от всички данни (за $\Delta\nu = 3 - 8$ cm^{-1}), отделно за всеки един от използваните методи в таблица 3.2.2 в дисертацията.

Средна позиция 1.0 за даден метод за търсене в таблица 3.2.2 означава, че за тези стойности на ΔA и $\Delta\nu$ методът дава всички 30 съединения от таблица 3.2.1 на първо място в хит-списъка (ХС). Стойност 1.03 означава, че точно едно съединение от тези 30 е второ в ХС, докато останилите 29 са идентифицирани на първо място в ХС. Стойност 1.07 има двузначно значение – или точно две съединения са на второ място в ХС, или точно едно на трето място в ХС, а останалите – на първо място.

Най-добре за идентификация работи правия метод за търсене по пикове – в широк диапазон на неопределеностите той дава неизвестното съединение като първи хит. Както и се очаква при увеличаване на $\Delta\nu$ се губи селективността на метода, а при най-малката стойност на $\Delta\nu = 3$ cm^{-1} се намалява разпознавателната (идентифициращата) способност, защото нарастват несъвпадналите пикове. При обратния алгоритъм това е ключово обяснение – ниският брой на съвпаднали пикове, **K**, оказва по правило по-голямо влияние на числото **B** от т. 3.1.2, отколкото на числото **A**, и съответно на хит-качествения индекс от вида **BAC**, отколкото на **ABC**. Причина за това е, че **HQI = BAC** има по-висока стойност за съединенията от библиотеката с малък брой пикове **N**, а при ниско **N**, **B = 9K/N** по-силно се влияе от намаляването на **K**, отколкото **A = 9K/M**.

Скаларното произведение на пикове работи зле само за три от тези 30 съединения: и в трите случая „непознатия" и референтния спектър имат значително различен брой пикове. Читателят може да се запознае със сравнението между „непознатия" и библиотечния спектър за три двойки спектри, които са дадени на фигура 3.2.2 в дисертацията.

За оптимални комбинации на ΔA и $\Delta\nu$ бяха избрани:
за право търсене: $\Delta A = 0.4 - 0.9$ a.u. и $\Delta\nu = 3 - 7$ см$^{-1}$

за обратно търсене: ΔA = 0.4 - 0.8 a.u. и $\Delta\nu$ = 3 - 7 $см^{-1}$

За третия алгоритъм ако се пренебрегнат три от съединенията се получават следните оптимални интервали за неопределеностите:

ΔA = 0.4 - 1.0 a.u. и $\Delta\nu$ = 4 - 7 $см^{-1}$

Почти еднаквите области на неопределеност за първите два алгоритъма са следствие от постановката на задачата - идентификация на чисти вещества. В този случай и двата алгоритъма работят еднакво добре, тъй като чистото вещество е частен случай на смес. По-лошото представяне на третия алгоритъм (в някои от случаите) за идентификация на чисти вещества не е проблем, тъй като той е въведен за търсене на спектри на смеси.

3.2.2. Търсене по ИЧ спектрална крива.

В предварителните изследвания бяха проведени две серии тестове с използването на (a) 27 спектъра от библиотеката **SR** на **Sadtler IR Search** и (b) 48 спектъра, измерени по стандартна методика в нашата лаборатория. Тези 75 спектъра бяха идентифицирани с помощта на библиотеката **Pl-Uni**, съдържаща спектри на всичките 75 вещества. Спектрите от серия (a) представляват спектри, измерени на друг апарат и в друга лаборатория, а тези от (b) - спектри, измерени на същия апарат. Резултатите от проведените търсения са систематизирани в таблица 3.2.3, където са представени броят и процентът на успешно идентифицираните вещества с всеки един от четирите метода. От таблицата се вижда, че спектрите, които са измерени на същия апарат се идентифицират (средно с 11%) по-успешно, отколкото спектрите от други лаборатории. Едно детайлно сравнение на спектралните криви на спектрите от библиотеката **SR** с тези от **Pl-Uni** показва несъответствие в положението на пиковете от порядъка на 3 до 6 $см^{-1}$. Това е общ проблем за всички системи за библиотечно търсене, който се преодолява успешно с използването на неопределеност по абсцисата при търсенето по ивици или с използването на методите на коефициента на корелация и този на скаларното произведение при търсене по спектрална крива.

Таблица 3.2.3. Брой и процент на успешно идентифицираните съединения от серии (a) и (b) с алгоритмите за търсене по спектрална крива: средно квадратичното отклонение (с.к.о.); средно абсолютно отклонение (с.а.о.), скаларно произведение (с.п.) и коефициент на корелация (к.к.).

	серия (a)				серия (b)			
методи	с.к.о.	с.а.о.	с.п.	к.к.	с.к.о.	с.а.о.	с.п.	к.к.
брой	21	16	22	23	42	35	44	46
%	78	59	82	85	88	73	92	96

Друга закономерност, която се наблюдава в таблицата, е подредбата по надеждност на методите в реда: к.к. > с.п. > с.к.о. > с.а.о. За изследването на тази подредба бяха разгледани спектралните криви на всички неуспешно идентифицирани спектри и сравнени с библиотечните. В дисертацията подробно са анализирани причините за тези лоши резултати.

В заключение, резултатите от проведените тестове за идентификация на органични съединения с помощта на търсене в библиотеки от ИЧ спектри могат да се обобщят по следния начин:

(a) когато се спазват изискванията за равенство на максималната ивица на спектъра на 10% T, методите за търсене по спектрална крива не се затрудняват при идентификация на изследваното съединение.
(b) при отклонение от това условие или при търсене на спектър, заснет в друга лаборатория (в сравнение с библиотечните), методът на коефициента на корелация на спектрални криви дава най-добри резултати, последван от този на скаларното произведение на спектрални криви.
(c) наличието на дълги хомоложни серии или на дълга поредица от съединения с много близка структура затруднява идентификацията на изследвания спектър.
(d) методите за търсене по пикове дават по-добри резултати, а и са по-бързи в сравнение с методите, използващи спектрална крива. По добрите резултати се дължат на неопределеността в положението на пиковете, задавана от потребителя.

3.3. Експериментална проверка на алгоритмите за търсене в библиотеки от Раман спектри.

Беше извършена основна проверка на приложимостта на алгоритмите за търсене на ИЧ спектри в ИЧ спектрална библиотека към търсене на Раман спектри в Раман библиотеки на 100 органични съединения - **Raman1**, която се поддържа от програмата **IRSS**, и **RaOpus1**, която се поддържа от софтуера **OPUS** на апаратите **VERTEX 70** и **RAM II**. Въпреки съществената разлика между ИЧ и Раман спектрите е известно, че Раман спектърът отразява в детайли структурата на съединението и затова се очаква идентификацията на веществата посредством техните Раман спектри да е възможна и надеждна.

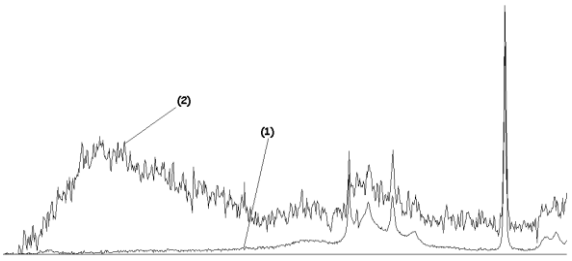
Създадените две спектрални библиотеки, **Raman1** и **RaOpus1**, позволяват сравняването на алгоритмите за търсене в спектрална библиотека, т.е. сравнението между алгоритмите програмирани в **IRSS** и тези в софтуера **OPUS**. За целта са заснети Раман спектрите на 29 съединения (т.н. *тестови спектри*) при мощност на лазера, различна от тази на съответните библиотечни спектри: те са дадени в таблица 3.3.1 в дисертацията. Това дава различно отношение сигнал/шум, което е съществено за тестовете. Търсенията на 29 от спектрите дават следните резултати:

a) 28 от тях се намират на първо място в списъка с резултати, генериран както с програмата **IRSS**, така и със софтуера на апарата.

b) Спектърът на **3-Тиофен-3-илметилен-3H-бензо[de]изхромен-1-он** при търсене с **IRSS** се появява на 5 позиция, докато софтуерът на апарата не го намира сред първите тридесет хита.

Три съединения, които имат спектри в Раман спектралните библиотеки, са заснети на Раман микроскопа като са използвани малки кристалчета от тях. Техните структури и мощността на лазера са дадени в таблица 3.3.2 от дисертацията. Целта на тези тестове е да се провери способността на алгоритмите за идентификация, когато са приложени към спектри с изключително лошо отношение сигнал/шум, каквото се получава при измерване с микроскоп. От трите измерени спектри на микроскопа, този на оксаловата киселина, даден на фигура 3.3.1, показва най-лошо отношение сигнал шум. Докато

спектрите на първите две съединения се идентифицират успешно с двата софтуера, но *стандартният алгоритъм* за библиотечно търсене в софтуера **OPUS** не успява да идентифицира третото съединение. При търсене с програмата **IRSS** оксаловата киселина излиза като първи хит.



Фигура 3.3.1. Раман спектрите на оксаловата киселина (съединение 3 от таблица 3.1.5). Означения:
(1) – от библиотеката;
(2) - измерен на Раман микроскопа.

Тези резултати, въпреки че са получени с малки по обем спектрални библиотеки – 100 спектъра, показват опасността от работа със софтуер, за алгоритмите на който няма подробно описание и се използва „сляпо".

3.4. Математически анализ на вибрационни спектри на смеси.

3.4.1. Теория

Спектърът на една смес ($\mathbf{M}_{1,K}$) може да се запише *при определени условия* като линейна комбинация от спектрите на отделните компоненти ($\mathbf{S}_{N,K}$):

$$\mathbf{M}_{1,K} = \mathbf{C}_{1,N}\,\mathbf{S}_{N,K}\,, \qquad (3.4.1)$$

където индексите в това матрично уравнение (както и в следващите) означават размерностите на съответните матрици. Тези необходими условия са: (1) концентрациите на съставките (компонентите) се намират в линейната работна област на зависимостта абсорбция-концентрация; (2) няма взаимодействие между компонентите в сместа, т.е. спектърът на сместа е линейна комбинация от техните спектри; (3) извършено е коригиране на базовата линия на спектъра на сместа и на спектрите на съставките в спектралната библиотека.

При дължина на оптическия път, равна на единица, коефициентите $\mathbf{C}_{1,N}$ са концентрациите на $\mathbf{N}^{\text{те}}$ компонента и могат да бъдат намерени при условие, че броят на спектралните признаци **K** е по-голям или равен на **N** и рангът на матрицата $\mathbf{S}_{N,K}$ е равен на **N**. Чрез умножението на уравнение (3.4.1) с обобщената обратна матрица на $\mathbf{S}_{N,K}$, $\mathbf{S}_{K,N}^{T}\ (\mathbf{S}_{N,K}\ \mathbf{S}_{K,N}^{T})^{-1}$, се получава (с $\mathbf{E}_{N,N}$ е означена единична матрица с размерност **N**):

$$\mathbf{M}_{1,K}\,\mathbf{S}_{K,N}^{T}\ (\mathbf{S}_{N,K}\ \mathbf{S}_{K,N}^{T})^{-1} = \mathbf{C}_{1,N}(\mathbf{S}_{N,K}\mathbf{S}_{K,N}^{T})\,(\mathbf{S}_{N,K}\mathbf{S}_{K,N}^{T})^{-1} = \mathbf{C}_{1,N}\,\mathbf{E}_{N,N} = \mathbf{C}_{1,N},$$

т.е.

$$\mathbf{C}_{1,N} = \mathbf{M}_{1,K}\ \mathbf{S}_{K,N}^{T}\ (\mathbf{S}_{N,K}\mathbf{S}_{K,N}^{T})^{-1} \qquad (3.4.2)$$

3.4.2. Идентификация на компонентите на смеси

Както беше споменато в литературния обзор, регресионното уравнение (3.4.2) е използвано при анализ на смеси. Основен техен недостатък е получаването на статистически различни от нула коефициенти за съставки, които не присъстват в сместа за сметка на отрицателните стойности на други коефициенти.

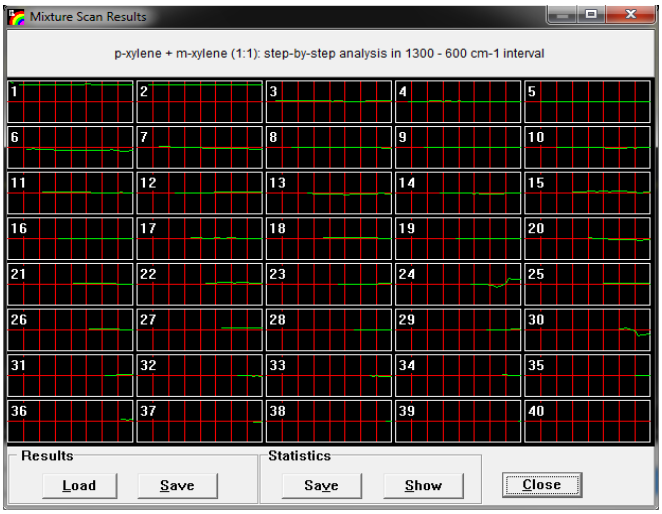
За избягване на този недостатък ние прилагаме поредица от регресионни изчисления на концентрациите по уравнение (3.4.2) с нарастващ брой на спектрите от хит-списъка. Първо се изчисляват коефициентите на участие на първия

спектър от хит-списъка по уравнение (3.4.2), после на първи и втори, след това тези на първи, втори и трети и т.н. до достигане на избрано от потребителя число на компонентите или получаването на линейна зависимост между използваните спектри от хит-списъка. Получава се една поредица от коефициенти (концентрации) c_k^m

$$\begin{array}{l} C_1^1,\ 0,\ \ 0,\ \ 0,\ \ldots.\ 0,\ \ 0 \\ C_1^2,\ C_2^2,\ 0,\ \ 0,\ \ldots.\ 0,\ \ 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ C_1^N,\ C_2^N,\ C_3^N,\ C_4^N,\ \ldots\ C_{N-1}^N,\ C_N^N, \end{array} \qquad (3.4.3)$$

в която горният индекс съответства на номера на изчислението, а долният - на този на спектъра. Нулите от горната част на диагонала се получават от постановката на задачата - допускането, че сместа се състои само от първите няколко компонента. Със c_k^m ще изначаваме псевдо-концентрацията на k-тото вещество в m-тата крива.

Построените графики на зависимостта на концентрациите c_k^m от броя на разглежданите компоненти **k** показват в повечето случаи стабилност само за компонентите, действително намиращи се в сместа. Под стабилност тук разбираме, изчисляваната концентрация да се променя слабо при нарастване на броя на спектрите от хит-списъка, които са включени в изчисленията. На фигура 3.4.1 са показани резултатите от изчисленията по уравнение (3.4.2) със спектрите от хит-списък, получен при търсене на ИЧ спектър на смес от **мета-ксилен** и **пара-ксилен** в съотношение 1:1 о.ч. в библиотеките **IR01-IR06** от общо 911 спектъра. Използвано е обратно търсене по пикове с параметри ΔA= 1.00 a.u., $\Delta\nu$ = 7 cm^{-1}, t_U = 0.01 a.u. и t_R = 0.03 a.u. Регресионният анализ е извършен с 40 спектъра от хит-списъка и 151 стойности на абсорбцията в интервала 1200 - 600 cm^{-1} за всички спектри. Хоризонталните линии в средата на малките прозорци са за стойностите c_k^m = 0.0. Прегледът на първите двадесет криви показва, че само криви #1 (**мета-ксилен**) и #2 (**пара-ксилен**) са приблизително прави линии, представляващи положителни стойности на "концентрацията". От коефициентите c_k^m могат да се изчислят средна стойност, $\bar{c}$, стандартно отклонение, **s**, и относително стандартно отклонение (r.s.d.), като се вземат само стойностите на c_k^m, без нулевите стойности, дефинирани *a priori*.



Фигура 3.4.1. Криви на зависимост на коефициен-тите, изчислени при регресионния анализ на първите 40 спектри от хит-списък, получен при търсене по пикове (обратен метод) на спектъра на смес от **мета-ксилен** и **пара-ксилен** в съотношение 1:1 о. ч. в библиотеките **IR01-IR06**. Графиките са дадени така, както излизат на екрана на програмата **IRSS**.

В таблица 3.4.1a са дадени тези стойности за веществата от хит-списъка, а в таблица 3.4.1b – резултатите за съставките на сместа и за съединението, което не присъства в сместа и има най-малко r.s.d. при положителна средна стойност на c_m^N и е сред първите 20 хита. Опитът от проведените изчисления сочи, че компонентите, които се съдържат в сместа *почти винаги* имат най-малки r.s.d.

Средните стойности, $\bar{c}$, **не показват** точната концентрация на компонентите, защото библиотечните спектри са заснети с различно и недокументирано количество вещество за твърдите проби, както и с различна и неизмерена дебелина на слоя за течните вещества. От уравнение (3.4.2) се вижда, че коефициентите c_m^N са безразмерни величини. Техните стойности нямат физически смисъл поради изложените по-горе причини и опитът показва, че те *не съвпадат* с нито едно изразяване на концентрацията на компонентите, но отношенията между тях са в добра корелация с обемните отношения на компонентите. Ето защо е по-правилно те да се наричат *псевдо-концентрации*. Изчисленото стандартно отклонение няма смисъла на това на серия от независими измервания, и поради тази причина не може да се използва за изчисляване на интервалната оценка на стойността на $\bar{c}_k$ по общоизвестните формули. В настоящата работа се разглежда единствено стабилността на получаваните криви c_m^N, която може да се оцени визуално от потребителя на програмата или количествено чрез стойностите на относителното стандартно отклонение. От тези оценки, според нас, може да се направят обективни заключения за качествения състав на сместа.

Таблица 3.4.1. Данни от регресионния анализ на първите 40 спектри от хит-списък, получен при търсене по пикове (обратен метод) на спектъра на смес от **мета-ксилен** и **пара-ксилен** в съотношение 1:1 о. ч. в библиотеките **IR01-IR06**. Означения: HL# - номер на съответния спектър в хит-списъка; **c** – псевдо-концентрация; **s** - стандартно отклонение; r.s.d. относително стандартно отклонение.

a) За първите девет хита, така както те се записват във файл с резултати. Трите звезди показват, че средните псевдо-концентрации са отрицателни.

```
p-xylene + m-xylene (1:1)
Number of Spectra:  40
H i t ' s   i d e n t i f i c a t i o n                    m e a n     st.dev.  r.s.d.
Hit # 1; HQI = 948; Sp #: IR05 3 m-Xylene                  0.976       0.033     3.4
Hit # 2; HQI = 838; Sp #: IR05 5 p-Xylene                  0.971       0.008     0.9
Hit # 3; HQI = 647; Sp #: IR03 183 3-Methylbenzyl cyanide  0.043       0.013    29.1
Hit # 4; HQI = 614; Sp #: IR06 30 Tetrachloroethylene      0.020       0.010    47.6
Hit # 5; HQI = 536; Sp #: IR03 87 Diphenylamine           -0.008 ***   0.012   149.4
Hit # 6; HQI = 525; Sp #: IR02 144 Citronellal            -0.164 ***   0.037    22.7
Hit # 7; HQI = 525; Sp #: IR04 83 1-Nitronaphthalene      -0.032 ***   0.054   166.3
Hit # 8; HQI = 515; Sp #: IR04 15 Trifluoroacetic acid    -0.011 ***   0.011    99.4
Hit # 9; HQI = 436; Sp #: IR01 10 (3,4-Dimethoxyphenyl)aceti  0.003       0.004   157.6
```

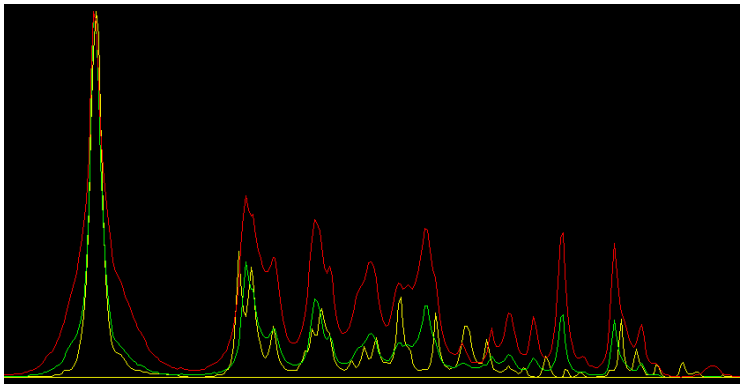
b) За съставките на сместта и съединението, което не присъства в сместа и има най-малка стойност на r.s.d.

Съединение	HL#	HQI	c	s	r.s.d. %
мета-ксилен	1	948	0.976	0.033	3.4
пара-ксилен	2	838	0.971	0.008	0.9
3-метилбензил цианид	3	746	0.043	0.013	29.1

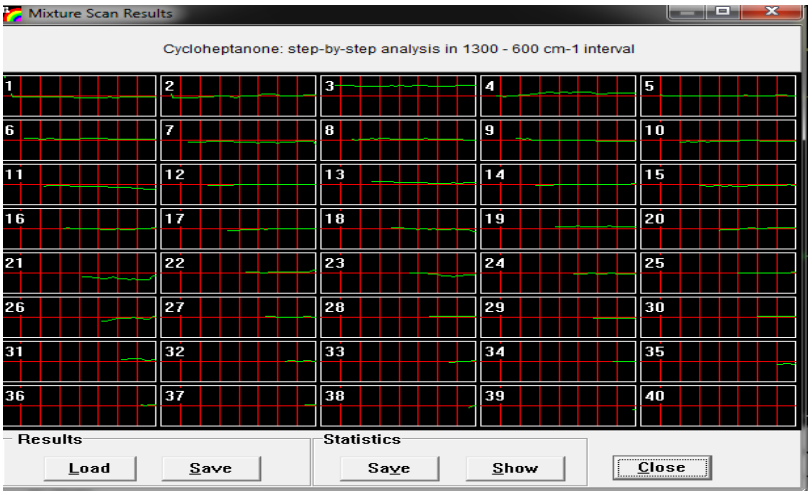
3.4.3. Идентификация на химични съединения.

Разработеният подход може успешно да се използва за идентификация на единично съединение. В този случай задачата се решава от изследователя при условие, че анализираното вещество се състои само от едно съединение.

Както беше споменато по-горе наличието на дълги хомоложни серии или поредици от близки по структура вещества затруднява идентификацията на веществата. Много често спектрите от получения хит-списък са значително подобни един от друг. Така например, спектърът на **циклохептанон**, съдържащ се в библиотеката **IR02** не се появява като първи хит при търсене в голямата библиотека **IR13484** - вижте таблица 3.2.1. Като първи хит при търсене по коефициент на корелация се появява спектърът на **циклооктанон** и причината е в библиотечния спектър в **IR13484**, заснет със значителна дебелина на тънкия слой. На фигура 3.2.2c са сравнени двата спектъра на **циклохептанона**, а на фигура 3.4.2 са показани тези два спектъра и спектърът на първия хит (**циклооктанон**) в спектралния интервал 1850-650 cm^{-1}. Пробемът може да се реши и със визуално сравнение на спектрите, но това е времеемка и уморителна процедура за спектроскописта, затова ако се приложи регресионното уравнение по схемата, описана в т. 3.4.1, последният ще бъде значително улеснен.



Фигура 3.4.2. Спектрите на **циклохептанон**: зеленият - от нашата библиотека **IR02**, и червеният - от **IR13484**, както и на първия хит (жълтият), **циклооктанон**, от **IR13484**.



Фигура 3.4.3. Резултати от регресионния анализ на първите 40 спектри от хит-списък, получен при търсене по коефициент на корелация на ИЧ спектър на **циклохептанон** в библиотеката **IR13484**, така както се показват от програмата **IRSS**.

При прилагане на регресионния метод върху първите 40 спектри от хит-списък, получен при търсене по коефициент на корелация и последващо

прилагане на серията регресионни уравнения дават резултатите, показани на фигура 3.4.3 и таблица 3.4.2, със най-стабилна крива за неизвестното съединение.

3.4.4. Експериментална проверка на алгоритъма за анализ на смеси.

В докторската дисертация на автора, анализът на смеси беше проверен с поредица от приготвени смеси в различни обемни отношения и с търсене в библиотеката **PL-Uni** от 608 ИЧ спектъра. Тези смеси сега са потърсени в библиотеките **IR01 - IR06** от общо 911 спектъра и в комбинация от тях и голямата библиотека от 13 484 спектъра (общо 14 395 спектъра). Целта е да се провери дали по-голямото структурно разнообразие, т.е. наличието на повече съединения, подобни по структура на съставките, не пречи на алгоритъма. Допълнително са заснети и спектрите на нови смеси, с които алгоритъмът е тестван с тези две комбинации от библиотечни спектри.

За всички търсения, ако не е изрично упоменато се използва обратно търсене по пикове с параметри ΔA= 1.00 a.u., $\Delta\nu$ = 10 cm^{-1}, t_U = 0.01 a.u. и t_R = 0.03 a.u. Първият толеранс е с максималната си стойност, тъй като съставката в сместа с по-малка концентрация дава пикове с по-малка абсорбция - по същата причина и прагът при извличане на пикове, t_U, е с по-малка стойност. Поредица регресионни изчисления се провежда в интервала 1300 - 600 cm^{-1}, освен ако не е споменат изрично друг спектрален интервал. Методиката, която препоръчваме за този анализ на смеси от две вещества, и която е приложена за тези изследвания, е следната: *след библиотечното търсене с горните параметри, се извършва регресионен анализ с първите четиридесет спектъра от хит-списъка, но се разглеждат само кривите на първите двадесет, тъй като кривите се скъсяват и стават безполезни. Първите две съединения с най-ниски r.s.d. и положителни средни псевдо-концентрации се приемат за съставките на бинерната смес*. Тук трябва да се отбележи, че наличието на два или повече спектъра на едно и също съединение в хит-списъка водят до понижаване на ранга на матрицата C_m^N и различно „смесване" на псевдо-концентрациите на двата спектъра в различните изчисления, а от там до неверни резултати за това съединение, затова вторият или третият и т.н. спектри се изтриват преди извършване на регресионния анализ.

В таблица 3.4.2 са дадени резултатите за петте смеси на **3-хептанон** и **1-хептанон**, когато са потърсени в 911 спектъра, а в таблица 3.4.3 – търсенето във втората комбинация от библиотеки. Преди да сравним тези пет двойки резултати е важно да се отбележи, че даже при търсенето в само $911^{те}$ спектъра девет от първите десет съединения в хит-списъка са много подобни (или, разбира се, еднакви) на съставките в сместта: това са (1) **3-хептанон**, (2) **3-нонанон**, (3) **2-метил-3-хептанон**, (4) **4-хептанон**, (5) **3-додеканон**, (6) **4-Decanone**, (7) **3-деканон**, (8) **5-додеканон**, (9) **дибутиламин**, (10) **3-тридеканон**. Въобще, анализът на тези смеси със стандартния алгоритъм за изваждане на спектри (чието приложение е описано в т. 3.4.5) ще бъде затруднен, поради значителното структурно подобие на двете съставки, което води до значително съвпадение на ивици им в техните ИЧ спектри.

Сравняването на търсенето в двата набора спектри (таблици 3.4.2-3 в дисертацията) показва, че двете съставки се идентифицират успешно, т.е. те са първите два хита, с най-ниско r.s.d. Съществената разлика е, че съставката с по-

малка концентрация излиза значително по-назад в списъка с резултати, когато търсенето е с повече библиотечни спектри. Това е резултат, който можеше да се предвиди, защото се очаква, че ивиците на основния компонент ще намаляват селективността на библиотечното търсене. Това не е проблем на самия регресионен метод, а на метода на библиотечно търсене.

За другата поредица от смеси, тази от **хексан** и **циклохексан**, споменатият проблем се задълбочава и съставката с по-малка концентрация въобще не е сред първите 50 съединения на хит-списъка. Единствено търсенето на сместа 1:1 о.ч. дава двете съставки сред първите 15 хита.

Поредица регресионни изчисления, може да се проведе в различен спектрален интервал и тогава по-правило r.s.d. на съставките запазва ниски стойности (но се получават различни r.s.d.), а съединенията, които не са съставки на сместа показват по-високи стойности на r.s.d. В таблица 3.4.5c от дисертацията са дадени резултатите на анилиз с използване на интервала 1300 - 500 cm^{-1} – вижда се, че r.s.d. на **1-пропанол,2-(циклохексил)** е нарастнало от 10.6% на 16.9%, докато промените на r.s.d. за съставките са много по-малки (за едната намалява, а за другата леко нараства). Определено може да се твърди, че ако се използват двата метода в комбинация (този за еднократно изчисление и този с поредица от изчисления) и с промяна на спектралния интервал, то идентификацията на компонентите на смеси става по-сигурна – това се вижда с този пример, ако се анализират едновременно резултатите от таблици 3.4.4-5.

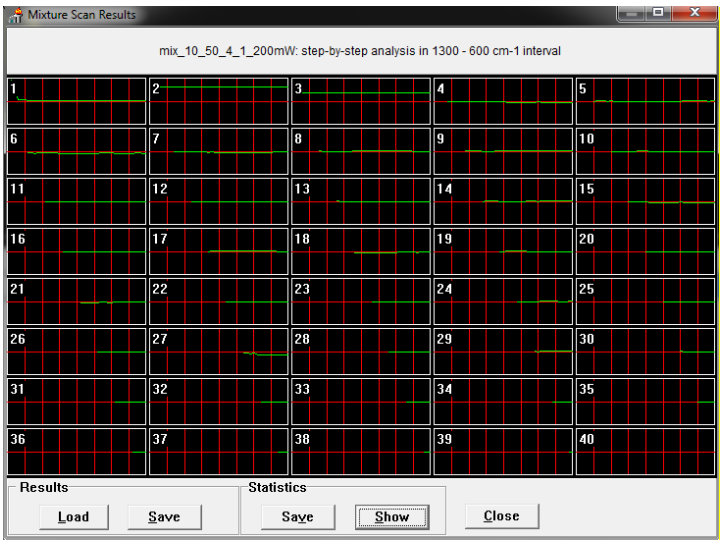
За другите приготвени смеси на **хексан** и **циклохексан,** с 1:4, 1:9, 4:1 и 9:1 о.ч., библиотечното търсене даже и в 911^{те} спектъра на библиотеките **IR01-IR06** пропуска компонента с по-малка концентрация сред първите 20 хита, в някои случаи и сред първите 50 хита. Това бяха двете серии смеси, които в докторската дисертация на автора бяха използвани за развитие на метода и неговото тестване: с използването на библиотека от 608 спектъра съставките на смесите и в двете серии бяха успешно идентифицирани, но сега с използване на 911 спектъра втората серия не може да се анализира и то не по причини на метода, а защото спектърът на съставката с по-малка концентрация не се появява в хит-списъка. С използване на значително повече спектри резултатите са даже по-лоши.

Трябва да споменем, че в литературата се докладва за идентифициране на компонентите на тройна смес от **хексан**, **циклохексан** и **толуен**, но повтаряне на неговите изчисления (вижте израза на стр. 73) даже и за бинарната смес от **хексан** и **циклохексан** показва, че методът не е приложим на практика и равните участъци, които трябва да се получат, не могат лесно да се отличат от другите спектрални интервали. Nyden също разглежда смес от **хексан** и **циклохексан** и неговият метод по същество е регресионен (но с ортонормализирани спектри) и работи толкова добре, колкото и нашия, при условие че съставките са в хит-списъка (той използва цялата библиотека, но тя е малко по обем), затова следващите серии смеси, които бяха приготвени са от тези съединения, както и от други алкилзаместени бензени – вижте таблица 3.4.6 в дисертацията. Използван е спектралният интервал 1700 - 500 cm^{-1}, тъй като спектърът на сместа има интензивни ивици над 1300 cm^{-1}. В първите три смеси, таблица 3.4.6 a), b) и c), компонентите се откриват без проблем. При шестата смес, f), спектрите на

съставките се появяват в хит-списъка, но **изопропилбензенът** не се идентифицира успешно – има трето по стойност r.s.d. При четвъртата и петата смес, d) и e), ксилените не се появяват в хит-списъка, ако се търси в голямата библиотека: причината е във високоинтензивните ивици на **изопропилбензена**, които доминират в спектъра на сместа. При търсене в нашите шест библиотеки от общо 911 спектъра и трите смеси, d), e) и g), се анализират успешно. Ако при анализа на шестата смес се промени спектралният интервал на 1600 - 500 cm^{-1} и отново се работи с голямата библиотека от 13,484 спектъра, то пречещото съединение (третото в таблица 3.4.6f) дава r.s.d., равно на 35.6%, докато двете съставки дават r.s.d. 16.1% и 3.4%.

Анализът на други две серии от смеси, пет смеси **бутанол** и **изо-бутанол** и пет смеси на **бензен** и **пиридин** (вижте следващата точка) показва успешна идентификация на компонентите: резултатите не са дадени тук, поради значителния обем на текста на дисертацията.

Интензитетът на Раман ивиците е също пропорционален на концентрацията и това позволява използване на Раман спектри за анализ на смеси с описаната по-горе методика. На фигура 3.4.4 са показани резултатите от изчисленията по уравнение (3.4.2) със спектрите от хит-списък, получен при търсене на Раман спектър на смес от **бензилацетон** и **циклопентанон** в съотношение 1:4 о. ч. в библиотеката **RaR** от 330 Раман спектъра. Двете съставки се появяват като хитове с номера 2 и 3 и се идентифицират успешно. В таблица 3.4.7 от дисертацията са дадени резултатите за тази смес и останалите четири смеси от тези два компонента. За последните две смеси компонентът с по-ниска концентрация не може да се идентифицира, защото не се появява в списъка с резултати: както споменахме това не е недостатък на регресионния метод, а на алгоритъма за търсене на спектъра на сместа. Другата серия от смеси със същите обемни отношения, но с компоненти **циклопентанон** и **бензилацетон** дава подобни резултати, които не са приведени с цел спестяване на място.



Фигура 3.4.4. Криви на зависимост на коефициентите, изчислени при регресионния анализ на първите 40 спектри от хит-списък, получен при търсене по пикове (обратен метод) на Раман спектъра на смес от **бензилацетон** и **циклопентанон** в съотношение 1:4 о. ч. в библиотеката **RaR**. Графиките са дадени така, както излизат на екрана на програмата.

Раман спектрите са значително по-шумни от ИЧ спектрите, затова при анализа на смеси е използвано обратно търсене по пикове с параметри ΔA= 1.00, $\Delta\nu$ = 11 cm^{-1}, t_U = 0.03 (по-висока стойност от тази за ИЧ спектрите) и t_R = 0.03 (четирите стойности без дадена размерност са ординатите на Раман спектрите,

които в литературата се дават като интензитет в произволни единици).

Като цяло недостатъците на спектралната идентификация на съставките на смеси се дължи на две основни причини: (1) изпускане на спектъра на компонента с по-ниска концентрация в хит-списъка, и (2) наличие на подобни спектри в хит-списъка, които трудно се различават с математическа процедура от всякакъв вид.

3.4.5. Анализ на смеси с процедура по изваждане на спектри.

Основно изискване за приложението на тази рутинна процедура е спектрите на всички съставки на сместа да присъстват в библиотеката. Спектърът на сместа се пропуска през системата за търсене. Първият спектър от списъка с резултати се приема за принадлежащ на един от компонентите на сместа. Той се умножава с коефициент, определен от потребителя, и се изважда от спектъра на сместа. Отрицателните стойности в получения спектър се премахват, спектърът се нормира между нула и единица по ординатата и се търси отново. Предполага се, че първият спектър от новия хит-списък представя втория компонент на сместа.

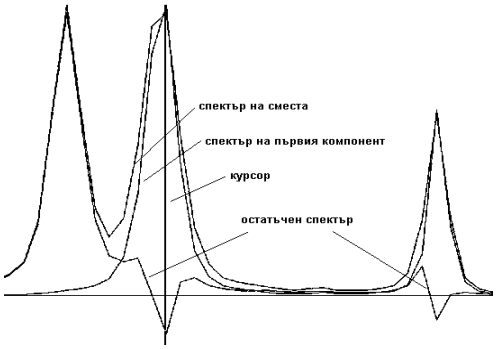
Остатък = Смес – Коефициент x Първи хит (3.4.4)

По този начин описана, процедурата по изваждане на спектри изглежда елементарна и ясна, но дори за смеси на вещества със значително различаващи се спектри може да даде грешни резултати. В литературата не се уточнява до каква степен да се провежда изваждането, споменава се само, че една или повече спектрални ивици трябва да бъдат нулирани. Друго усложнение се явява при компоненти с подобни спектри и припокриващи се ивици, което води до прекалено изваждане и загуба на ивици от втория компонент. В резултат той може да не се появи начело на хит-списъка или да се изгуби напълно.

За идентификация на чисти вещества търсенето по пълна спектрална крива дава сравнително добри резултати. Но за спектри на смеси се очаква сравняването с цялата крива да дава неточни резултати и компонентите на сместа да не са в първите две места от хит-списъка. От друга страна при търсене по пикове стойностите за неопределеност (tolerance) по вълново число и по абсорбция, оптимизирани за чисти вещества, ще са неподходящи за идентификация на съставките на смеси, заради намаляването на относителния интензитет на ивиците и изместването им по абсцисата. По тези причини за търсенето на компонентите на смеси е необходимо да се намерят нови стойности на толерансите. За целта са приготвени пет смеси на **бутанол** и **изо-бутанол** и пет смеси на **бензен** и **пиридин** и двете серии са със състав 1:4, 1:9, 4:1 и 9:1 о.ч. Техните ИЧ спектри са измерени на Perkin-Elmer 1750 FT-IR и са потърсени в библиотеката **IR1000**. Неопределеността по вълново число ($\Delta\nu$) се изследва в границите от 3 до 20 cm^{-1}. Ограничение по отношение на интензитета на абсорбцията води до лоши резултати, затова за ΔA се избира максималната стойност от 1.0 a.u. Прагът за подбиране на пикове (threshold) е свален от 0.03 a.u. за чистите до 0.01 a.u. за смеси, за да се намират пиковете на компонентите с по-малка концентрация в спектъра на сместа. Резултатите дават основание за оптимални граници на $\Delta\nu$ да се предложи интервалът от 8 до 11 cm^{-1}.

За проверка на процедурата и за оптимизиране на надеждността й са анализирани десет математически приготвени спектъра и десет реално заснети

смеси. Всички спектри са регистрирани в нашата лаборатория и са търсени в библиотеката **IR1000**. За математическите смеси може да се твърди, че притежават всички свойства на реални проби, тъй като спектрите на чистите вещества, от които са приготвени, се различават от спектрите на съответните библиотечни чисти вещества. Разликите са в базисната линия и ширината на ивиците и се дължат на различните условия на заснемане на спектрите. Тези изкуствени смеси са съставени чрез просто събиране на съответните спектри на компонентите, умножени по числата, които представляват обемните отношения. Те представляват пет спектъра на смес от **2-нитробензалдехид** и **индол-3-карбоксалдехид** в обемни отношения 9:1, 4:1, 1:1, 1:4 и 1:9 и пет спектъра на смес от **3,3-диметил-2-бутанон** и **пропиофенон** в обемни отношения 9:1, 4:1, 1:1, 1:4 и 1:9. За първите пет смеси спектрите на съединенията са регистрирани в таблетки KBr, а другите пет – в капилярен слой. Прагът на подбиране на пикове от спектъра на сместа е 0.01 a.u., търсенето е реализирано с обратен алгоритъм при $\Delta\nu$ = 9 cm^{-1} и ΔA = 1.0 a.u. Приложена бе рутинната процедура за анализ на смеси от един от съавторите в публикацията по дисертацията [D9], на когото бе неизвестен съставът на смесите. Получените резултати показват, че идентификацията не е еднозначна и в осем от случаите вторият компонент не е разпознат. Двете смеси с правилно идентифицирани компоненти на първите места в съответните хит-списъците са с отношение 1:1.



Фигура 3.4.5. Спектри на сместа, на предполагаемия първи компонент и на тяхната разлика. В спектъра на разликата се наблюдават т.нар. "крила". Спектрите са показани в интервал, около две от ивиците на спектъра на сместа.

Внимателният анализ на тези незадоволителни резултати показа няколко причини за неправилната идентификация на съставките. Първо, остатъчният спектър след изваждането е прекалено шумен и се наблюдават т.нар. "крила" (wings), причинени от различните ширини на ивиците в спектъра на сместа и в библиотечния. Второ, спектроскопистът обикновено избира една ивица и се стреми да я нулира при изваждане на спектрите. Няма гаранция, че избраната за изваждане ивица не се съдържа и в спектъра на втория компонент на сместа или не припокрива друга негова ивица. В такъв случай се губи съществена информация в остатъчния спектър. Трето, ясен критерий за прекратяване на изваждането не се използва. Колкото по-малък е коефициентът от уравнение (3.4.4), толкова повече спектрални ивици от първия хит остават; колкото по-голям е този коефициент, толкова по-малко ивици на втория компонент се запазват. Присъствието на "крила" (фигура 3.4.5.) пречи на търсенето, но е неизбежно.

За преодоляване на тези проблеми формулирахме три евристики:

(1) По време на изваждането се наблюдават повече от една ивици. Това са

ивиците с най-близки относителни интензитети и в спектъра на сместа, и в референтния;

(2) Изваждането се провежда докато ивиците, избрани по евристика (1), дадат равни положителни и отрицателни "крила", както е показано на фиг. 3.4.5;

(3) За второто търсене, търсенето на остатъка, се поставя по-висок праг за подбиране на пикове. Експериментите показват, че стойности между 0.03 и 0.05 a.u., вместо 0.01 a.u. са по-подходящи, защото остатъчният спектър е значително по-шумен.

За проверка на тези евристики са използвани десет смеси. Праговете на подбиране на пикове за търсенето на спектъра на сместа и за търсенето на остатъчния спектър са съответно 0.01 и 0.05 a.u., използва се обратно търсене с $\Delta\nu$ = 9 cm^{-1} и ΔA = 1.0 a.u. Компонентът се счита за идентифициран, ако се намира на първо място в съответния хит-списък. Резултатите са представени в таблица 3.4.12.

Таблица 3.4.12. Анализ на компонентите на смеси. В скоби е даден съответният индекс на подобие (HQI). (част от оригиналната таблица)

#	Смес	Обемно отношение	Идентифициран първи компонент (HQI)	Коефициент от уравнение (3.4.4)	Идентифициран втори компонент (HQI)
1.	бутанол и и-бутанол	1:9	и-бутанол (979)	1.02	-
2.	бутанол и i-бутанол	1:4	и-бутанол (978)	0.99	бутанол (648)
3.	бутанол и и-бутанол	4:1	бутанол (968)	0.96	-
4.	о-ксилен и м-ксилен	1:1	о-ксилен (949)	0.86	м-ксилен (888)
5.	м-ксилен и п-ксилен	1:1	м-ксилен (938)	1.10	п-ксилен (879)
6.	и-пропилбензен и о-ксилен	1:1	и-пропилбензен (949)	1.15	-
7.	и-пропилбензен и m-ксилен	1:1	и-пропилбензен (958)	1.08	м-ксилен (848)
8.	и-пропилбензен и р-ксилен	1:1	и-пропилбензен (948)	1.13	м-ксилен (878)

За седем от десетте смеси и двата компонента са правилно идентифицирани. Грешките в смеси 1 и 4 са предвидими и обясними. В смес 1 след изваждането се получава буквално плосък спектър, което е добра илюстрация на пределните възможности на метода спрямо концентрациите на компонентите на смеси. За смес 8 вторият компонент (о-ксилен) също не е идентифициран: спектърът му се намира чак на 43-то място във втория хит-списък. Причината е, че спектърът на **о-ксилен** е подмножество на спектъра на **изо-пропилбензен** (първият идентифициран компонент). Най-интензивната ивица в спектъра на **изо-**

пропилбензена е при 700 cm^{-1} и не се припокрива с ивици на **о-ксилена**. При използването на тази ивица за изваждане на спектрите, ивиците за **о-ксилен** също са били извадени. Съставките на смес 2 са успешно идентифицирани, въпреки че основният компонент е 80% от обема на цялата смес.

От таблицата се вижда още, че коефициентът от уравнение (3.4.4) е около 1.00. Този факт се обяснява с това, че се изважда първо спектърът на по-разпространения компонент от сместа, а самите спектри са нормирани в интервала 0.0 – 1.0 a.u. Това наблюдение може да се приеме като един вид допълнително уверение в правилността на процедурата на изваждане, но само при липса на пълно припокриване на ивиците на двете съставки – вижте примера, който използва фигури 3.4.6-7. Тъй като остатъчният спектър е шумен, имаше предположения, че търсенето му по спектрална крива ще дава по-добри резултати. Такова подобрение обаче не бе наблюдавано, вероятно заради следите от спектъра на първия компонент. Като цяло резултатите от тази част може да се обобщят със създадените три нови правила (евристики), дадени по-горе и четвърта евристика:

(4) Стойността множителя от уравнение (3.4.4) трябва да е около единица при липса на пълно припокриване на ивиците на двата компонента.

3.5. Анализ на структурите на съединенията от хит-списъка с метода на най-близките съседи.

3.5.1. Метод на най-близките съседи – теория.

Методът на *k^{te} най-близки съседа* (kNN, k-nearest neighbors) е отдавна добре изследван и разработен метод за обработка на резултатите от библиотечно търсене. Новост в разработеното от нас приложение на kNN метода е използването на вероятности при оценка на надеждността на набора от подструктури, които се изследват дали присъстват или отсъстват в дадено съединение, чрез търсене на неговия ИЧ спектър в спектрална библиотека.

В настоящата реализация един *kNN класификатор* е математически алгоритъм, който трансформира ИЧ спектъра в две условни вероятности, които представляват вероятността съответната химическа подструктура да присъства (P_1) или отсъства (P_2) в изследваното съединение, при условие че търсенето в спектралната библиотека дава съответния брой хитове, които съдържат химичната подструктура. Основната предпоставка на метода, която позволява ефективно създаване на класификаторите важи в много голяма степен за ИЧ спектри. А тя *отново* е, че ***подобни по структура съединения дават подобни ИЧ спектри,*** което от своя страна определя, че съединенията от получения *списък с резултати при библиотечно търсене* (hitlist) са подобни по структура на изследваното съединение.

Двете гореспоменати функции предварително са получени при библиотечно търсене на извадка от спектри (тестваща извадка) в обучаващата извадка. Тази изключително времеемка процедура е извършена само един път и е необходимо нейното повтаряне, само ако съставът на спектралната библиотека се промени.

3.5.2. Създадени kNN класификатори на ИЧ спектри.

За изследване на работоспособността на предложената модификация на kNN метода бяха избрани 20 химични подструктури или дескриптори, които характеризират структурата на едно органично съединение. Те са дадени в таблица 3.5.2 от дисертацията, заедно с броя на съединенията от голямата библиотека **IR13484**, в които те присъстват – $\mathbf{N_1}$. Много от подструктурите на класификаторите се съдържат в по-малко от половината от структурите на библиотечните спектри и традиционното използване на kNN метода с т.н. "вот на мнозинството" (majority vote) би довело до изключително грешни резултати. Същото се отнася и до структурните дескриптори, които присъстват в много повече от половината съединения – например **метиленова група** притежават 11 285 от общо 13 484 съединения, т.е. при съставяне на една случайна извадка от библиотечни структури 84% от съединенията в нея ще притежават този структурен дескриптор. В същата таблица са дадени и параметрите на класификаторите, които са генерирани със сравнение на спектрални криви с метода на коефициента на корелация и брой на хитове в списъка с хитове, $\mathbf{N_H = 50}$.

Два много важни параметъра определят ефикасността на класификаторите, която ефикасност се определя със степента на класификация при точност на класификация, равна на 90% - това са използваната спектрална мярка за сравнение и броят на хитовете в списъка с хитове. За определяне на най-добрата мярка за спектрално подобие са сравнени степените на класификация за двата класа, $\mathbf{R_0}$ и $\mathbf{R_1}$, при точност на класификация, равна на 90%. Ако поне за една от спектралните мерки за подобие не може да се намери $\mathbf{R_k}$ при $\mathbf{P_k = 90\%}$, е взета стойността при $\mathbf{P_k = 85\%}$ или даже $\mathbf{P_k = 80\%}$, ако нямаме и 85% вероятност. Данните от тези сравнения са дадени в таблица 3.5.3, при което е приложен t-тест (paired t-test) със степени свобода 19 и избрана статистическа сигурност $\mathbf{P = 90\%}$: интегралната граница при двустранна постановка на задачата е $\mathbf{t(19,0.90) = 1.73}$.

Таблица 3.5.3. Сравнение между класификациите с kNN класификатори, генерирани с различни мерки на спектрално подобие. Числените стойности в таблицата са t-критерия. Означения: LS – средното квадратичното отклонение между спектрите, AV – средното абсолютно отклонение, SP – скаларно произведение, и CC – коефициент на корелация. Ако числото в таблицата е положително и по-голямо от интегралната граница, $\mathbf{t(19,0.90) = 1.73}$, то съответният метод в първия ред е по-добър от този в първата колона, и имаме обратното, ако числото е отрицателно и неговата абсолютна стойност е по-голяма от 1.73.

a) сравнение на степента на класификация за първи клас, R_0

Спектрално подобие	LS	AV	SP
AV	**-4.2**		
SP	-1.0	1.6	
CC	-1.5	1.1	-1.3

b) сравнение на степента на класификация за втори клас, R_1

Спектрално подобие	LS	AV	SP
AV	**3.0**		
SP	-1.4	-1.4	
CC	-0.6	**2.8**	1.0

От таблица 3.5.3 се вижда, че когато се сравняват степените на класификация за спектрите от нулевия клас, единствената статистически отличима разлика е

между LS и AV, в полза на AV. Докато от сравнението на R_1 се виждат следните подредби LS > AV и AV > CC. Това доста объркващо сравнение не показва, че има метод, който да преобладава над другите и по двата критерия.

Анализът в дисертацията на тези данни от 3.5.3 показва, че е невъзможно да се избере оптимален алгоритъм за търсене по спектрална крива, нито определен оптималния брой на хитовете, които се обработват от KNN класификаторите за различните структурни фрагменти от таблица 3.5.2. За практически приложения се използват 20 класификатора с N_H = 50, тъй като този брой хитове се оказва оптимален за класификацията с използване на максимална обща подструктура, вижте т. 3.6.

3.6. Анализ на структурите на съединенията от хит-списъка с концепцията на максимална обща подструктура.

В тази част е описан подход за класификацията на ИЧ спектри на органични съединения чрез използване на концепцията за максимална обща подструктура (МОП) на две химични структури. Когато непознатото химично съединение не може да се идентифицира с помощта на търсене на неговият ИЧ спектър в спектрална библиотека, полученият хит-списък се анализира с помощта на подхода, описан в т. 3.6.2. За структурите на съединенията от хит-списъка се изчисляват най-често срещаните в тях подструктури, наречени от нас *характеристични подструктури*. Нашите експерименти показват, че те могат да бъдат една добра основа за разкриването на структурата на непознатото съединение.

3.6.1. Използване на максимална обща подструктура на двойките хитове.

Максимална обща подструктура на две химични съединения се нарича най-голямата подструктура, която е обща за тях. В структурната химия МОП намира приложение предимно за оценяване на подобието между две структури. Друго приложение на МОП е анализът на резултатите от търсене в библиотека от спектри.

Програмата **ToSiM** съдържа алгоритъм за определянето на МОП на две дадени химични структури, които са представени със своите таблици на свързаност. Алгоритъмът намира поредица от "атом-атом" и "връзка-връзка" съответствия (matches) в двете молекули, като започва от произволна двойка съответстващи си атоми (по един от всяка молекула) и продължава със сравнението на връзките и атомите, следвайки разклоненията на едно топологично дърво. В резултат се получава свързана подструктура, чиято топология и тип на атомите (възлите) отговаря на определена част от едната и от другата структура. Потребителят на програмата може да въведе следните параметри или ограничения при сравнението: (a) да /не/ се проверява типът на съответстващите си атоми, (b) да /не/ се проверява типът на съответстващите си връзки, (c) да /не/ се разглеждат водородните атоми в двете подструктури, (d) всички хетероатоми да /не/ се приемат за идентични, и (e) въвеждане на минималния брой неводородни атоми в максималната обща подструктура.

МОП на две химични сруктури е една добра мярка за тяхното подобие, но когато броят на сравняваните структури нараства, МОП губи своето значение като

такава. За n^{Te} структури от хит-списъка са възможни $n(n-1)/2$ двойки структури, и съответно толкова МОП се изчисляват от програмата. За всяка подструктура се преброяват нейните присъствия (честота) в структурите на хит-списъка, n_i, и МОП се подреждат по своя ранг R_j:

$$R_j = (1 - f) \cdot n_j / n + f \cdot A_j / A_{max}, \quad (3.6.1)$$

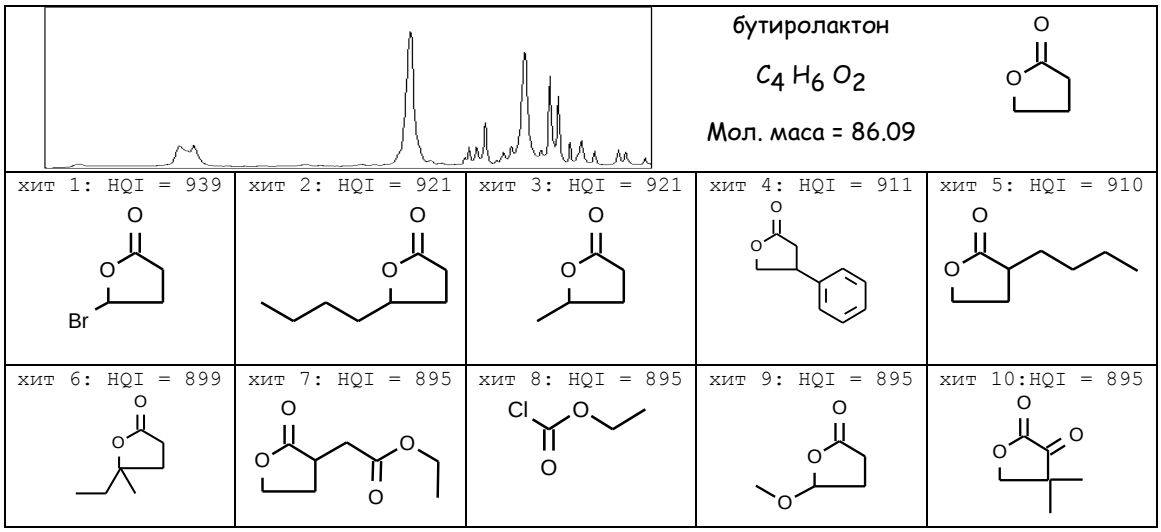
където A_i е броят на неводородните (тежки) атоми в съответната МОП, A_{max} е максималният брой на неводородни атоми във всички n изследвани структури, а f е коефициент, определян от потребителя.

Алгоритъмът за анализ на структурите от хит-списъка е представен на фигура 3.6.1, дадена само в дисертацията. Всички търсения на ИЧ спектри са извършени в библиотеката **IR13484**, както и спектрите на изследваните "непознати" съединения са избрани от нея. При всички резултати от търсенията е премахвана първата структура в хит-списъка, което на практика означава, че поставената пред изследователя задача е да се определи структурата на химично съединение, което няма спектър в библиотеката. Проведени са експерименти със всичките седем алгоритъма за спектрално търсене, които са реализирани в програмната система, за да се определи този от тях, който дава най-добри резултати. Също така е изследвано влиянието върху получаваните резултати на броя на структурите от хит-списъка, за които се провеждат изчисленията на МОП.

От всички параметри (ограничения) при изчисляване на набора от МОП ние сме изследвали влиянието върху резултатите на коефициента f и ограничението (c). Ограниченията (a) и (b) са положителни, а (d) - отрицателно, което означава, че две (под)структури се приемат за еднакви, ако всички атоми, различни от водородните, и всички химични връзки в едната от тях съответстват на атомите и връзките от другата. Това положение, според нас, съответства на физическата същност на ИЧ спектроскопия. Минималният брой на неводородните атоми в МОП е избран равен на минималната възможна стойност - четири.

3.6.2. Предварителни изследвания.

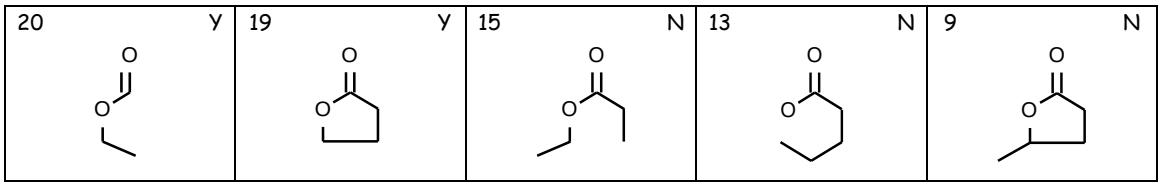
Резултатите от анализа на хит-списъка, получен при търсене на спектъра на **бутиролактон** са типични за молекули с ниска молекулна маса. Извършено е търсене в библиотеката по метода на коефициента на корелация, уравнение (3.1.5). ИЧ спектър на **бутиролактона** и структурите на първите 10 съединения от хит-списъка са дадени на фигура 3.6.2. Петчленният лактонен пръстен присъства в осем от структурите (девет, ако се пренебрегне sp^2 хибридизацията на един от въглеродните атоми в десетия хит), а естерната група - във всичките десет съединения. Получаваните резултати зависят от големината на анализирания хит-списък. Това влияние може да се демонстрира при сравнение на резултатите от МОП анализ на 20 и 50 структури, съответно фигури 3.6.3 и 3.6.4, а първите 10 хита са дадени на фигура 3.6.2. На първата фигура са дадени първите пет характеристични подструктури, а на втората - първите 15; и на двете фигури те са подредени по честотата на тяхното срещане в структурите на хит-списъка. Изобразените подструктури да се възприемат без съответните водородни атоми, защото МОП анализът е извършван с пренебрегване на водородните атоми в сравняваните структури.



Фигура 3.6.2. Първите 10 структури от хит-списък, получен при търсене на спектъра на "непознатото" съединение **бутиролактон**.

В първия случай **бутиролактонният пръстен** се съдържа в 19 от структурите на хит-списъка, а фрагментът C–C–O–C=O -- във всичките 20. Останалите три намерени подструктури са с честота 15, 13 и 9, но *не се съдържат* в изследваната структура. Докато във втория случай 10 от общо 15^те^ характеристични подструктури се съдържат в **бутиролактона**; сред тях са първите девет.

При използването само на първата структура от хит-списък, МОП анализът ще даде точно нея като характеристична подструктура, която ще е грешна, ако нямаме спектър на изследваното вещество в библиотеката. При МОП анализ на всички структури от библиотеката (другият краен случай) бихме получили подструктури, характеристични за библиотеката като цяло, но не и за изследваното съединение. Ето защо съществува някакъв оптимален брой на сравняваните структури. Нашите предварителни изследвания показаха, че този брой е някъде около 40 до 70 структури и той зависи от конкретната изследвана структура и вида и размера на спектралната библиотека. В тези предварителни изследвания сме се ограничили на 50 структури, тъй като с повишаване на техния брой изчисленията нарастват в геометрична прогресия и силно зависят от големината, разклонеността и цикличността на сравняваните структури.



Фигура 3.6.3. Първите пет характеристични подструктури, намерени при обработката на хит-списък от 20 структури, получен при търсене на спектъра на **бутиролактон**. За всяка подструктура е даден броят на нейните появи в структурите на хит-списъка. **Y/N** - подструктурата присъства/отсъства в изследваната структура.

Варирането на **f** в интервала (0.0, 1.0) с 0.1 показа, че ако искаме верните (*reliable*) подструктури да са в началото на получавания списък, то трябва да се работи с **f = 0.0**, т.е. подструктурите да се подреждат *изцяло* по тяхната честота на поява в хит-списъка. Това на практика означава предимство за по-

разпространените подструктури в хит-списъка и напълно отговаря на опита на специалистите по ИЧ спектроскопия, за които съединенията от хит-списъка, *взети заедно* са една добра оценка на структурата на неизвестното вещество. От друга страна, малките структури са по-ниско информационни, т.е. няма да намаляват значително броя на структурите, генерирани от структурния генератор. Ето защо в една от следващите точки е разгледан системен подход за оптимизиране на МОП по коефициента **f** и размера **n** на използвания хит-списък.

Да предположим, че сме подбрали случайно 50 спектъра от цялата библиотека (13 484 спектъра) и те съставят хит-списъка. Ако честотата на срещане на дадена подструктура в библиотеката е **p**, то се очаква приблизително същата честота и в избрания хит-списък, понеже той е случайна извадка от библиотеката. Вероятността за наличие на **k** подструктури в хит-списък от **n** спектъра е оценявана в тази предварителна работа с биномиалното разпределение:

$$p(k) = [n!/(k!(n-k)!)]\ p^{k}\ (1-p)^{n-k} \qquad (3.6.2)$$

Правилната оценка е с хипергеометричното разпределение, тъй като изборът на едно съединение в списъка с резултати намалява броя на останалите библиотечни спектри с единица, и съответно така се променя вероятността **p**. Но за голяма библиотека от 13 484 спектъра и голям списък с резултати, 50 на брой, получаваните оценки по двете разпределения на практика съвпадат: в дисертацията този въпрос е подробно обяснен с примери и една фигура.

Нарастването на честотата на срещане на подструктурата в хит-списъка, **k/n**, над дадена граница показва, че хит-списъкът не е случайна извадка от библиотеката, а е подбран по специални критерии (в случая библиотечно търсене), които облагодетелстват появата на разглежданата подструктура в него. *Един от тези критерии* е приликата на търсения спектър с тези от хит-списъка - прилика, която се пренася и върху съответните структури на последния. Ако подструктурата се среща n_j пъти, то сумата α от уравнение (3.6.3) дава вероятността за грешка от първи род, т.е. когато нулевата хипотеза (хит-списъкът е случайно избран) е вярна, ние я отхвърляме. Ако имаме много малка стойност на α то можем да отхвърлим нулевата хипотеза и да приемем, че подструктурата се среща в хит-списъка поради някакви обективни причини.

$$\alpha = \Sigma\ p(k);\ k = n_i\ \ldots\ n \qquad (3.6.3)$$

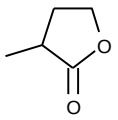
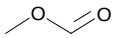
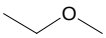
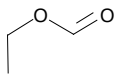
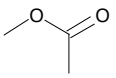
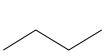
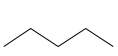
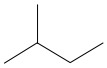
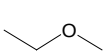
Например, изчисленото ниво на значимост по формула (3.6.3) за 12^та подструктура от фигура 3.6.4 с **p = 3909/13484 = 30.0%** и **k/n = 27/50 = 0.54** по уравнения (3.6.2) и (3.6.3) е α **= 0.000184**. Всичките останали подструктури от фигурата имат по ниско ниво на значимост α. Проведените експерименти със спектрите на други съединения показват, че всички получавани характеристични структури (18 на брой) са статистически значими със $\alpha < 0.001$.

Повечето характеристични ивици в ИЧ спектър се дължат на трептения на фрагменти в молекулата от вида **X-H**, ето защо би било уместно при провеждане на МОП анализа да се отчита броят на водородните атоми при съответните тежки атоми. Нашите изследвания показаха, че това *не води* до по-добри резултати. Другият фактор, който в значителна степен определя вида на получаваните резултати, е използваната при библиотечното търсене спектрална мярка на

подобие. Нашите предварителни изследвания показаха, че алгоритмите за търсене по пикове дават по-лоши резултати от тези, които използват спектрална крива. От четирите алгоритъма, програмирани от нас - уравнения (3.1.2)-(3.1.5) - този, който използва коефициента на корелация показва най-добри резултати.

3.6.3. Сравнение на максималните общи подструктури, генерирани при търсене на ИЧ и мас-спектри.

Описаният подход за намиране на МОП от структурите на съединенията на хит-списъка може да се приложи към други видове спектроскопия.

неизвестна структура	сп. мет.	n	характеристични подструктури и тяхната честота 1	2	3	4	5	*средна честота*
	ИЧ	50	46 	44 	42 	42 	40 	86 %
$C_5H_8O_2$	MS	36	26 	18 	16 ***невярна*** 	13 	11 	47 %

Фигура 3.6.11. Пет характеристични структури (в дясно), които са най-разпространени в списъка с резултати, получен при търсене на ИЧ и мас-спектрите на структурите във първата колона. Означения: **n** – размер на списъка с резултати; над подструктурите са дадени честотите на тяхното срещане в хит-списъците. Съкратен вариант на таблицата.

За това изследване е използвана спектралната мярка коефициент на корелация на спектрални криви, уравнение (3.1.5). Получените резултати могат да се обобщят на фигура 3.6.11.

Първият факт, който се вижда е, че обработката с МОП на ИЧ хит-списъците дават характеристични подструктури, които по-често се срещат в хит-списъците, отколкото тези получени с мас-спектри – сравнете последната колона по двойки. Горното сравнение е в полза на ИЧ спектроскопията – стойностите са толкова различни, че даже не е необходим статистически тест за доказване на предното твърдение. По-голяма честота на вярна характеристична подструктура означава, че повече от съединенията в хит-списъка са подобни по структура на „неизвестното", чийто спектър е бил потърсен в библиотеката.

Вторият извод е по отношение на структурния състав на характеристичните подструктури: тези от ИЧ списъците изобилстват на хетероатома кислород и затова са по-добри като качество. Обаче, най-важният фактор е големината на фрагментите в списъка **GOODLIST**, но той както се вижда от фигура 3.6.11 е съезмерим при двата спектрални метода – даже по големина характеристичните подструктури просто си „съответстват" в единия и другия спектрален метод.

Допълнително, две от характеристичните подструктури, получени с мас-спектри са неверни. Интересно е, че обработката на мас-спектралните хит-списъци с концепцията на МОП е направена подобно на ИЧ хит-списъците – без използване на водородните атоми, въпреки че в мас-спектрите йоните имат точно определена маса, която без съмнение се определя и от водородните атоми в тях.

3.6.4. Оптимизиране на параметрите на алгоритъма за прилагане на концепцията за максимална обща подструктура.

Твърде много на брой параметри на МОП алгоритъма определят вида на получаваните резултати. Както споменахме по-горе в т. 3.6.2 едно от директните ограничения (параметри) на алгоритъма е следнито: (c) да /не/ се разглеждат водородните атоми в двете подструктури. Предвид теоретичните основи на ИЧ спектроскопията изборът на ограниченията (a) и (b) може да е само положителен, а (d) – отрицателно: и видът на връзките, чрез техните силови константи, и видът на атомите, чрез техните маси изрично определят вълновите числа на ИЧ ивиците, а от там и цялостния вид на ИЧ спектъра. Експериментите с ограничение (c) показаха, че МОП трябва да се генерират без да се взимат под внимание, съответните водородни атоми. Минималният брой на неводородни (тежки) атоми в МОП се задава по избор на потребителя: малките по размер МОП са много по-достоверни, но по-безполезни при структурната генерация, а големите МОП по правило са по-малко верни (не присъстват в изследваната структура). Използваното число от нас във всички изчисления е поне 4 тежки атома в генерираните МОП.

Освен тези пет параметъра на самия МОП алгоритъм има три много важни параметъра – (1) видът на спектралната мярка, (2) големината на използвания хит-списък, **n**, и (3) параметърът **f** от уравнение (3.6.1). Ако се вземе само първия хит (**n = 1**) и концепцията за МОП се обобщи, то МОП ще е самата структура на първия хит, и ако нямаме вярна идентификация, то МОП ще е погрешна. При избор на хит-списък с размер на спектралната библиотека (**n = 13484**, в нашия случай), получените МОП ще характеризират самата спектрална библиотека, а не резултатите от спектралното търсене. Затова се очаква, че има някакво число, което е оптимална стойност за големината на обработвания хит-списък, **n**, и то е между 1 и **n**. Същите разсъждения могат да се приложат и за коефициента **f** от уравнение (3.6.1). При **f**, равно на нула в уравнение (3.6.1), на първите места в списъка на МОП са най-често срещаните подструктури, а те очевидно са по-малки от останалите МОП (макар и по-достоверни, както показват нашите експерименти) и по-малко полезни като ограничения при структурната генерация, а при **f = 1**, алгоритъмът поставя най-големите структури на първите места в списъка на МОП, според техния ранг. Очевидно, колкото една МОП е по-голяма, толкова по-рядко ще се среща в списъка с резултати, и съответно ще е по-ненадеждна (казано по друг начин – по-нехарактеристична). Ето защо, целта е да се намерят стойности на двата параметъра, които дават в средно най-верни и най-полезни характеристични подструктури в първите места на списъка с МОП.

Биха могли да се изнамерят много числови критерии, наречени от нас мерки за ефикасност или само ефикасност (measure of effectiveness, effectivity measure), които описват полезността на генерираните МОП, но всички те трябва да отчитат поне техния размер и вярност. Един от най-простите на вид е предложеният от нас критерий по уравнение (3.6.4), в което уравнение сумата е по всички генерирани МОП (**k** = 18 в нашия случай: това е ограничение на програмата **ToSiM**).

$$\mathbf{E = \Sigma\ p_i\ n_i\ A_i\ /(k\ n\ A);\quad i = 1,\ 2,\ \ldots\ \ k,} \qquad (3.6.4)$$

където **k** е броят на разглежданите МОП, **n** е размерът на хит-списъка, **A** е размерът (в брой тежки атоми) на „неизвестното" съединение, $\mathbf{n_i}$ и $\mathbf{A_i}$ са честота и размерът (в брой тежки атоми) на генерираните МОП. В знаменателя **k** нормира сумата по броя генерирани МОП, **n** - по броя на използваните хитове, а **A** неглижира влиянието на размера на „неизвестната" структура.

Числата $\mathbf{p_i}$ са така наречените от нас наказателни коефициенти (penalty coefficients). Естесвен избор за тях са двойката стойностите +1 и -1 или +1 и 0, при вярна и невярна МОП, съответно. В настоящата работа е избран първия вариант, $\mathbf{p_i = +1}$ или $\mathbf{p_i = -1}$. Ясно става, че структурното разнообразие на отделните МОП не се отчита от този коефициент (3.6.4), а то е от много голямо значение.

За определяне на оптималните стойности на **f** и **n**, 10 съединения, таблица 3.6.2 в дисертацията, бяха избрани на случаен принцип в ролята си на „неизвестни" – те са структурно разнообразни и техният размер варира в широки граници. Техните ИЧ спектри бяха потърсени в спектралната библиотека **IR13484**, съответните хит-списъци бяха обработени с програмата **ToSiM** при вариране на **f** и **n**, и за получените набори от характеристични подструктури бяха определяни мерките за ефикасност - първият хит е премахван от хит-списъка, защото очевидно е „неизвестното" съединение.

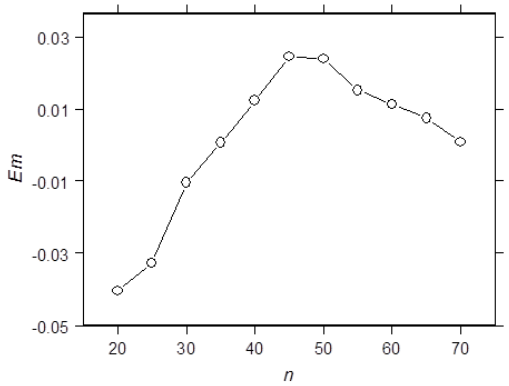
Един много важен параметър е видът на използваната спектрална мярка за подобие, която непосредствено определя вида на структурите на хитовете – нашите експерименти (т. 3.6.2) показаха, че спектралните мерки за подобие по уравнения (3.1.4) и (3.1.5) дават най-добри резултати, с лек превес на (3.1.5). Но това бяха несистемни наблюдения, затова тук са сравнени пет спектрални мерки за подобие – тези, които използват спектралните криви по уравнения (3.1.2) – (3.1.5) от т. 3.1.3, означени в тази част и публикация [D7] като **SD**, **AD**, **SP** и **CC**, както и сравнението на пикове с правия алгоритъм от т. 3.1.2, означено като **PM**.

ИЧ спектрите на десетте съединения са потърсени в спектралната библиотека **IR13484** и по методиката, описана в т. 3.6.1, са определени $18^{те}$ максимални общи подструктури за всяко едно от тях. Размерът на хит-списъците е 50 структури, а коефициентът **f** от уравнение (3.6.1) е **0.0**, т.е. МОП са сортирани по тяхната честота на срещане сред структурите на $50^{те}$ хита. Получените ефикасности са дадени в таблица 3.6.3 от дисертацията, където също са изчислени медианата, първия и третия квартил, както и средната стойност на резултатите за всяка една от спектралните мерки.

Тези стойности на ефикасността дават възможност да се сравнят спектралните мерки. По техните медиани и средни стойности мерките се подреждат по ефикасност така: **CC > AD > SD > SP > PM**, но при използване на теста на Wilcoxon единствените статистически отличия (със статистическа сигурност, равна на 90%) са между **CC** и **SP**, както и всички мерки по уравнения (3.1.2) – (3.1.5) са статистически по-добри от търсенето по пикове. Затова при останалите изследвания е използвана спектралната мярка за подобие *коефициент на корелация на спектрални криви*, изчислявана по уравнение (3.1.5).

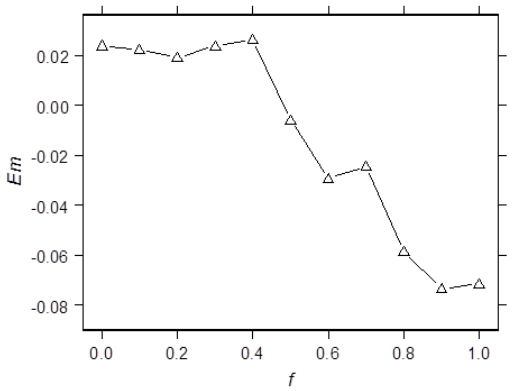
Зависимостта на средната ефикасност, **Em**, от броя хитове е представена на фигура 3.6.12. Двете най-високи стойности са за 45 и 50 хита$^{\#}$, факт, който бе

определен от нас при нашите предварителни изследвания – там беше прието, че за около 50 хита се получават най-добри резултати.



Фигура 3.6.12. Средната ефикасност, $\mathbf{E_m}$, като функция от големината на хит-списъка, **n**. Спектралният метод е коефициента на корелация, а **f** = 0.0.

Зависимостта на средната ефикасност от коефициента **f** в уравнението за ранга на характеристичните подструктури (3.6.1) е представена на фигура 3.6.13. Вижда се че най-добри резултати се получават за **f** в интервала [0.0-0.4]. Интересно е, че от редицата проверки на МОП концепцията, за които споменахме в бележка под линия по-горе, стигнахме до извода, че този интервал е [0.0-0.3].



Фигура 3.6.13. Средната ефикасност, $\mathbf{E_m}$, като функция от коефициента **f** от уравнението за ранга на МОП (3.6.1).

Обяснението за почти хоризонталния интервал между **f = 0.0** и **f = 0.4** е, че двете противоположни тенденции в уравнението за ефикасността – намаляване на $\mathbf{n_i}$ и повишаване на $\mathbf{A_i}$ - се изравняват в този интервал. При стойности на **f** над 0.4 вече наказателните коефициенти $\mathbf{p_i}$ оказват влияние – появяват се по-големи, но грешни МОП.

3.6.5. Друг ранк за сортиране на максималните общи подструктури.

Подреждането на характеристичните подструктури по уравнение (3.6.1) не отчита позициите на съединенията от хит-списъка, които съдържат съответната МОП. Например, нека хит-списъкът е с големина **n = 50** и честота на присъствието на дадена МОП е 4. Очевидно е, че има значение дали например 1, 2, 3 и 4 хит съдържат подструктурата или това са хитове 47, 48, 49 и 50: спектроскопистът ще приеме първия случай за много по-достоверен. За отчитане на тези позиции рангът за сортиране на подструктурите е изменен и се изчисява по формулата:

$$Rh_j = (1-f)\frac{\sum_{k=1}^{n_j}(n+1-h_k)}{\sum_{k=1}^{n}(n+1-k)} + f\frac{A_j}{A_{\max}} \quad , \qquad (3.6.5a)$$

където h_k е позицията в хит-списъка на съответната структура, която съдържа получената МОП; останалите означения са като тези при уравнение (3.6.1).

Вижда се в числителя на (3.6.5), че колкото позицията h_k на хит-структурите е по-голяма, толкова по-малка стойност ще има R_j. Ако всички хитове съдържат съответната МОП, то числата h_k са равни на 1, 2 ... n, и отношението на двете суми става единица, както това е в (3.6.1), където ще имаме $n_j/n = 1$.

Най-интересното свойство на новия ранг Rh_k е, че стария ранг R_k е частен случай на Rh_k, като се получава от него при условие, че структурите, които съдържат характеристичната подструктура **j**, са точно по средата на хит-списъка. В дисертацията е дадено математическото доказателство за това свойство.

Сравнението на двата ранга е направено в две публикации по дисертацията - [D21] и [D8], при които е използвана програма **ToSiM** за обработка на хит-списъка по стария ранг и програма, написана от Николай Кочев, която обработва по-новия ранг. Авторът няма претенции към програмирането на МОП алгоритъма, но алгоритъмът използва една евристика (вижте описанието и в дисертацията), която неимоверно ускорява изчисляването на МОП, и която е предложена от автора (така както и новия ранг).

Като тестови съединения са избрани същите десет от таблица 3.6.2. Тъй като евристиката (вижте забележката под линия) води до намирането на само една от всички МОП, еднакви по размер, то не е задължително намерената МОП да съвпада при използване на нашата програма и програмата **ToSim** – затова резултатите са сравнявани и за двете програми, когато се използва стария ранг от уравнение (3.6.1). Резултатите са дадени в таблица 3.6.4 от дисертацията, където е показан броят на първите верни, T_{18} (от top correct), МОП в сортирания списък по стария и по новия ранг и с използването на нашата програма и програмата **ToSiM**. От тези данни се вижда, че новата реализация на МОП алгоритъма с нашата програма (колона **B**) дава съизмерими резултати с реализацията му в програмата **ToSim** (колона **A**). Сравнението на колони B и D от таблица 3.6.4 би показало дали новият ранг дава по-добри резултати по този показател T_{18}, който е броя на първите верни МОП в списъка. За съжаление, прилагането на теста на Wilcoxon дава $T_- = -18.5$ и $T_+ = 36.5$, което показва че данните са статистически неразличими поне с ниво на значимост 10% ($T_{min} = 11$).

Този неопределен резултат провокира изследването да продължи с избора на 100 случайни съединения от библиотеката **IR13484**, които да бъдат т.н. „неизвестни" съединения и чийто спектри са потърсени в библиотеката. Използвана е същата методика, описана по-горе в досегашните МОП изследвания с тази библиотека. Размерът на хит-списъка е 50 структури, коефициентът **f** = 0.3, а от сравняването на 1225 двойки структури в хит-списъка се запазват първите 50 характеристични подструктури, сортирани по R_k или Rh_k. За всяка една МОП се проверява дали тя е вярна или не, и се съставят два списъка от 100 числа T_{50} – за стария ранг и за новия. Прилагането на теста на Wilcoxon дава статистическо отличие на двете серии с ниво на значимост $\alpha < 0.001$, в полза на новия ранг.

3.7. Класификация на мас- и ИЧ спектри по химични подструктури.

В тази част е описана класификацията на мас- и ИЧ спектри, с помощта на модела на изкуствените невронни мрежи (ИНМ). За мас-спектри те се прилагат

директно върху спектъра, който се разглежда като **N**-мерен хемометричен образ. Аналогично се разглежда и изходът от класифициращия алгоритъм – също многомерен образ, който в конкретния случай съдържа информация за определени химични фрагменти в, или молекулната формула на, съединението. За ИЧ спектри алгоритъмът се прилага върху набор от спектрални признаци, които се изчисляват от съответния спектър.

3.7.1. Директна класификация на спектри с ИНМ.

Невронните мрежи с обратно разпространение на грешките са изградени от няколко слоя неврони: входен слой, няколко скрити слоя и един изходен слой – в дисертацията е описан подробно модела на ИНМ, която е използвана за изследванията в тази част - ИНМ с един скрит слой, с право разпространение на сигналите и обратно разпространение на грешките.

3.7.2. Резултати от класификацията на мас-спектри по химични подструктури.

Описаният в т. 3.7.1. модел на ИНМ с право разпространение на сигналите и обратно разпространение на грешките беше използван за класификацията на мас-спектри по химични подструктури и по молекулна формула. В таблица 3.7.1 от дисертацията са дадени $16^{те}$ химични подструктури или структурни характеристики и процентите на съединенията, които ги притежават (percentage occurrence).

Класификацията бе оптимизирана по два параметъра – броя скрити неврони, NHN, и представянето на спектралната информация. Първият параметър определя броя на коефициентите в ИНМ, чийто стойности съдържат спектро-структурните корелации. При малък брой скрити неврони ще имам малък брой коефициенти $\mathbf{w21}_{i,j}$ и $\mathbf{w32}_{j,k}$ (и съответните офсети θ_j и θ_k) и получените корелации няма да са добри, а при голям брой на невроните, съответно коефициентите, ще имаме много добро обучение, което ще „хване" локалните тенденции в обучаващата извадка (ОИ), но няма да може да обобщи (generalize) основните тенденции в корелациите и така няма да може да предсказва подструктурите от спектрите на тестващата извадка (ТИ) или други спектри, които не са включени в обучаваща извадка.

В настоящото изследване ние избрахме да изследваме как следните преобразувания на мас-спектрите ще повлияят на предсказващата способност. Представянето на спектрите беше избрано да е по следните пет формули:

$$I_j = (A_j/100)^2 \qquad (3.7.3a)$$

$$I_j = A_j/100 \qquad (3.7.3b)$$

$$I_j = \mathrm{sqrt}(A_j/100) \qquad (3.7.3c)$$

$$I_j = \log_{10}(A_j)/2 \text{ and } I_j = 1 \text{ if } A_j = 1 \text{ and } I_j = 0 \text{ if } A_j = 0 \qquad (3.7.3d)$$

$$I_j = 1 \text{ if } A_j = 0 \text{ and } I_j = 0 \text{ if } A_j = 0 \qquad (3.7.3e)$$

Съответните спектрални признаци по уравнения (3.7.3a-e) могат да се нарекат квадратични спектри, сурови спектри, коренувани спектри, логаритмувани спектри и двоични (бинарни) спектри. При всичките пет преобразувания винаги най-

високият пик в мас-спектъра (който има $A_j = 100\%$) ще има стойност единица на ординатата.

В дисертацията е описана подробно методиката на изследванията, а тук само ще спомена, че за ефективността на обработката на спектрите се получи следната последователност, в която предсказващата способност намалява:

сурови > логаритмувани > двоични > коренувани > квадратични

Всичките 140 спектъра бяха използвани и за 28-пъти кръстосана валидация (k-fold cross-validation). От проведените 28 пъти обучения върху 135 спектъра и съответно валидация с останалите 5 бяха изчислени предсказващите способности на модела, които са дадени също в таблица 3.7.1.

Въпреки показаните високи предсказващи способности (от 76.1% за карбонилната група до 98.6% за пиридин), тези резултати не са толкова добри. В таблицата нарочно са дадени и двете стойности PO_0 и PO_1, въпреки че тяхната сума е 100%, за да може да се прецени от читателя, че например за фрагмента ФУРАНОН имаме $PO_0 > PA$, което означава, че въобще нямаме класифициращa способност: например ако даден изследовател твърди, че нямаме фуранонов фрагмен, то той ще е прав в 91.5% от случаите (това число е PO_0), докато с ИНМ ще познаем в 87.3% от случаите (това число е PA). Тук трябва да се отчете, че PA е изчислявана като изходът от ИНМ между 0.4-0.6 се приема за неверна класификация. Анализът на тези задоволителни резултати, както и пребиваването на автора в лабораторията на проф. Вармуца ни насочиха към използването на друг модел за класификация на спектрална информация, който е взаимстван от работите на Вармуца и сътрудници, но е приложен с редица подобрения към класификацията на ИЧ спектри по подструктури – вижте т. 3.7.3.

Освен по подструктури мас-спектрите са класифицирани и по молекулна формула с друга, отделна ИНМ по схемата. В таблица 3.7.2, представена само в дисертацията, са дадени съответните предсказания, направени отново с 28-пъти кръстосана валидация. Прави впечатление ниската предсказателна способност, но ако се разгледат в подробности четирите подтаблици ще се види, че грешките в предсказването на молекулната формула са малки по отклонение. Главният извод от тези изследвания бе, че за разкриване на спектро-структурните корелации е необходима голяма и представителна спектрална библиотека. В заключение ще спомена, че тези изследвания имат своето значение за проверка на програмирания от автора модел на ИНМ.

3.7.3. Алгоритъм за създаване на класификатори на ИЧ спектри.

Моделът за тази класификация е взаимстван от литературата, където е използван за класификация на мас-спектри с ЛДА и ИНМ с радиално разпределени функции, но са направени редица промени и са въведени нови спектрални признаци, изчислявани от ИЧ спектрите.

На фигура 3.7.2, дадена само в дисертацията, са представени процедурите за създаването на класификатори. *Създаването на обучаваща и тестваща извадка* е първата стъпка от изчисляването на класификатора. На този етап двете извадки се състоят от пиковите таблици, извлечени от ИЧ спектри. Следващ етап е *изборът на спектрални признаци*, които осигуряват оптимална класификация на ИЧ

спектри. Наборът от спектрални признаци за дадена поредица от химични съединения съдържа неявно определена част от спектро-структурните корелации, и колкото по-голяма е тази част, толкова по-добри са избраните спектрални признаци. В настоящата работа сме използвали два типа спектрални признаци, които се изчисляват от пиковите таблици на ИЧ спектри. Първият вид е наречен от нас *интервален признак* и се дефинира като интензитетът на максималния пик в даден спектрален интервал. Вторият вид е т.н. *логаритмичен признак* и се дефинира като логаритъм от отношението на двата най-високи пика в даден спектрален интервал. Математически спектралните признаци се изразяват така:

$$\mathbf{INT}(\nu_1, \nu_2) = \begin{cases} A_{max}/100.0 \\ 0.0, \text{ ако няма пикове в } (\nu_1, \nu_2) \end{cases} \qquad (3.7.4)$$

$$\mathbf{L12}(\nu_1, \nu_2) = \begin{cases} [a - \lg(A_{max}/A_{sec})]/a;\ a = 2.0 \\ 0.0, \text{ ако има по-малко от 2 пика в } (\nu_1, \nu_2) \end{cases} \qquad (3.7.5)$$

където $\mathbf{A_{max}}$ и $\mathbf{A_{sec}}$ са интензитетите на максималната, и съответно, втората по интензитет ивица в спектралния интервал (ν_1, ν_2).

Логаритмичните спектрални признаци отчитат факта, че някои химични групи дават няколко ивици в даден характеристичен интервал. Във формула (3.7.5) признакът е отместен, така че да има максимална стойност при равни интензитети на двете ивици - фигура 3.7.3.

пикове					няма пикове
$\mathbf{L12}(\nu_1, \nu_2)$	1.00	0.85	0.50	0.00	0.00

Фигура 3.7.3. Логаритмичните признаци, **L12**.

В интервала 4000 - 400 $см^{-1}$ при вариране на ν_1 и ν_2 с 1 $см^{-1}$ е възможно изчисляването на $3600 \cdot 3559/2 \approx 6.5\ 10^6$ признака само от единия вид. Ето защо е необходимо провеждането на специален подбор на признаци, който да осигури оптимален набор за дадена класификационна задача. Ние приемаме, че характеристичните интервали на поглъщане на съответните структурни елементи са едно добро начално приближение на съответните интервали, в които се определят признаците по формули (3.7.4) и (3.7.5). По този начин се ограничаваме с изследването на значително по-малък брой признаци - максимум от порядъка на 70 хил. За даден структурен елемент предварително се избират характеристичните му интервали по литературни източници. За всеки признак **k** се изчисляват средните му стойности за обектите от първия ($\mathbf{A_{1,k}}$) и втория клас ($\mathbf{A_{2,k}}$), и съответните стандартни отклонения ($\mathbf{S_{1,k}}$ и $\mathbf{S_{2,k}}$). От тези статистики се изчислява следното отношение:

$$\mathbf{F_k = g \cdot (A_{2,k} - A_{1,k})^2 / (S_{2,k}^{\ 2} + S_{1,k}^{\ 2})} \qquad (3.7.6)$$

В уравнение (3.7.6) единствената разлика от известното отношение на Фишер е членът **g**, който е равен на $\mathbf{sign(A_{2,k} - A_{1,k})}$. Както ще бъде показано в дискусията по резултатите на проведената класификация, отчитането на този знак

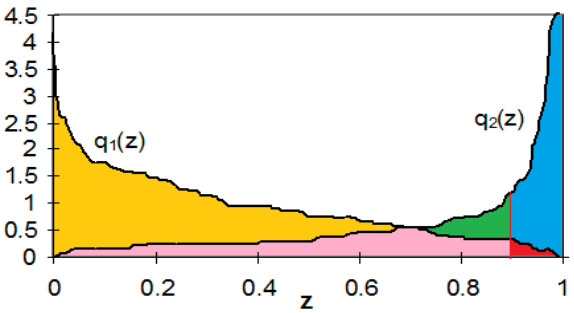
позволява отхвърлянето на редица неправилни спектро-структурни корелации от вида: "отсъствието на ивица в даден спектрален интервал показва наличието на дадена химична подструктура" и "присъствието на ивица в даден спектрален интервал показва отсъствието на дадена химична подструктура".

За всеки начален интервал програмата подрежда признаците от даден вид по намаляващ ред на техните F_k. Спектралният признак с максимално отношение F_k е най-подходящ за класификацията по съответния структурен елемент. Допълнително програмата избира други признаци, следвайки тяхното сортиране, но това са признаци, чиито интервали се припокриват слабо с интервалите на вече избраните признаци. Обикновено за даден начален интервал се получават от един до три признака от даден вид. Следва обучение на ИНМ (вижте т. 3.7.1) или изчисляване на коефициентите на функцията на ЛДА. Двете извадки се нормират по колоните (т.е. по спектралните признаци) по следната формула:

$$X_{j,k}^{new} = (X_{j,k} - A_k) / S_k, \qquad (3.7.7)$$

където A_k *и* S_k са средните стойности и стандартните отклонения на признаците **по всички спектри в обучаващата извадка**. Към двете извадки (матрици) се добавя по една колона за изходите от ИНМ: последните са избрани да са равни на 0.0 за съединенията от първия клас (непритежаващи дадения структурен елемент) и 1.0 - за тези от другия клас. Редовете на обучаващата извадка (вече матрицата $X_{J,K+1}$) се разбъркват и ИНМ се обучава до достигане на даден критерий за спиране на обучението. В резултат се изчисляват коефициентите на ИНМ, които дават най-добро разделяне на обектите от двата класа в обучаващата извадка.

В настоящата работа е използвана ИНМ с право разпространение на сигналите и обратно разпространение на грешките. Броят на невроните в единствения скрит слой, който дава най-добра класификация е различен за всеки един структурен елемент и е определен експериментално. Като критерий за спиране на обучението е използвана относителната промяна на средно-квадратичната грешка, **MSE**, на изходите от мрежата.



Фигура 3.7.4. Плътност на разпределение на изходите от невронната мрежа за обектите от първи и втори клас, $q_1(z)$ и $q_2(z)$ като функция от стойността на изхода z. Обектите са от тестващата извадка при създаване на класификатор за фенил.

Тъй като при обучение ИНМ се "старае" да преобразува пространството на признаците в две числа, 0.0 и 1.0, то плътностите на разпределение на изходите от ИНМ за двата класа биха изглеждали подобно на тези на фигура 3.7.4. Ако обект от пространството на признаците се "пропусне" през невронната мрежа (или функцията на ЛДА) и тя дава изход z. От този изход за всеки един спектър (набор от признаци) от тестващата извадка могат да се изчислят две условни вероятности - $P_1(z)$ и $P_2(z)$, които представляват вероятността обектът да е от първи клас при условие, че изхода от ИНМ $Out < z$, и съответно от втори клас при условие, че $Out > z$. Тези вероятности се изчисляват от площите под съответните криви

(тези площи съответстват на функцията на разпределение) и се приемат *за точност на класификацията* (*precision*).

Точната оценка на кривите от фигура 3.7.4, а от тях и на вероятностите, е затруднена от малкия брой на тестовите обекти. Ето защо ние извършваме директно оценка на тези вероятности по формулите:

$$P_1(z) = N_{1,1}/(N_{1,1} + N_{2,1})\ 100\% \qquad (3.7.10a)$$

$$P_2(z) = N_{2,2}/(N_{2,2} + N_{1,2})\ 100\% \qquad (3.7.10b)$$

където $N_{1,1}$, $N_{2,1}$, $N_{1,2}$ и $N_{2,2}$ са определен брой обекти от тестващата извадка, които са статистически оценки на площите под съответните криви: за $z = 0.9$ с използване на цветовете на фигура 3.7.4 може се каже, че тези числа са оценки на следните площи: $N_{1,1}$ – розовата и кафявата, $N_{2,1}$ – розовата и зелената, $N_{1,2}$ – червената, $N_{2,2}$ – синята и червената. Програмата изчислява $N_{k,m}$ чрез преброяване на обектите от клас k $(k = 1, 2)$, които дават изход от тази страна на z, където е целевият изход на клас m $(m = 1, 2)$. На фигурата $N_{1,2}$ означава броя обекти от първия клас, които дават изход по-голям от изхода на непознатия обект z.

Друга също важна характеристика на проведената класификация е броят обекти от първи и втори клас, които ще бъдат класифицирани с точност по-голяма от дадена прагова стойност, определена от потребителя - обикновено 90% или 95%. Този брой, изразен в проценти от броя на обектите на съответния клас се нарича *степен на класификация* (recall). Следвайки означенията на фигура 3.7.4, степените на класификация за съответните класове, R_1 и R_2 се дават с уравненията:

$$R_1(z) = N_{1,1}/N_1\ 100\% \qquad (3.7.11a)$$

$$R_2(z) = N_{2,2}/N_2\ 100\% \qquad (3.7.11b)$$

където N_1 and N_2 са съответно броят на обектите от първи и втори клас.

Видът на кривите от фигура 3.7.4 е различен, когато се прилага ЛДА. Тогава те изобразяват плътностите на разпределение на преобразуваните признаци (scores) на обектите от двата класа и разликата е, че кривите не са ограничени между 0 и 1 по абсцисата. Двете величини - точност и степен на класификация, обаче се изчисляват по същата схема и формули от нас при прилагането на ЛДА както тези, когато се използват ИНМ. В нашите разработки, същият подход за оценка на точността (достоверността) на подструктурите е приложен и (1) за метода на най-близките съседи (т. 3.5), както и (2) при интерпретационно търсене в спектрални библиотеки от напълно отнесени ^{13}C ЯМР спектри (т. 3.8).

3.7.3.1. Подбор на спектрални признаци от ИЧ спектри.

За изчисляване на *интервалните признаци* интервалът 4000 - 400 cm^{-1} е разделен на 256 интервала, чиито пореден номер k и дясна граница ν_2 са свързани с уравнение (3.7.6) и за всеки един интервал е приложена формула:

$$k = (6.0\ \nu_2^{1/2} - 120.0), \text{ закръглени до цяло число} \qquad (3.7.12)$$

Ширината на интервалите нараства с нарастване на вълновото число. Ако всички признаци бъдат използвани за обучение на ИНМ, то най-голямо влияние на класификацията ще имат признаците с най-голяма дискриминираща (класификационна) способност. Една оценка на тази способност е отношението на

Фишер, т.е. това е изразът (3.7.6), в който имаме **g = 1.0**. Ето защо за класификацията на дадена подструктура са генерирани всички 256 признака и от тях са избрани 20^{те} с най-голямо отношение на Фишер за обектите от двата класа в обучаващата извадка.

Броят на *експертните признаци* зависи от броя на характеристичните интервали на класифицираната химична подструктура. С характеристичните интервали са генирирани признаци от интервален и логаритмичен тип по уравнения (3.7.4) и (3.7.5). Първите два типа признаци са избрани от нас за сравнение на тяхната класификационна способност с въведените от нас оптимизирани експертни признаци. За целта за всяка химична подструктура от таблица 3.7.6 са създадени ЛДА класификатори с използване на трите вида признаци. Допълнително за всяка една подструктура са създадени ИНМ класификатори с използване на оптимизираните експертни признаци.

3.7.3.2. Надеждност на признаците.

Трите вида спектрални признаци се основават на различни концепции и допускания, които определят техния избор и начин на използване.

(а) Интервалните *признаци* с граници, изчислявани по уравнение (4.7.12) са въведени като алтернатива на използването на признаци с еднакви по големина спектрални интервали. Разширяването на интервалите с нарастване на вълновото число отразява факта, че в нискочестотната област на ИЧ спектър има повече и по-тесни ивици, както и че там положението на ивиците отразява в по-голяма степен химичното обкръжение на отделните структурни елементи, т.е. тази област е по-богата на структурна информация. За всяка една от химичните подструктури от таблица 3.7.6 ние сме избрали по 20 признака от този вид, най-добре разделящи обектите по класове в обучаващата извадка. Тези 20 спектрални признака са използвани при създаване класификатора от първия вид за съответната подструктура. Както беше споменато по-горе те са избрани по отношението на Фишер, което определя разделящата способност на спектралния признак без оглед на знака на разликата между средните стойности на признака за отделните класове. Например 14 от тези 20 признака за класификацията по метилова група са с отрицателно отношение $\mathbf{F_k}$.

Прегледът на **F**-отношенията на тези признаци за останалите подструктури също показва, че повечето от тях са отрицателни. Това на практика означава, че ако се проведе класификация по тези подструктури с използването на тези признаци класификационният алгоритъм ще използва признаците с най-голямо отношение на Фишер, като в това число попадат и някои от тези с отрицателно **F**-отношение, вижте таблица 3.7.4 от дисертацията. Тъй като признаците показват височината на най-интензивната ивица в даден интервал, то използваните спектро-структурни корелации за признаците с отрицателно **F**-отношение ще са от вида "ако *има* пик в даден интервал, то *нямаме* дадена подструктура" и "ако *няма* пик в даден интервал, то *имаме* дадена подструктура": това са две твърдения напълно противоречащи на теорията на ИЧ спектроскопия. На практика създадените 20 ЛДА класификатора с използване на тези признаци показват добра предсказваща способност, но тя е добра средностатистически само за тази

извадка от спектри и тези класификатори ще са напълно безполезни за произволно съединение.

Една възможна причина за тези лъжливи отрицателни спектро-структурни корелации е корелацията между наличието на редица структурни елементи в обучаващата извадка. Доколкото обучаващата (а и тестващата) извадка е случаен подбор от съответните класове на библиотеката от спектри, то тези отрицателни корелации са пряк резултат от състава на самата библиотека от спектри. Например в обучаващата извадка (250 + 250 обекта) за метиловата група в клас 1 (няма метил) 189 структури съдържат бензеново ядро, а в клас 2 - само 146. Броят на орто-, мета-, пара- и монозаместените бензени в съединенията от двата класа са 40, 14, 50, и 55, и 26, 8, 43 и 30, съответно. Един практичен начин да се избегнат тези нежелателни отрицателни спектро-структурни корелации е да се избират признаците по стойността на **F**-отношението от уравнение (3.7.6), както това е направено за третата група от спектрални признаци. Проблемът с лъжливите корелации е значително по-сериозен когато те са с положителен знак, т.е. в клас 2 по-често се среща друга подструктура отколкото в клас 1. В този случай единствено ограничаването на интервалите на спектралните признаци в характеристичните интервали на съответната класифицирана подструктура могат да доведат до избягване на голяма част от тези лъжливи корелации. Този подход е реализиран при втората и третата група от признаци.

(b) *Експертните признаци* по уравнения (3.7.4) и (3.7.5) за класификацията по дадена подструктура са генерирани с нейните характеристични интервали. Изчислените **F**-отношения за тях са с ниска, а в някои случаи и с отрицателна стойност. Една от причините за това е самата дефиниция на характеристичните интервали, които обхващат ивиците на подструктурата при почти всички съединения, които я съдържат. Това от своя страна прави интервалите достатъчно широки, че да съдържат ивици от други структурни елементи, което води до появата на ивици в съответния интервал и при обектите от клас 1 (подструктурата отсъства).

Повечето логаритмични експертни признаци имат отрицателни или положителни, но много ниски по стойност **F**-отношения. Причина за това е, че в повечето характеристични интервали има само една ивица на подструктурата, а логаритмичните признаци са различни от нула при наличието поне на две ивици в техния интервал.

Таблица 3.7.5. Експертните признаци, оптимални за класификацията на ИЧ спектри по 3 от подструктурите от таблица 3.7.6. Границите на спектралните интервали са в cm^{-1}.

метил			**метилен**			**бензен**		
F	тип	интервал	F	тип	интервал	F	тип	интервал
0.506	INT	2996 2944	1.017	INT	2984 2927	0.102	INT	3070 3055
0.236	INT	2887 2868	0.573	INT	2940 2839	0.331	INT	1617 1588
0.147	INT	1379 1357	0.112	INT	2861 2790	0.758	INT	1550 1471
0.545	L12	3015 2927	0.103	INT	1369 1364	0.260	INT	870 820
0.322	L12	2945 2868	2.257	L12	3007 2855	0.617	INT	838 739
0.292	L12	1404 1357	0.212	L12	1468 1367	0.353	INT	716 670
						0.426	L12	3136 3023
						1.672	L12	1616 1476
						0.410	L12	883 783
						0.725	L12	766 670

(с) Оптимизираните експертни признаци, генерирани за три от $20^{те}$ подструктури са дадени в таблица 3.7.5 (за $20^{те}$ – в дисертацията). Всички тези признаци имат положителни и високи **F**-отношения, които са винаги по-големи от **F**-отношенията на другите типове признаци за дадения структурен елемент: разбира се, това е следствие от начина на техния избор. Получените ЛДА тегла са положителни, или ако са отрицателни те са близки до нулата, т.е. те са статистически неотличими от нула.

Оптимизираните интервали на признаците напълно съответстват на експертните знания, натрупани в областта на отнасяне на ивиците в ИЧ спектри.

3.7.3.3. Ефективност на класификаторите на ИЧ спектри.

Приложеният класификационен метод с използването на бинерни класификатори изисква напълно различен подход за оценка на тяхната ефективност. Обикновено потребителят на една автоматична система очаква висока точност на извършваната интерпретация, а това от своя страна води до редица случаи, в които системата не дава решение, тъй като не е надмината праговата точност. Критерий за ефективността на такива системи е процентът на обектите, които ще бъдат класифицирани с точност по-голяма от праговата, т.е. за всеки отделен класификатор критерии за неговата ефективност са степените на класификация за клас 1 и 2 съответно, при определена точност на класификацията. В настоящата работа тази точност е приета за 90%, и сравняваните числови величини са степените на класификация при 90% точност. Допълнително за всеки един класификатор е определена точността на класификация за обектите от втория клас при степен на класификация, равна на 50%.

В таблица 3.7.6 от дисертацията са представени някои параметри на създадените класификатори (които са четири на брой за всяка една подструктура): броят на спектралните признаци, оптималният брой на скритите неврони за ИНМ, степените на класификация при 90% точност за двата класа и точността на класификация за обектите от втория клас при $P_2 = 50\%$. За повечето подструктури степените на класификация и критерия A_{50} са с високи стойности, с изключение на класификаторите за **–CH=CH–** и **–NH–**. Като правило от ИЧ спектър се доказва с по-голяма сигурност отсъствието на даден структурен елемент, отколкото неговото присъствие. Това се потвърждава и при сравнение на степените на класификация за двата класа, които за повечето подструктури са в отношение $R_1 > R_2$, с някои изключения.

За обективно сравнение на класификаторите, използващи различни видове признаци и различните класификационни методи (ЛДА или ИНМ) е приложен t-тест (paired t-test), с който са сравнени техните R_k $(k = 1, 2)$ и A_{50}. Данните от сравнението са представени в таблица 3.7.7. Степените свобода са 19 и при избрана статистическа сигурност $P = 90\%$ интегралната граница при двустранна постановка на задачата е $t(19,0.90) = 1.73$. Всяка една стойност на t-критерия от таблицата по-голяма от 1.73 показва превъзходство на признаците/метода от съответния ред над тези от съответната колона, а при $t_{кр} < -1.73$ имаме обратното. От таблица 3.7.7 се вижда, че t-критериите, изчислени с последните две стойности са приблизително еднакви. Сравнението на R_1 показва, че използването на оптимизираните експертни признаци от метода на ЛДА дава най-

добра класификация на обектите от клас 1; съответните стойности от таблицата са по-големи от интегралната граница. Използването на другите два вида признаци от метода на ЛДА дава статистически неотличими резултати.

Таблица 3.7.6. Характеристики на класификацията на три от подструктурите при използване на различни типове признаци и методи: NF - брой използвани признаци; NH - оптимален брой на скритите неврони на ИНМ; R_1, R_2 - степен на класификация при 90% точност за двата класа; A_{50} - точност на класификация за втория клас при R_2 = 50%.

подструкура 1	Признаци / методи 2	NF/ NH 3	R_1 4	R_2 5	A_{50} 6
	интерв./ЛДА	20	25.6[a]	31.2	83.1
метил	експ./ЛДА	6	45.2[a]	7.6	66.3
	опт. експ./ЛДА	6	56.4[a]	34.0	85.7
	опт. експ./ИНМ	6 / 16	68.4[a]	41.2	86.6
	интерв./ЛДА	20	3.6	44.0	86.6
метилен	експ./ЛДА	6	43.2	52.8	89.9
	опт. експ./ЛДА	6	63.6	39.6	88.2
	опт. експ./ИНМ	6 / 5	64.8	49.6	89.4
	интерв./ЛДА	20	52.8	46.8	89.4
бензен	експ./ЛДА	7	64.8	14.0	81.7
	опт. експ./ЛДА	10	73.6	42.8	88.5
	опт. експ./ИНМ	10 / 6	70.8	50.0	89.9

[a]степен на класификация при точност = 80%.

За обектите от клас 2 използването от ЛДА на интервалните и оптимизираните експертни признаци дава статистически неотличими резултати (съответните стойности 0.85 и 1.58 са по-малки от интегралната граница 1.73), а класификацията с експертните признаци дава значително по-лоши резултати (съответните четири стойности от таблица 3.7.7, b) и c) са по-големи по абсолютна стойност от 1.73).

Значително по-лошите резултати при използване на експертните признаци могат да се обяснят по следния начин. Както беше споменато по-горе, дефиницията на характеристичните интервали изисква те да са достатъчно широки, че да включват ивиците на съответния структурен елемент в почти всички съединения, които го съдържат. При тестване на изчисления модел това от своя страна води до появата на редица ивици в тези интервали, дължащи се на други подструктури от съединенията на тестващата извадка. По този начин се повишава числото на обектите от клас 1, които са класифицирани към клас 2, т.е. повишава се числото $\mathbf{N}_{1,2}$ в уравнение (3.7.10b). Съответно се намалява точността на класификация при обектите от клас 2, което води до по-ниската степен на класификация при изискваната от нас 90% точност. Подобрението на класификационната способност при използване на 20^те интервални признаци се дължи на техния избор от 256 признака от този вид, с изискването тези 20 признака да имат най-високата дискриминираща способност за обектите от обучаващата извадка. Но както беше изяснено по-горе, сред тях има признаци, съдържащи лъжливи спектро-структурни корелации, така че тяхното използване за класификация на непознат обект (спектър) не е оправдано. От друга страна

оптимизирането на границите на интервалите води до нови граници, които са един компромис между двете изисквания: интервалът да съдържа ивици на спектрите от клас 2 и интервалът да не съдържа ивици на спектрите от клас 1.

Таблица 3.7.7. Сравнение между класификациите с използване на ЛДА и 1) интервални, 2) експертни и 3) оптимални експертни признаци, и ИНМ и последния вид признаци. Сравнението се извършва с използването на t-тест: положителната му стойност показва предимството на класификацията от първа колона над тази от първи ред.

a) сравнение на степента на класификация за първи клас, $\mathbf{R_1}$

тип признаци/метод	интервални/ЛДА	експертни/ЛДА	опт. експертни/ЛДА
експертни/ЛДА	0.62		
опт. експертни/ЛДА	4.44	6.29	
опт. експертни/ИНМ	6.91	5.27	1.53

b) сравнение на степента на класификация за втори клас, $\mathbf{R_2}$

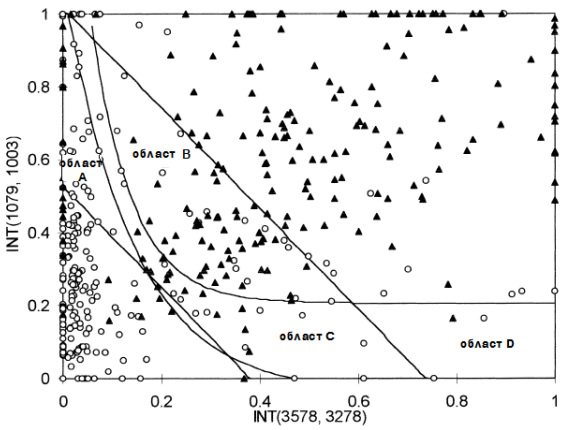
тип признаци	интервални/ЛДА	експертни/ЛДА	опт. експертни/ЛДА
експертни/ЛДА	-3.10		
опт. експертни/ЛДА	0.85	4.04	
опт. експертни/ИНМ	6.55	6.54	3.85

c) сравнение на стойностите на $\mathbf{A_{50}}$ от таблица 3.7.6

тип признаци	интервални/ЛДА	експертни/ЛДА	опт. експертни/ЛДА
експертни/ЛДА	-3.35		
опт. експертни/ЛДА	1.58	4.02	
опт. експертни/ИНМ	4.48	5.16	2.78

3.7.3.4. Сравнение на ИНМ с метода на ЛДА.

Оптимизираните експертни признаци са избрани за сравнение на ИНМ с метода на ЛДА. В таблица 3.7.7 са дадени съответните t-критерии. Двете стойности 3.85 и 2.78 показват, че ИНМ класифицира обектите от клас 2 по-добре от ЛДА. За клас 1 разликата ($\mathbf{t_{кр} = 1.53}$) е статистически отличима, със статистическа сигурност $\mathbf{P = 86\%}$. Действително 16 от стойностите за R_1 за ИНМ класификаторите са по-големи от тези за ЛДА класификаторите (вж. таблица 3.7.6). По добрата класифицираща способност на ИНМ в сравнение с линейните класификационни методи се обяснява с нелинейната разделяща повърхност, с която ИНМ разделят обектите от съответните класове. Този факт може да бъде илюстриран с ЛДА и ИНМ класификаторите за подструктурата първичен алкохол, чиито математически зависимости са разгледани подробно в дисертацията. Тъй като имаме само два признака, то уравнения на разделящите повърхнини могат да се изобразят като криви в равнината (A_1, A_2), определена от две перпендикулярни оси, на които са нанесени абсорбциите на ивиците в двата интервала - вижте фигура 3.7.5. Обектите от горната дясна част на фигурата се класифицират и от двата класификатора към втори клас (първичен алкохол), докато тези от долната лява част - към първи клас: така би разсъждавал и експертът-спектроскопист, защото да е съединението първичен алкохол, то трябва да има силни по интензитет ивици в двата спектрални интервала.



Фигура 3.7.5. Области в спектралното пространство, за които ИЧ спектри се класифицират по подструктурата "първичен алкохол" с точност, равна на 90%. Правите линии са за ЛДА класификатора, а другите за ИНМ класификатора. Съответните леви линии са за първи клас, а десните - за втори клас. Запълнените триъгълници изобразяват съединенията, които са първични алкохоли, а празните кръгчета – тези, които нямат CH_2OH групата.

Обектите от областите A и B на фигура 3.7.5 се класифицират от ЛДА класификатора с точност по-малка от 90%, докато ИНМ класификатор ги класифицира към съответния клас с точност по-голяма от 90%. Ето защо степените на класификация при 90% точност са съответно $R_1 = 62\%$ и $R_2 = 56\%$ за ЛДА класификатора, и $R_1 = 82\%$ и $R_2 = 84\%$ за ИНМ класификатора. Обектите от област C *не се класифицират* и от двата класификатора при прагова точност, равна на 90%, докато тези от област D - *само* от ИНМ класификатора. Последният факт също показва предимството на ИНМ, защото спектрите, които се изобразяват в областта D близо до абсцисата притежават слаба по интензитет ивица в интервала 3578 - 3278 cm^{-1}, което не е характерно за ν(O-H) ивицата на първичните алкохоли.

3.8. Интерпретационно търсене в библиотеки от напълно отнесени ^{13}C спектри.

Този вид библиотечно търсене намира подструктури, които са част от библиотечните структури и имат сигнали, които са близки до част от сигналите на търсения спектър. За разлика от търсенето за идентификация и по подобие, интерпретационното библиотечно търсене (ИБТ) не съставя хит-списък на съединения, които имат подобни спектри със спектъра на неизвестното съединение, а използва допускането, че ако два спектъра имат обща част, то съответните структури имат също обща част. ИБТ е замислено да служи като самостоятелно приложение в спектроскопията, но най-успешно би се прилагало като модул към комплексна система за разкриване на структурата на органични съединения с помощта на ЯМР спектри. При коректни резултати от ИБТ се повишава количеството информация, генерирана от модула за интерпретация и съдържаща се в ограниченията към структурния генератор, а това увеличава бързината на структурното генериране и се намалява броят на генерираните структури.

При интерпретационното търсене няма ограничения за размера или вида на предсказаните подструктури: двете са ограничени само чрез структурите на съединенията в библиотеката. Второ, процедурата не се влияе от цялостното структурно подобие между референтното съединение и непознатото. Този контраст с търсенето по подобие е главната разлика между търсенето по подобие и интерпретационното библиотечно търсене. Трето, на изхода на търсенето всяка предсказана подструктура е представена като включена в структурата на

референтното вещество, от което тя е възпроизведена. Това от своя страна, при визуално преглеждане на референтните структрури, може да предложи обяснение на една или повече от предсказаните подструктури, а може също да разкрие нещо повече за класа на веществата, към който непознатото принадлежи. Във варианта, предложен от нас, извличаните подструктури са с отнесени сигнали на всеки един от въглеродните атоми и това ускорява неимоверно структурната генерация, за разлика от стария алгоритъм, чиито извлечени подструктури са без сигналите на неизвестния спектър, което на практика забавя структурната генерация.

3.8.1. Интерпретационно търсене – спектри, подструктури и оценка за тяхната надеждност.

3.8.1.1. Търсене по сигнали в инвертни файлове.

Интерпретационно библиотечно търсене използва ^{13}C-ЯМР спектъра, който се дава като стойности на химичното отместване на сигналите и тяхната мултиплетност, която се дължи на съседните водородни атоми. Параметрите, които ограничават алгоритъма за търсене, са: (1) неопределеността (толерансът) на съвпадане на сигналите (**Tol**), изразена в милионни части (ppm), и (2) минималният брой на въглеродни атоми в предсказаната подструктура (**m.n.c.**).

За всяко химическо отместване (**Sh**) от непознатия спектър интервалът на съвпадение на стойностите на химичното отместване в референтния спектър се определя като **Sh ± Tol**. В резултат на инвертно търсене в индексни файлове, само тези референтни структури са отбелязани като вход за следващата стъпка в алгоритъма, които имат брой сигнали в референтния спектър, съвпадащи с непознатите сигнали, и брой сигнали в непознатия спектър, съвпадащи с тези на референтния, по-голям или равен на предварително определен праг, **m.n.c**.

3.8.1.2. Извличане (генериране) на подструктурите.

Всяка от референтните структури, която е била отбелязана в предишната стъпка се обработва по този алгоритъм. Целта на този алгоритъм е да се намери една или повече свързани подструктури, които са част от референтната структура, и чиито въглеродни атоми са свързани помежду си с директна химическа връзка или през химически връзки и хетероатом. Също така референтните сигнали на въглеродните атоми в разширената подструктура трябва да съвпадат еднозначно (one-to-one match) със сигналите на непознатия спектър. Всички тези условия правят субструктурното разширяване изключително сложно. Обикновено, един сигнал от непознатия спектър може да съвпадне с повече от един сигнал от референтния. Включването на всички такива референтни атоми в подструктура не е основателно, защото не може да има един сигнал от непознатия спектър, присвоен (matched) към няколко атома от подструктурата – това не е само некоректно от спектроскопска гледна точка, но и структурният генератор **Sesami** няма да даде решения. Отнасянето на непознатите към референтните сигнали се избира, така че да дава най-малко средно квадратично отклонение (root mean squared deviation, **RMSD**); това отнасяне е наречено отнасяне, с минимално средно квадратично отклонение. Параметърът **sRMSD** е важна променлива във всяка една функция за определяне на надеждността (reliability) на субструктурата, защото

колкото то е по-малко, толкова по-близко е химичното обкръжение на подструктурата в референтната структура до това в неизвестната подструктура и затова вероятността подструктурата да е вярна е по-голяма. В резултат подструктурата се представя заедно с едно-към-едно отнесените непознати сигнали и по-късно те се използват за определяне на няколко други параметъра, характеризиращи подструктурата.

3.8.1.3. Оценка на надеждността на подструктурите.

Потребителят на ИБТ очаква генерираните подструктури да присъстват в непознатата структура. На практика някои от тях са неверни, т.е. те не присъстват в непознатата структура. Всяка генерирана подструктура може да бъде описана с набор от нейни собствени параметри и някои от тези параметри или комбинация от тях могат да бъдат използвани за сортиране на подструктурите, така че първите няколко от тях да са верни. Принципите на статистическия анализ дават база за намирането на подходяща вероятностна функция за оценка на верността на получените подструктури и нейното оптимизиране, така че тя има максимална ефективност: под ефективност тук се разбира за дадена точност на класификация функцията да класифицира по-голям брой верни подструктури, т.е. по-висока *степен на класификация* (recall), вижте т. 3.7.3. Статистическият подход има две изисквания: (1) голям списък от подструктури, предсказани чрез алгоритъма, за всяка от които е известно дали е вярна или невярна, и (2) набор от параметри (променливи), характеризиращи предсказваните подструктури, които влияят върху предсказаната вярност. Обикновено се използва многопроменлива логистична регресия (logistic regression analysis, **LoRA**), в която изходите са ограничени между единица (вярна подструктура) и нула (невярна такава). Изкуствените невронни мрежи (ИНМ) се използват като алтернатива на **LoRA** за оценяването на вероятностната функция на подструктурите.

Създаването на вероятностна функция за надеждност на подструктурите в настоящата работа е извършено за голямата библиотека **LAST** и са тествани няколко комбинации от параметрите на подструктурите с цел последните да бъдат сортирани по реда на тяхната надеждност (вярност).

Подструктурни параметри. Те са дефинирани (съставени) на основата на химическата интуиция и опита на спектроскопистите относно приложимостта им по проблем, поставен с въпроса: дали предсказаната структура е вярна или невярна? Това ще рече, че техният химически и/или спектроскопски смисъл е известен предварително и влиянието им върху надеждността на подструктурите е донякъде изяснено. Двата хистограмни параметъра отчитат донякъде, че колкото един сигнал е по-разпространен в библиотеката, толкова по-често се появява при съвпадането на химичните отмествания и е по-вероятно да имаме подструктура с него. Всички тези параметри са дадени в таблица 3.8.2.

Въпреки големия брой на тези променливи, 23, за статистическа оценка на функцията на надеждност бяха генерирани извънаредно голям брой подструктури, и затова в крайния стадий на изследванията бяха въведени нови параметри, наречени от нас скалирани.

Таблица 3.8.2. Описание на подструктурните параметри (шест са изпуснати за краткост).

1. uNA: брой на атомите в непознатото съединение; 2. uNS: брой на синглетите в спектъра на непознатото съединение; 5. uNQ: брой на квартетите в спектъра на непознатото съединение;
6. rNA: брой на атомите в библиотечното (референтното) съединение; 7. rNS: брой на синглетите в спектъра на библиотечното съединение; 10. rNQ: брой на синглетите в спектъра на библиотечното съединение;
11. sNA: брой на атомите в подструктурата; 12. sNB: брой на връзките в подструктурата; 13. sNS: брой на синглетите в спектъра на подструктурата; 16. sNQ: брой на квартетите в спектъра на подструктурата;
17. sRMSD: минималното средно-квадратично отклонение между химичното отместване на сигналите на неизвестното съединение и това на тези в подструктурата; 18. sFV: броят на свободните валенции на подструктурата; 19. sLO: броят на библиотечните съединения, които съдържат подструктурата; 20. sRS: броят на генерираните подструктури при търсене с определена неопределеност; 21. sSO: броят на присъствията на подструктурата в другите подструктури, генерирани със същата неопределеност;
22. sHS – хистограмна променлива (за нейното описание вижте текста в дисертацията); 23. sIHS - обратна хистограмна променлива (за нейното описание вижте текста в дис.);

Скалирани параметри. Те са дадени в таблица 3.8.3 и представляват отношения на параметрите от таблица 3.8.2 или тяхно нормиране по други параметри на подструктурата – например по броят въглеродни атоми в нея. Общият брой на параметрите от таблици 3.8.2 и 3.8.3 е 48, което голямо число не позволява използването на всичките параметри за всички извадки от подструктури.

Таблица 3.8.3. Скалираните входни параметри, използвани като променливи на функцията за надеждност. Означения: **sNC** - брой на въглеродните атоми в подструктурата, другите параметри са описани в таблица 3.8.2.

Общи скалирани параметри	**Скалирани хистограмни параметри**
1. sRelFV = sFV / sNB 2. sLibFreq = sSO / sLO 3. sSubFreq = sSO / sRS	4. sHSC = sHS / sNC 5. sIHSC = sIHS / sNC
Структурни скалирани параметри, 1 тип	**Структурни скалирани параметри, 2 тип**
6. rRelSize = rNA / uNA 7. rRelSings = rNS / (uNS + 1) 8. rRelDoubs = rND / (uND + 1) 9. rRelTrips = rNT / (uNT + 1) 10. rRelQuars = rNQ / (uNQ + 1) 11. sRelSize = sNA / uNA 12. sRelSings = sNS / (uNS + 1); 13. sRelDoubs = sND / (uND + 1); 14. sRelTrips = sNT / (uNT + 1); 15. sRelQuars = sNQ / (uNQ + 1);	16. rRelSize = uNA / rNA 17. rRelSings = uNS / (rNS + 1) 18. rRelDoubs = uND / (rND + 1) 19. rRelTrips = uNT / (rNT + 1) 20. rRelQuars = uNQ / (rNQ + 1) 21. sRelSize = sNA / rNA 22. sRelSings = sNS / (rNS + 1); 23. sRelDoubs = sND / (rND + 1); 24. sRelTrips = sNT / (rNT + 1); 25. sRelQuars = sNQ / (rNQ + 1);

Като изходен параметър или т.н. *зависима променлива* се използват стойностите 1 или 0, за вярна или невярна подструктура, съответно.

3.8.1.4. Отнасяне на сигналите на неизвестния спектър към въглеродните атоми в подструктурите.

Съвременните структурни генератори използват 1D- и 2D-ЯМР спектри, без които разкриването на структурата на големи съединения е невъзможна, въпреки структурната информация, получавана от другите спектрални методи. При програмата **Sesami** присъствието на даден структурен фрагмент, който се явява един вид ограничение (constraint), се проверява още в процеса на генериране на структурата, преди тя да е получена напълно – това се описва на английски с термина „prospectively", като се прави разлика с проверката в крайния резултат, наречено „retrospectively". Този вид проверка налага изискването всеки един въглероден атом в структурния фрагмент да има асоцииран (присвоен) поне един сигнал от ^{13}C-ЯМР спектъра на неизвестното съединение. В програмата **Sesami** не пречи да имаме няколко неизвестни сигнала, асоциирани към един въглероден атом, стига от тези множества да може да се определи едно-към-едно съответствие между сигналите в непознатия спектър и въглеродните атоми във фрагмента.

От няколко възможни схеми ние избрахме към даден въглероден атом да се асоциират всички сигнали от спектъра на неизвестното съединение с дадена мултиплетност, появяващи се в даден интервал около съответния референтен сигнал, <u>като този интервал е с променлива ширина</u>, която зависи от разстоянието на въглеродния атом до най-близката свободна валентност.

3.8.2. Структура на библиотеката от отнесени ^{13}C-ЯМР спектри.

В тази точка в дисертацията е подробно описана структурата на библиотечните файлове и форматът на запис на структурата с отбелязване на обкръжението на атомите, ароматността и тавтомерността на химичните връзки.

3.8.3. Функция на надеждност на подструктурите, генерирани при интерпретационно търсене в библиотека от отнесени ^{13}C-ЯМР спектри.

Такава функция с използване на всички параметри от таблици 3.8.2 и 3.83 е получена само за голямата библиотека, **LAST**, докато подходът приет за спектралната библиотека **PhyChem** е различен – вижте дискусията в т. 3.8.6. Една от причините е, че тази дейност е изключително ресурсоемка (в компютърно време и дейности на изследователя) и при промяна или създаване на спектрална библиотека трябва да се създаде нова функция на надеждност.

3.8.3.1. Проблемите с функцията на надежност и някои други неясноти.

В дисертацията е дадена история на прилагане на различни подходи (неуспешни по една или друга причина), която история е не само поучителна, но и разкрива в голяма степен проблемите с интерпретационното библиотечно търсене и самата същност на метода.

Всички тези проблеми наложиха нов подход, който се описва така:

- всеки спектър от библиотеката се потърсва в нея, като се изключва неговата референтна структура;

- извършват се гореописаните търсения при 21 различни стойности на толеранса: 0.0, 0.1, ... 2.0 ppm.

- при всяка една стойност на толеранса се регистрират само подструктурите с максимално отклонение, равно на използвания толеранс;

- за всяка една от тези 21 извадки от подструктури се изчисляват техните параметри от таблици 3.8.2 и 3.8.3 и се съставят 21 отделни извадки от параметри и изходни променливи, с които се изчисляват 21 различни функции на надеждност;

- подструктурата се счита за вярна, ако се съдържа в „неизвестната" подструктура, като не се взимат под внимание отнесените сигнали в минимално sRMSD;

- допълнително са проведени статистики за определяне на отнасянето на сигналите на „неизвестния" спектър към въглеродните атоми в генерираната подструктура при ИБТ.

В резултат на прилагане на **третия подход** бяха генерирани 16 690 896 подструктури, от които 6 890 800 са верни – таблица 3.8.7. Както се вижда от таблицата процентът на верните подструктури намалява при увеличаване на неопределеността на сравнение на сигналите, **Tol**. От 38 225 спектъра в библиотеката, 37 866 са генерирали при ИБТ поне една подструктура – това е значителна част (99%) от тях.

Tol (ppm)	NS	NCS	PCS (%)	NUC
0.0	74,199	74,055	99.8	25,279
0.1	78,930	77,896	98.7	8,220
0.5	309,536	265,738	85.9	15,122
1.0	605,186	332,017	54.9	20,202
1.5	1,212,□08	422,370	34.8	23,822
2.0	2,044,446	524,550	25.7	26,202
Общо	**16,690,896**	**6,890,800**	**41.3**	**37,866**

Таблица 3.8.7. Брой на генерираните субструктури (**NS**) като функция от неопределеността на сравнение на сигналите (**Tol**). Означения: NCS – брой на коректните подструктури; **PCS** – процент на коректните подструктури; **NUC** – брой на „неизвестните" спектри, които са генерирали поне една подструктура. (Пълната таблица за всички 21 толеранси е дадена в дисертацията)

3.8.3.2. Регресионен модел на функцията на надеждност.

Първоначалното разбиране в групата на проф. Мънк бе да се използват логистична регресия (**LoRA**) или изкуствените невронни мрежи (ИНМ) и изходът от тях да се приеме за вероятността подструктурата да присъства в „неизвестната" структура. Оказа се, че за практически цели, при работа с извадка от 600 000 обекта и 48 променливи, понякога програмата **SAS** „увисва", когато е стартирана с опцията избор на коефициенти, особено ако се работи с регресионен модел с кръстосани членове. Във всички случаи моделът на ИНМ показва по-добри степени на класификация от **LoRA**. Въпреки значителните подобрения на степента на класификация, при големи извадки (при големи толеранси) и ИНМ не дават добри резултати, ето защо се наложи промяна на начина на обучение на ИНМ – използване на различни тегла при разпространение на грешката, в зависимост от стойностите на целевия изход, 1 и 0.

Ако се провежда класификация с ИНМ между два класа (както е поставена задачата тук) и се иска ИНМ да дава много надеждна класификация само за обектите на единия клас, то може обучението на ИНМ мрежа да се промени с включването на различни тегла при разпространение на грешката, $(T_k - Y_k)$, така както е във формула (3.8.7).

$$\delta_2 = (T_k - Y_k)\ Y_k\ (1 - Y_k)\ W_T, \qquad (3.8.7)$$

където W_T са въведените тегловни коефициенти (тегла, weights), които са различни за двата класа – зависят от целевия изход, T_k: подразбира се от поставената задача, че $W_1 < W_0$.

Нашите теоретични разглеждания показаха, че вместо сумата S във формула (3.8.5) от дисертацията се минимизира сумата S′ в уравнение (3.8.8).

$$S'^2 = \Sigma\ (T_k - Y_k)^2\ W_T\ /\ K \qquad (3.8.8)$$

където T_k и Y_k са целевия и реалния изход от ИНМ за обекта k, а сумата е по всички обекти, $k = 1, 2, \ldots K$, в обучаващата извадка.

На фигура 3.8.2 в дисертацията е показана зависимостта на степента на класификация (recall) за три точности на класификация, R = 90%, 95%, 99%, и за три от 21^ве извадки.

3.8.3.3. Изчисляване на функцията на надеждност.

Както се спомена по-горе бяха съставени 21 функции на надеждност за 21 стойности на максималното отклонение, което съвпада с толеранса на търсене: 0.0, 0.1, ... 2.0 ppm. Всички спектри от библиотеката бяха потърсени в цялата библиотека при тези стойности на толеранса, но бяха съставени три извадки: обучаваща извадка (ОИ), тестваща извадка (ТИ) и валидираща извадка (ВИ), всяка от тях с размер от почти 1/3 от броя библиотечните спектри – 12 741. С ОИ бе извършвано обучение на ИНМ. С ТИ – се изчисляваха статистиките за точността и степента на класификация по уравнения (3.7.10b) и (3.7.11b), съответно. ВИ бе използвана за независима проверка на получените зависимости от проведените обучения и статистики. Изборът на променливи бе направен с „проба и грешка", и с отчитане на големината на съответната ОИ и при следене на обучението на ИНМ. В таблици 3.8.8 и 3.8.9 в дисертацията са дадени съответно избраните входни променливи и оптималният брой на скритите неврони.

3.8.3.4. Характеристики на функцията на надеждност.

Статистическите характеристики на 21^ве функции на надеждност, които заедно съставят общата функция на надеждност са дадени в таблица 3.8.10 (тук е съкратен вариант). В таблицата в колони 7 – 18 са дадени статистически оценки за функциите на надеждност при точности на класификация 90, 95 и 99% за тестващата извадка – тези стойности са означени с P_T (от test или threshold) и имат значението на прагови стойности. ИНМ се обучава с обектите от ОИ, но точността на класификация като функция от изхода от ИНМ се изчисляват с обектите от ТИ.

Таблица 3.8.10[a)]. Характеристики на съответните извадки от данни, с които са разработени функциите, които оценяват надеждноста на подструктурите. Означения в заглавния ред/колона: $\mathbf{P_T}$ – точност на класификация; **S** – извадка (**s**ample); **L** – обучаваща извадка (**l**earning set); **T** – тестваща извадка (**t**est set); **V** – валидираща извадка (**v**alidation set); **A** – цялата извадка (**a**ll); **Tol** – неопределеност при библиотечното търсене (**tol**erance); **NO** – брой обекти в съответната извадка (number of objects); **NC** – брой коректни подструктури (number of correct); **C** – процент на коректните подструктури, което е степен на класификация (**r**ecall); **NU** – брой на „непознатите" спектри, които са генерирали поне една подструктура (**n**umber of **u**nknowns); $\mathbf{P_E}$ / % – оценка на точността на класификация. (Това е съкратена таблица)

						P_T = 90 %				P_T = 95 %				P_T = 99 %			
Tol / ppm	**S**	**NO**	**NC**	**C/%**	**NU**	**NC**	**C / %**	**NU**	**P_E / %**	**NC**	**C / %**	**NU**	**P_E / %**	**NC**	**C / %**	**NU**	**P_E / %**
0.0	L	24 434	24 393	99.8	8 382	-	-	-	-	-	-	-	-	-	-	-	-
	T	24 562	24 515	99.8	8 402	-	-	-	-	-	-	-	-	-	-	-	-
	V	25 203	25 147	99.8	8 495	-	-	-	-	-	-	-	-	-	-	-	-
	A	74 199	74 055	99.8	25 279	-	-	-	-	-	-	-	-	-	-	-	-
0.1	L	26 576	26 386	99.3	2699	-	-	-	-	-	-	-	-	-	-	-	-
	T	25 573	25 295	98.9	2721	-	-	-	-	-	-	-	-	-	-	-	-
	V	26 781	26 215	97.9	2800	-	-	-	-	-	-	-	-	-	-	-	-
	A	78 930	77 896	98.7	8 220	-	-	-	-	-	-	-	-	-	-	-	-
0.5	L	107 206	91 629	85.5	5 009	90 043	98.3	4 973	88.5	84 792	92.5	4 548	94.7	60 643	66.2	2 958	99.5
	T	99 521	86 374	86.8	5 055	84 765	98.1	5 006	90.0	79 339	91.9	4 551	95.0	57 039	66.0	2 910	99.0
	V	102 809	87 735	85.3	5 058	86 177	98.2	5 016	88.4	80 344	91.6	4 583	93.4	56 289	64.2	2 935	98.7
	A	309 536	265 738	85.9	15 122	260 985	98.2	14 995	-	244 475	92.0	13 682	-	173 971	65.5	8 803	-
1.6	L	458,265	148,555	32.4	8,117	69,392	46.7	3,271	91.1	43,215	29.1	1,898	96.6	15,022	10.1	738	99.7
	T	452,013	146,887	32.5	8,134	66,254	45.1	3,279	90.0	41,385	28.2	1,875	95.0	14,401	9.8	680	99.0
	V	456,705	146,199	32.0	8,112	64,933	44.4	3,302	89.5	40,634	27.8	1,880	95.4	14,052	9.6	679	99.1
	A	1,366,983	441,641	32.3	24,363	200,579	45.4	9,852	-	125,234	28.4	5,653	-	43,475	9.8	2,097	-
1.7	L	515,741	157,301	30.5	8,307	65,818	41.8	2,499	91.3	42,605	27.1	1,455	97.0	7,482	4.8	330	99.9
	T	503,847	152,796	30.3	8,341	60,810	39.8	2,481	90.0	38,025	24.9	1,447	95.0	6,666	4.4	289	99.0
	V	513,049	154,101	30.0	8,282	60,847	39.5	2,489	89.3	38,058	24.7	1,455	94.7	6,831	4.4	299	99.8
	A	1,532,637	464,198	30.3	24,930	187,475	40.4	7,469	-	118,688	25.6	4,357	-	20,979	4.5	918	-
1.8	L	571,141	164,048	28.7	8,415	69,761	42.5	2,851	92.4	46,753	28.5	1,747	97.3	19,013	11.6	661	99.6
	T	554,416	160,196	28.9	8,448	64,182	40.1	2,919	90.0	42,344	26.4	1,733	95.0	17,953	11.2	627	99.0
	V	569,517	159,303	28.0	8,468	63,381	39.8	2,923	89.0	41,412	26.0	1,767	94.4	17,945	11.3	617	98.3
	A	1,695,074	483,547	28.5	25,331	197,324	40.8	8,693	-	130,509	27.0	5,247	-	54,911	11.4	1,905	-
2.0	L	684,443	178,774	26.1	8,714	76,603	42.8	2,791	93.5	58,769	32.9	1,895	97.5	31,525	17.6	1,026	99.6
	T	669,774	171,313	25.6	8,744	68,484	40.0	2,782	90.0	51,258	29.9	1,853	95.0	25,495	14.9	961	99.0
	V	690,229	174,463	25.3	8,744	70,653	40.5	2,682	89.4	53,572	30.7	1,835	94.0	28,197	16.2	985	98.2
	A	2,044,446	524,550	25.7	26,202	215,740	41.1	8,255	-	163,599	31.2	5,583	-	85,217	16.2	2,972	-

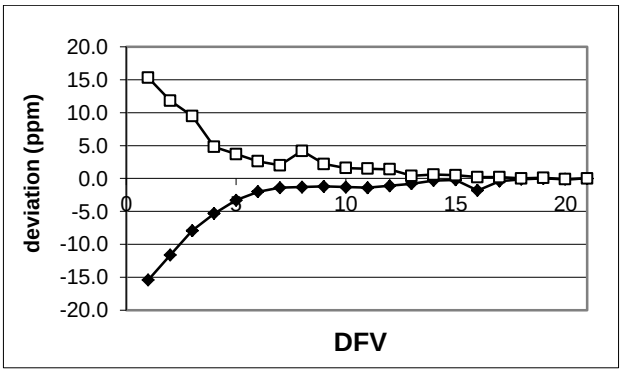
.

За стойностите на изхода от ИНМ, които дават P_T = 90, 95 и 99% се изчисляват съответните стойности на точността за другите две извадки, **ОИ** и **ВИ**. Вижда се, че тези стойности (в колони 10, 14 и 18) за валидиращата извадка са много близки до праговите стойности P_T, което означава, че ИНМ не само се обучава, но и предсказва правилно за обекти, невключени в **ОИ** и **ТИ**.

Главният извод, който може да се направи от таблицата, е че степента на класификация е много добра – в работата за мас-спектри Вармуца и Вертер пишат [**A4**], че класификатор, който има 30% степен на класификация (recall) при точност на класификация **P** = 90% се счита за полезен (използваем, "potentially useful"). От таблицата се вижда, че за **P** = 90% степента на класификация е над 38%. Това, заедно с факта че 99% от съединенията в библиотеката са дали поне една генерирана подструктура при ИБТ прави настоящите изследвания приложими в спектроскопската практика. Също така, степента на класификация е много еднаква за трите извадки, което отново потвърждава извода в предния параграф.

3.8.4. Параметри при отнасяне на сигналите на неизвестния спектър към въглеродните атоми в подструктурите.

Възприета беше идеята да се използват интервали, центрирани около сигналите в референтната структура и променлива дължина, която зависи от разстоянието до свободните валентности в генерираната подструктура. За целта за всяка една от 21^{ве} валидиращи извадки беше извършено субструктурно търсене на верните подструктури в съответната „непозната" структура. Бяха отбелязвани всички съответствия (matches) между подструктурата и „непозната" структура и отклоненията в сигналите бяха регистрирани с отбелязване и на дистанцията до най-близката свободна валентност, **DFV**, която се има стойност 1, ако въглеродният атом има свободна валентност (FV), с 2 - ако той няма FV, но негов съсед има и т.н. Тук при подструктурното търсене е задължително да се използват обкръженията на атомите, дискутирани в 3.8.2 в дисертацията, иначе ще се получат неверни интервали.



Фигура 3.8.3. Първи и 99^{ти} процентил на отклоненията между сигналите на неизвестния спектър и подструктурата като част от референтната структура. c) – за 2.0 ppm. Означения: ♦1^{ви} процентил; □ 99^{ти} процентил. Дадена е само част c) – за толеранс от 2.0 ppm.

Беше извършена статистика и се определи първият и последният процентил на данните за всяка една стойност на **DFV**. За три от валидиращите извадки те са дадени на фигура 3.8.3 в дисертацията: тук това е направено само за Tol = 2.0 ppm. Вижда се, че ширината на интервала между двата процентила намалява с повишаване на разстоянието до свободните валентности. Това намаляване е следствие от това, че изясненото химично обкръжение на атомите нараства, и

също така напълно съвпада с идеята за сигналите на въглеродните атоми в центрирани подструктури. Тези два процентила определят интервал в който попадат 98% от данните и затова могат да се вземат като оценка на търсения около референтния сигнал интервал, който да се използва за отнасяне на неизвестните сигнали към въглеродните атоми в подструктурата. За практически приложения беше решено тези интервали да се направят напълно симетрични и незначително да се разширят, и да се използват само данните за валидиращата извадка с **Tol** = 2.0 ppm.

3.8.5. Приложение на метода за търсене на спектри на природни съединения.

Разкриването на структурата на природни съединения е значително по-сложна задача от тази по потвърждаване на структурата на съединения, получени при химичен синтез – вижте т. 1.2. В тази част е описана практическа проверка на използването на подхода за ИБТ към спектрите на природни съединения, изолирани от растения.

През 2012 г. групата по молекулна спектроскопия започна сътрудничество с д-р Петко Бозов, който се занимава с изолиране на природни съединения и разкриване на тяхната структура. Ето защо представлява интерес дали методът на интерпретационното търсене би бил от помощ при тази дейност. За проверка на това предположение, десет спектъра на природни съединения (таблица 3.8.12), изолирани от д-р Петко Бозов и съавтори бяха потърсени в библиотеката **LAST** и получените резултати обработени по същия начин, описан в тази част.

За десетте спектъра се генерират от 154 до 5 878 подструктури (1 720 средно), с вероятност на първата подструктура в списъка от 95.5% до 100% (98.8% средно). Само в един от случаите, първата подструктура е невярна; нейната вероятност е 95.5%, което е под 97%, праговата стойност дискутирана по-горе. Останалите 9 верни подструктури са предсказани с вероятност в интервала 97.1% до 100% (99.1% средно). Може да се види, че вероятността на всички верни подструктури надвишава праговата стойност от 97%, т.е. имаме 100% степен на класификация на верните подструктури. Оценката на праговите стойности на надеждността също е близка до съответната прагова.

Структурите на тестваните съединения са дадени на в първа колона на таблица 3.8.12 в дисертацията; във втора колона са дадени първите подструктури от списъците, получени при интерпретационното търсене в библиотеката **LAST**. Подструктурите са представени, вмъкнати (embedded) в съответната структура на референтното съединение. Размерът на десетте тествани структури (изчислен като брой тежки атоми) варира между 25 и 43 (34 средно) и тези на 10^{te} предсказани на първо място в списъка подструктури от 6 до 20 (9 средно). Размерът на деветте верни подструктури е между 17% и 56% от съответния размер на „неизвестната" структура. Тези проценти на практика означават, че подструктурата определя (фиксира) голямо количество структурна информация от съответната структура, още повече, че наличието на кислороден атом в подструктурата в две трети от случаите увеличава нейното информационно съдържание.

Внимателното разглеждане на извлечените подструктури разкрива някои специфики на интерпретационното библиотечно търсене. Първо, ако в

структурата присъства странична верига, тази молекулна част е предсказана като подструктура с висока вероятност както е в случая с **скуталпини А** и **Е** (става въпрос за фрагмента 2-**метилбутанов естер**) и **спленолиди А** и **В** и **сплендидин** (фрагментът, който съдържа 3-**фуранил**). Второ, тъй като алгоритъмът изгражда подструктури от атоми на референтната структура, които не са магнитно еквивалентни, получената като краен резултат подструктура от спектъра на **скуториенталин Е** съдържа само половината бензеново ядро. Въпреки това опитният химик ще заключи че фрагментът **естер на канелената киселина** е част от търсената структура. Трето, ако референтното и неизвестното съединение имат подобна структура, предсказаната подструктура е голяма. Такъв резултат е получен за **формил метил олеанолат** – предсказаната подструктура съдържа 20 тежки атома, които съставят 56% от всички атоми на „неизвестната" структура. Както може да се види, „неизвестната" и референтната структури се различават само на две места, които (заедно с близкото си обкръжение) са изключени от референтната структура и не се съдържат в генерираната подструктура.

3.8.6. Директно сортиране на подструктурите по тяхната надеждност.

Функцията, която оценява подструктурите по тяхната коректност е извлечена при интерпретационно търсене на спектри от библиотеката **LAST** в тази същата библиотека. Параметрите, които са вход във функцията бяха дадени в таблици 3.8.2 и 3.8.3 и анализът на техните стойности показва, че те са зависими директно от размера на библиотеката. Това се вижда най-явно при параметър **sLO**: ако чисто формално удвоим спектралната библиотека (същите спектри, но записани два пъти), то този параметър при всички подструктури ще се удвои. От същия недостатък „страдат" и някои от другите параметри. Параметри, които не зависят директно от размера на библиотеката са параметри 1 – 18 от таблица 3.8.2 в дисертацията и всички параметри от таблица 3.8.3 без споменатите **sHSC** и **sIHSC** (№ 4 и 5). Разбира се, всички параметри без първите пет от таблица 3.8.2 зависят директно от състава на библиотеката, но се очаква, че при малки промени на спектралната библиотека тяхната промяна няма да е значителна. Но „преминаването" от една голяма библиотека (каквато е **LAST**) към малка библиотека (от 1000 спектъра – **PhyChem**) е свързано с драстична промяна на размера (над 38 пъти), както и в разлика в състава на библиотеките. А тази промяна в стойностите на параметрите би трябвало по математически съображения да направи неизползваема функцията на надеждност, т.е. да я лиши от свойството „преносимост".

За изследване на преносимостта на функцията на надеждност бяха съставени две извадки от по 100 спектъра, които не присъстват в библиотеките **LAST** и **PhyChem**. Тези двеста спектъра са на вещества, изолирани от растения; спектрите и съответните структури са публикувани в списание Phytochemistry (2001-2002 година, томове 58-60). Тъй като от дадена публикация са взети по няколко ^{13}C-ЯМР спектри, то те алтернативно по техния ред са представени в обучаващата извадка или в тестващата извадка – по този начин се избягва донякъде нарочната подредба на подобни по структура съединения, която се прави от авторите с цел прегледност на публикацията. За спектри от валидиращата извадка са избрани

отново тези на десетте съединения, изолирани от д-р Бозов и съавтори (вижте т. 3.8.5 и таблица 3.8.12 в дисертацията).

Стоте спектъра избрани за обучителна извадка генерират при интерпретационно търсене в спектралната библиотека **PhyChem** списъци от подструктури с размери между 3 и 858 подструктури (151 средно), а размерът на списъците, получени за спектрите на тестващата извадка, варира от 8 до 800 подструктури (134 средно). За оценка ефективността на вероятностната функция, подструктурите от списъците на спектрите в обучаващата и съответно тестващата извадка са обединени в два отделни списъка от 15 136 и 13 424 подструктури, съответно 47.5% и 46.1% от които са верни. Вероятността за коректност на подструктурите във всеки списък е определена чрез функцията на надеждност, след което подструктурите в двата списъка са сортирани по тази стойност. Изчислена е оценка на ефективността на функцията за 90%, 95% и 99% вероятност по схемата, докладвана в т. 3.8.5. Тази оценка на вероятността се дефинира от броя на верните подструктури като процент от общия брой подструктури, за които функцията на надеждност е дала коректност с вероятност над избран от нас праг (90%, 95% и 99%). Резултатите за двете извадки са представени в таблица 3.8.13, от където може да се види, че стойностите на съответната оценка са много по-ниски от зададената прагова вероятност. Това означава че вероятностната функция не работи добре. Създаването на нова функция би могло да бъде решение на разгледания проблем, но това изисква сложна статистика и е време- и ресурсоемка процедура. От друга страна цел на групата по молекулна спектроскопия е разширяване на създадената библиотека с допълнителни спектри: до сега други 500 спектъра са набрани от статитии на същото списание.

Таблица 3.8.13. Оценка на преносимостта на функцията на надеждност. Означения: **LS** и **TS** – обучаваща и тестваща извадка; P_T – праг на надеждността; **NCS** и **NIS** – съответно, брой на коректните и неверните подструктури, които са предсказани с надеждност над дадения праг; P_E – оценка на надеждността.

Извадка	P_T = 90 % [b]			P_T = 95 %			P_T = 99 %		
	NCS	NIS	A_E, %	NCS	NIS	A_E, %	NCS	NIS	A_E, %
LS	4 470	3 168	58.5[c]	1 966	827	70.4	669	231	74.3
TS	3 750	2 792	57.3	1 592	697	69.6	544	222	71.0

Алтернатива на вероятностна функция е прилагането на някакъв вид сортиращ критерий съставен също от подструктурните параметри от таблици 3.8.2 и 3.8.3. Критерият би трябвало да сортира верните подструктури в началните позиции на списъка с подструктури без да е свързан с оценка на вероятността, зависеща от вече създадената вероятностна функция. Създаването на такъв критерий не изисква много време и/или сложна статистика и ефективността му лесно може да бъде проверена с помощта на малка тестваща извадка. Допълнително, определена вероятностна функция лесно може да бъде извлечена от стойностите на критерия за всички подструктури в списъците, които съставят обучаващата извадка.

Всички 48 подструктурни параметри описани в таблици 3.2.2 и 3.2.3 са сравнени според тяхната способност да разграничават верните от неверните подструктури в обучителната извадка. Четири от тях, посочени в таблица 3.2.14, показаха най-добри резултати. Допълнително, неопределеността при сравнение

на сигналите (**Tol**) е много добър параметър, тъй като голяма част от верните подструктури се генерират при малка нейна стойност. **Tol** силно корелира с параметъра **sRMSD**, ето защо само единият от тях е използван при съставянето на критерия.

Таблица 3.8.14. Параметри, които участват в сортиращия критерий (**sRank**).

sRMSD: минималното средно-квадратично отклонение между химичното отместване на сигналите на неизвестното съединение и това на тези в подструктурата;
sFV: броят на свободните валенции на подструктурата;
sLO: броят на библиотечните съединения, които съдържат подструктурата;
sSO: броят на присъствията на подструктурата в другите подструктури, генерирани със същата неопределеност;

Няколко комбинации от параметрите в таблица 3.8.14 заедно с други параметри са тествани, в резултат на което бе избран критерият описан от уравнението:

$$\mathrm{sRank} = \mathrm{f}\,\frac{\mathrm{sLO}\cdot\mathrm{sSO}}{\mathrm{LibSz}} * \frac{\mathrm{sNA}}{\mathrm{sFV}} + (1-\mathrm{f})(2-\mathrm{Tol}), \qquad (3.8.9)$$

където **LibSz** е размерът на библиотеката (1000 в конкретния случия), **sNA** броят атоми в подструктурата и **f** – тегло, с което се взимат двете части на критерия (този параметър е оптимизиран в интервала от 0.0 до 1.0).

Променливите са разпределени в числителя или знаменателя, с положителен или отрицателен знак, в зависимост от тяхната позитивна или негативна разграничаваща способност: колкото по-висока е стойността на сортиращия критерий, толкова по-висока е вероятността подструктурата да бъде вярна. Параметърът **sNA** е поставен в числителя с цел сортиране на големите подструктури в началото на списъка. Тук е редно да спомена, че значението на този параметър не се заключава само с избора на големи подструктури. Допълнителни изследвания (вижте следващата точка) показаха, че поставен в знаменателя на квадрат той повишава достоверността на подструктурите. Едно от възможните обяснения, е че големите подструктури, които се генерират с тази библиотека, са сравнително локализирани в единия край на референтното съединение като имат малки на брой свободни валенции. Каквото и да е мистериозното влияние на този параметър, то неизбежно е свързано със структурни корелации в структурите на библиотечните съединения *и докато* тези корелации са налични в тази библиотека този параметър има своето „участие" във всеки един сортиращ критерий *за тази библиотека*.

Новата вероятностна функция е съставена, чрез използване стойностите на **sRank** критерия за всички подструктури в обучителната извадка. Процедурата е еквивалентна на тази в т. 3.8.3 с единствената разлика, че стойността на критерия замества стойността на изхода на невронната мрежа. По същия начин [вижте формула (3.7.11b)], е изчислена чувствителността на класификация (показателя recall) за прагова вероятност 90%, 95% и 99%. Оптималната стойност на коефициента **f** в уравнение (3.8.9) е потърсена чрез варирането му в интервала от 0.0 до 1.0 в 10 стъпки и сравняване чувствителността на класификация за прагова вероятност от 99%. На фигура 3.8.5 в дисертацията са представени графиките на

зависимостта на чувствителността от тегловния фактор за трите прагови вероятности и за двете извадки. Зависимостите R_{90}, R_{95} и R_{99} от f при обучителната и тестваща извадка съвпадат почти напълно по техния характер на поведения (наклон, максимуми, стойности) и кривите за R_{99} и в двете извадки имат оптимум при f = 0.6. Разбира се, кривите за другите две прагови вероятности от 90% и 95% имат максимуми съответно при f = 8 и f = 7 (и в двете извадки), но според нас евентуалният потребител на софтуера за интерпретационно търсене би използвал надеждност от 99%, затова f = 0.6 се препоръчва от нас като оптимална стойност. Съответстващата чувствителност (recall) за двете извадки е много близка и сравнително висока: за ОИ: R_{90} = 45.2%, R_{95} = 37.5% and R_{99} = 25.6%; за ТИ: R_{90} = 41.8%, R_{95} = 34.0% и R_{99} = 25.3%.

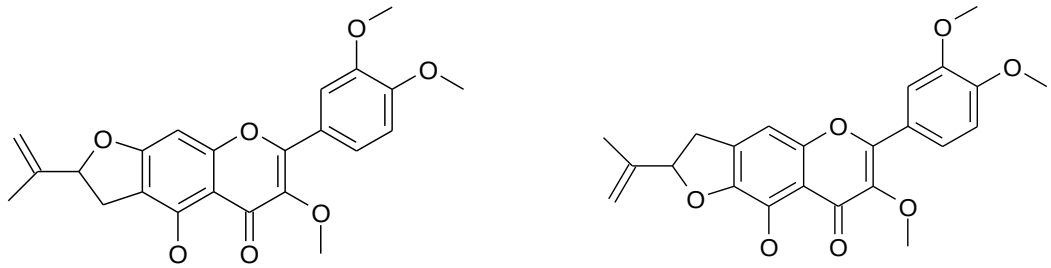
Сортиращият критерий е приложен за спектрите в обучителната и тестващата извадка, като резултатите показват, че 95 от 100 списъка с подструктури в обучителната извадка сортирани по този критерий дават вярна подструктура на първо място в списъка. За спектрите от ТИ те са 92 от 100. Десетте спектъра от валидиращата извадка генерират списък с резултати с размер от 37 до 680 подструктури (194 средно). В четири от случаите първата подструктура е невярна. Всички подструктури, първи по ранг в списъка с подструктури, заедно с потърсената структура са показани на таблица 3.8.14; отново предсказаната подструктурата е представена с удебелени химични връзки в структурата на референтното съединение. Този резултат е незадоволителен, ако се сравни с резултата от интерпретационното търсене в голямата **LAST** библиотека и използването на функцията на надеждност, която е изработена за търсене в нея.

Както споменахме в началото на т. 3.8 интерпретационното търсене има претенцията да се справя много по-добре от търсенето по подобие в спектрална библиотека – и на теория, и както показва нашият практически опит, за получаване на добри резултати не се изисква наличието на библиотечни съединения подобни по структура на неизвестното, а само съединения, които имат обща част (подструктура) между неизвестната и някои от библиотечните структури. Но използването на параметрите **sLO** и **sSO** в сортиращия критерий (а и във функцията на надеждност) за ИБТ и положителната корелация на техните стойности с надеждността на подструктурите води до неминуемото изискване в библиотеката да има значителен набор от съединения, които имат обща подструктура между неизвестната и някои от библиотечните структури.

3.9. Оценка и трансформация на резултатите от интерпретацията към програма за структурна генерация.

При анализ на структурите на съединенията от хит-списъка с концепцията на максимална обща подструктура (МОП) бяха дадени няколко примера за структурна генерация (т. 3.6): получените МОП бяха използвани в т.н. **GOODLIST**, една по една или заедно. Беше разгледан и пример с поставяне на подструктура в списъка **MACROATOM**. В настоящата част в дисертацията са разгледани особеностите на използване на информацията от компютърната интерпретация при работа с два структурни генератора – **MolGen** и **Houdini**. Поради липса на място тази част е значително съкратена в методичната си част като са оставени само резултатите.

Бяха избрани 12 съединения, които са използвани като независима валидираща извадка за оценка на функцията на надеждност, която е генерирана за голямата библиотека **LAST**. Структурите на тези 12 съединения са дадени на фигура 3.9.8 в дисертацията, а резултатите от ИБТ - в таблица 3.9.6 в дисертацията. Най-лоши са резултатите за Седкатрин А, като нито една от подструктурите, предсказани с надеждност по-голяма от 90%, не е вярна. За всички останали съединения подструктурата, предсказана с най-висока надеждност е вярна. За 11 съединения от таблицата (без **Седкатрин А**) структурната генерация с използване на първата подструктура значително намалява броя на генерираните структури, както и времето за структурна генерация. В статията по дисертацията [D10] са разгледани подробно резултатите от структурната генерация с използване на 1D- и 2D-ЯМР спектрите на **веллокуерцетин**, както и на подструктурите, генерирани при ИБТ: при използване на две от подструктурите (№ 24 и 30, с P = 94.07% и 92.89%) в структурния генератор **Houdini** се получават само две структури, които са дадени на фигура 3.9.9.



Истинската структура

Изомер на истинската структура

Фигура 3.9.9. Резултати от структурния генератор **Houdini** при използване на 1D- и 2D-ЯМР спектрите на **веллокуерцетин**, както и с две от подструктурите, генерирани при ИБТ.

3.10. Друг поглед към търсенето по подобие в библиотеки от вибрационни спектри.

Всички методи за интерпретация на спектрална информация се основават на връзката между спектрите и съответните структури, изразена с абстрактното уравнение (1.1.1). Тази връзка най-явно си проличава при търсенето в спектрални библиотеки по подобие и е в основата на метода на най-близките съседи и прилагането на концепцията за максимална обща подструктура при обработка на хит-списъците с резултати. Дори и интерпретационното библиотечно търсене излича подструктури от референтни структури, които са донякъде подобни на структурата на търсеното съединение..

3.10.1. Изследване на връзката между структурно и спектралното подобие за ИЧ и Раман спектри на едни и същи съединения.

Съставянето на две библиотеки на 185 органични съединения, **IRRa** от техните ИЧ спектри и **RaIR** от техните Раман спектри позволява да се сравни търсенето по подобие във всяка една от тях. Допълнително беше съставена и спектрална библиотека на същите 185 органични съединения, **IRRaAu**, в която „спектърът" е съставен от съответните ИЧ и Раман спектри, като първите

стойности на ординатата са от ИЧ спектъра, а следващите стойности на ординатата – от Раман спектъра. Това може да се опише формално със следните уравнения:

$$A_k^{au} = A_k^{IR} \text{ и } A_{k+801}^{au} = A_k^{Ra} \text{ за } k = 1 \ldots 801$$

Очевидно е, че в новата библиотека, **IRRaAu**, „спектърът" е предствен с 1602 стойности. Тази библиотека ще наричаме „комбинирана / разширена библиотека".

Набор от 500 подструктури, дефинирани от проф. Вармуца и предоставени ни от него като SDF файл, са използвани за изчисляване на т.н. структурен пръстов отпечатък (fingerprint) на съединенията от спектралните библиотеки. В програмата **SubMat** се зареждат тези подструктури и библиотечните структури и се изчислява матрица, която съдържа единици и нули. Ако означим съответния матричен елемент с $d_{k,m}$, то $d_{k,m}$ = 1 означава, че **k**-тата структура съдържа **m**-тата подструктура; при стойност нула – че не я съдържа. Подобието на две библиотечни структури **k** и **m** се изчислява с индекса на Танимото.

Спектрална библиотека	Коефициент на корелация
IRRa	0.513
RaIR	0.479
IRRaAu	0.564
IRRa / RaIR	0.573[a)]

Таблица 3.10.1. Коефициент на корелация между структурното подобие на всички двойки структури и спектралното подобие на всички двойки спектри в съответната спектрална библиотека.

[a)] Това е коефициентът на корелация между спектралното подобие на всички двойки спектри в ИЧ библиотеката и всички двойки спектри в Раман библиотеката.

По-сигурна оценка за информационното съдържание с повече данни може да се извърши по следния начин: (1) определя се структурното подобие по уравнение (3.10.3) на всички двойки структури в една от библиотеките, (2) за всяка една от трите библиотеки се определя спектралното подобие на всички двойки спектри с уравнение (3.1.5), което е коефициент на корелация на спектрални криви. (3) изчислява се коефициентът на корелация между структурното подобие на двойките структури и спектралното подобие на двойките спектри.

В таблица 3.10.1 са дадени коефициентите на корелация (Pearson correlation coefficients) на тези 17 020 двойки спектрално/структурно подобие. Без съмнение всички тези коефициенти на корелация в таблицата са статистически значими. Колкото даден коефициент има по-висока стойност, толкова по-добре съответните спектри описват структурата на съединенията. Най-добре отразяват структурното подобие спектрите на комбинираната библиотека, след това ИЧ и накрая Раман библиотеката. Последният коефициент на корелация, 0.573, показва че спектралното подобие на двойките ИЧ спектри не е същото като това на двойките Раман спектри. Подробна дискусия за спектралните причини за това е дадена в следващата точка.

3.10.2. Съвместна спектрална база от данни, съставена от ИЧ и Раман спектри на едни и същи съединения.

Както споменахме, комбинираната библиотека от ИЧ и Раман спектри, **IRRaAu**, не може да се поддържа от програмата **IRSS**. В настоящата част предлагаме една интересна алтернатива на комбинираното търсене.

Алтернативата е двойките ИЧ и Раман спектри на библиотеките **IRRa** и **RaIR** да се осреднят и получените „спектри" да съставят нова библиотека, наречена от нас **IRRaAv**: това се улеснява от това, че двата типа спектри са представени при едни и същи вълнови числа и ординатата им е нормирана в интервала 0 - 1. Още повече, че в програмата **IRSS** има интерфейс за изваждане на спектри, който лесно се преобразува в такъв за събиране на спектри.

Освен осредняване са възможни и други аритметични комбинации на ординатите на двата вида спектри. Всички те са дадени с уравнения (2.10.5-8).

осредняване: $A_k^{new} = (A_k^{IR} + A_k^{Ra})/2$ **(3.10.5)**

произведение: $A_k^{new} = (A_k^{IR} \times A_k^{Ra})$ **(3.10.6)**

изваждане: $A_k^{new} = (A_k^{IR} - A_k^{Ra})$ **(3.10.7)**

изваждане и събиране: $A_k^{new} = (A_k^{IR} + A_k^{Ra}) \times (A_k^{IR} - A_k^{Ra})$ **(3.10.8)**

С тези уравнения бяха създадени четири библиотеки от комбинирани спектри: **IRRaAv**, **IRRaPr**, **IRRaSu** и **IRRaSA** и е оценено търсенето по подобие, подобно на схемата, която е използвана за съставяне на данните в таблица 3.10.1 от дисертацията.

Спектралното търсене е проведено с коефициент на корелация на спектрални криви, уравнение (3.1.5). Както се вижда от таблицата, „спектрите" в осреднената библиотека по-добре съответстват на структурата на съединенията, отколкото спектрите в оригиналните ИЧ и Раман библиотеки. Произведението на спектри не се представя добре и според нас причината за това е, че при съединения с център на симетрия, според правилото за алтернативна забрана, би се получил изключително странен спектър, с ивици съизмерими с шума, освен в областите, където има случайно съвпадение на ивици. Двата вида изкуствени спектри, в които участва разликата между ИЧ и Раман спектрите, показват най-ниски коефициенти на корелация. Причина за това е, че редица ивици, с приблизително еднакви относителни интензитети в ИЧ и Раман спектрите, се нулират при изваждането. Изненадващото в случая е, че коефициентите на корелация за трите последно дискутирани преобразувания не са толкова ниски, но първо в библиотеката има малко на брой съединения с център на симетрия, и второ относителният интензитет на ивиците е коренно различен за повечето ивици в ИЧ и Раман спектрите. От таблици 3.10.1-2 се вижда, че разширените ИЧ - Раман спектри показват по-добра корелация със структурите на съединенията, отколкото осреднените спектри: коефициентите на корелация са съответно 0.564 и 0.538.

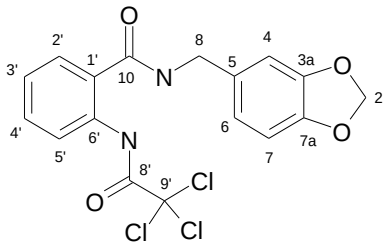
3.11. Отнасяне на ЯМР спектри на органични съединения.

В тази част се докладват резултати, които са получени чрез прилагане на някои от компютърните методи. Синтезът или изолирането на съединенията са извършени от колеги-химици, а приносът на автора е в отнасяне на ^{1}H- и ^{13}C-ЯМР спектрите.

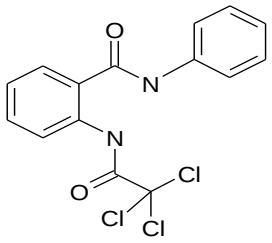
3.11.1. На синтезирани съединения.

1. N-(1,3-бензодиоксол-5-илметил)-2-[(трихлороацетил)амино]бензамид

На фигура 3.11.1 са дадени структурите на двете съединения **1** и **2**. Те са синтезирани отново от д-р Жан Петров.



1



2

Фигура 3.11.1. Структурата на съединения **1** и **2**. Номерацията е за целите на отнасянията.

Таблица 3.11.1. ^{1}H и ^{13}C NMR спектралните данни, както и ^{1}H-^{1}H COSY и HMBC корелациите за **1** [600.13 MHz (^{1}H) и 150.903 MHz (^{13}C)][a].

атом	δ (^{13}C) ppm	DEPT[b]	δ (^{1}H) ppm	мултиплетност (J, Hz)	^{1}H-^{1}H COSY[b]	HMBC[b]
2	100.81	CH_2	5.99	s		3a, 7a
3a	147.19	C				
4	108.04	CH	6.92	d (1.5)	8 [c], 6 [c], 7 [d]	6, 7a, 8
5	132.51	C				
6	120.64	CH	6.81	dd (7.9,1.5)	7, 4[c], 8 [c]	4, 7a, 8
7	107.96	CH	6.86	d (7.9)	6, 4[d]	5, 3a
7a	146.14	C				
8	42.41	CH_2	4.41	d (5.9)	9, 4[c], 6 [c]	4, 5, 6, 10
9 (NH)			9.50	t (5.8)	8	10, 8 [c]
10 (C=O)	168.00	C				
1′	121.13	C				
2′	128.26	CH	7.94	dd (7.9,1.3)	3′, 4′[c]	4′, 6′, 10, 1′[d]
3′	124.83	CH	7.33	td (7.8,0.9)	4′, 2′, 5′[c]	1′, 5′, 2′[c], 4′[d], 6′[d]
4′	132.72	CH	7.65	td (7.8,1.2)	5′, 3′, 2′[c]	2′, 6′, 5′[c]
5′	120.23	CH	8.38	dd (8.4,0.8)	4′, 3′[c]	1′, 3′, 6′[c], 10[d]
6′	137.14	C				
7′ (NH)			13.17	s		
8′ (C=O)	159.31	C				
9′	92.66	C				

[a)] В разтвор на DMSO-d6. Всички тези отнасяния са в съгласие с COSY[b)], HMQC и HMBC спектрите. [b)] За съкращенията вижте началото на автореферата. [c)] Тези корелации са слаби. [d)] Тези корелации са изключително слаби.

В таблица 3.11.1 е дадено пълното отнасяне на протонния и въглеродния спектри на **1**. Отнасянето на сигналите подвърждава структурата и молекулната формула $C_{17}H_{13}Cl_3N_2O_4$.

2. N-фенил-2-[(трихлороацетил)амино]бензамид

В таблица 3.11.2 е дадено пълното отнасяне на протонния и въглеродния спектри на **2**.

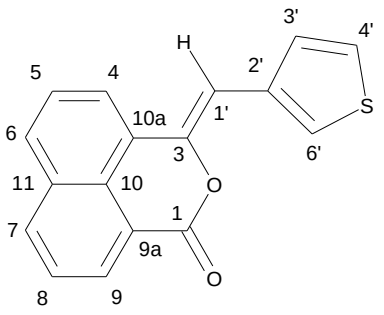
Таблица 3.11.2. ^{1}H и ^{13}C NMR спектралните данни, както и ^{1}H-^{1}H COSY и HMBC корелациите за **2** [600.13 MHz (^{1}H) и 150.903 MHz (^{13}C)][a].

атом	δ(^{13}C) ppm	DEPT	δ(^{1}H) ppm	мултиплетност (J, Hz)	^{1}H-^{1}H COSY	HMBC
1	138.62	C				
2	121.77	CH	7.71[d]		4[b]	4, 6
3	129.19	CH	7.40[d]		2, 6	1, 5
4	125.01	CH	7.17	tt (7.4,1.2)		2, 6, 1[b], 3[b], 5[b]
5	129.19	CH	7.40[d]		2, 6, 3, 5	1, 3
6	121.77	CH	7.71[d]		4[b]	2, 4
7(NH)			10.65	s		8, 2, 6
8(C=O)	167.34	C				
1′	123.61	C				
2′	129.70	CH	8.03	dd (7.9,1.2)	3′, 4′[c]	8, 4′, 6′, 1′[c], 5′[c]
3′	125.46	CH	7.43	td (7.7,1.1)	2′, 4′, 5′[c]	1′, 5′, 4′[b]
4′	133.17	CH	7.70[d]	td (7.9,1.5)	2′[c], 3′, 5′	2′, 6′
5′	121.49	CH	8.29	dd (8.3,0.9)	4′, 3′[c]	1′, 3′, 6[b], 2′[c]
6′	137.31	C				
7′(NH)			12.44	s		5′, 8′, 1′[b]
8′(C=O)	159.82	C				
9′	93.16	C				

[a] В разтвор на DMSO-d6. Съкращенията са като в таблица 3.11.1. [b] Тези корелации са слаби. [c] Тези корелации са изключително слаби. [d] Поради препокриване на сигналите тези стойности са взети от HSQC спектъра.

3. **3-(3-Тиенилметилен)-1H,3H-нафто-[1,8-c,d]-пиран-1-он***

Съединение е синтезирано от д-р Марин Маринов и проф. Нейко Стоянов по нов метод, с по-голям добив и доколкото имам информация от М.М. то е ново съединение. На фигура 3.11.3. е дадена неговата структура, като трябва да се има предвид, че ние не определяме дали екзоцикленатa двойна връзка е E или Z, но все пак изчертаването е възможно само за един от тези изомери – в случая Z заместена двойна връзка.



Фигура 3.11.3. Структурата на съединение **3**. Номерацията е за целите на отнасянията. Молекулната формула е $C_{17}H_{10}O_2S$.

Пълното отнасяне на сигналите на **3** е дадено в таблица 3.11.3.

* Отнасянията са публикувани в статия по дисертацията [D17].

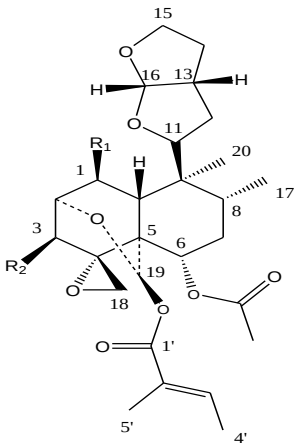
Таблица 3.11.3. ^{1}H и ^{13}C NMR спектралните данни, както и ^{1}H-^{1}H *COSY* и *HMBC* корелациите за **3** [600.13 MHz (^{1}H) и 150.903 MHz (^{13}C)][a].

Атом	δ (^{13}C) ppm	DEPT[b]	δ (^{1}H) ppm	Мултиплетност (J, Hz)	^{1}H-^{1}H *COSY*[b,e]	*HMBC*[b,f]
1 (*C*=*O*)	159.69	*C*				
3	144.89	*C*				
4	121.55	*CH*	8.26	dd (7.5, 0.5)	5, 6[c], 1′[c]	3, 6, 9a[d], 10, 11[d]
5	127.42	*CH*	7.73	t (7.8)[h]	4, 6	10a, 11
6	127.98	*CH*	8.08	d (8.1)[i]	4[c], 5, 7[d]	4, 7, 10, 11[c]
7	134.38	*CH*	8.35	dd (8.3, 0.8)	6[d], 8, 9[d]	9, 9a[d], 10, 11[c]
8	126.88	*CH*	7.78	dd (8.2, 7.2)	7, 9	9[c], 9a, 11
9	129.13	*CH*	8.33	dd (7.2, 1.1)	7[d], 8	1, 7, 10
9a	119.16	*C*				
10	127.44	*C*				
10a	123.62	*C*				
11	132.05	*C*				
1′	103.29	*CH*	7.18	*s*		1, 3, 10a, 2′[d]
2′	135.00	*C*				
тиофенов	128.84	*CH*	7.66[g]	m		
тиофенов	126.21	*CH*	7.66[g]	m		
тиофенов	125.58	*CH*	7.92	m		

[a)] В DMSO-d_6 разтвор. [b)] За съкращенията вижте началото на дисертацията. [c)] Тези корелации са слаби. [d)] Тези корелации са изключително слаби. [e)] Тиофеновите протони показват силни и слаби *COSY* и *HMBC* корелации между тях, които са споменати в текста. [f)] H-1′ показва силни корелации с два от трите протона към *C*-3′, *C*-4′ or *C*-6′. [g)] Тези стойности са взети от мултиплета H в ^{1}H-spectrum на фигура 3.11.4. HMQC дава 7.67 ppm и 7.66 ppm. [h)] ^{1}H сигналът на H-5 е триплет с разширени сигнали, вместо да е триплет от дублети. [i)] ^{1}H сигналът на H-6 е дублет с разширени сигнали, вместо да е дублет от дублети.

3.11.2. На съединения, изолирани от растения.

И трите съединения, на които са отнесени от нас спектрите, са изолирани от д-р Петко Бозов. Те са *нео*-клеродани от класа на дитерпените. Структурите на **4**, **5** и **6** са дадени на фигура 3.11.6.



Фигура 3.11.6. Структурите на **скутекиприн А**, **4**, **неоаюгапирин А**, **5**, и **скутегалерин А**, **6**.
скутекиприн А: R_1 = H, R_2 = H
неоаюгапирин А: R_1 = H, R_2 = OH
скутегалерин А: R_1 = OH, R_2 = H

4. Скутекиприн А

Пълното отнасяне на сигналите на **4** е дадено в таблица 3.11.4.

Таблица 3.11.4. ^{1}H и ^{13}C NMR спектралните данни, както и ^{1}H-^{1}H *COSY* и *HMBC* корелациите за **4** [600.13 MHz (^{1}H) и 150.903 MHz (^{13}C)][a].

Атом	δ(^{13}C) ppm	DEPT	δ(^{1}H) ppm	мултиплетност (J, Hz)	^{1}H-^{1}H COSY	HMBC
1	28.45	CH_2	2.36 (α) 1.60[e] (β)	m m	1β, 2β, 10β, 3α[c] 1α, 10β	2[c] – [f]
2	67.27	CH	4.18 (β,eq)	m	1α, 3α[β], 3β[c]	4[c], 19[c]
3	36.90	CH_2	2.55 (α) 1.79 (β)	dt (13.8,2.3) dd (–, 2.7)[d]	2β[β], 3β, 1α[c] 2β[c], 3α	1[c], 2[c] 4[b], 5[c], 6[c], 18[c]
4	60.66	C				
5	41.46[g]	C				
6	68.35	CH	4.62 (β)	dd (11.6, 4.5)	7α, 7β	4[b], 5[b], 7[b], 19, $CH_3\underline{C}O$,
7	33.16[h]	CH_2	1.65[e](α,ax) 1.38 (β,eq)	m ddd(12.6,4.4,2.7)	6β, 7β, 6β, 7α, 8β	– [f] 5[b] or 9[b], 6[c], 8[c]
8	35.16	CH	1.64[e](β)	m	7β, 17	– [f]
9	41.85[g]	C				
10	40.80	CH	2.04 (β)	dd (11.4,4.3)	1α, 1β	1, 2[c], 4, 9, 11[c], 19[c], 20
11	86.02	CH	4.08 (α)	dd (10.9, 5.8)	12α, 12β	8[b], 10, 20[b]
12	33.53[h]	CH_2	1.65[e] (α) 1.97 (β)	m m	11α,12β 11α, 12α, 13β[c]	– [f] none
13	41.85	CH	2.85 (β)	m	12β[c],14β[c],16β[c]	none
14	32.64	CH_2	1.68[e](α) 2.15 (β)	m m	14β, 15 14α, 13β[c], 15	– [f] 12[b], 13[b], 15[b]
15	68.31	CH_2	3.88	m, 2H	14α,14β	13[b], 16[b]
16	108.28	CH	5.64 (β)	d (5.1)	13β[c]	11,13,15,12[b] or 14[b]
17	16.71	CH_3	0.90	d (6.5)	8β	6[c], 7, 8, 9
18	50.20	CH_2	2.44 (A) 3.00 (B)	d (4.4) d (4.4)	18B 18A	3[c], 4[c] 4[c]
19	91.46	CH	6.81	s		4, 6, 1′
20	14.06	CH_3	1.16	s		8,5 or 9 or 10,11
1′C=O	166.39	C				
2′	128.89	C				
3′	138.46	CH	7.10	qq (7.1, 1.4)	4′, 5′[c]	1′[b], 4′[b], 5′[b]
4′	14.57	CH_3	1.81	dq (7.1, 1.2)	3′, 5′[c]	1′[c], 2′, 3′
5′	11.94	CH_3	1.89	qui = dq (1.2)	3′[c], 4′[c]	1′, 2′, 3′
$CH_3\underline{C}O$	170.07	C				
$\underline{C}H_3CO$	21.03	CH_3	1.80	s		$CH_3\underline{C}O$

[a)] В разтвор на $CDCl_3$; δ_{ref} 7.26; ^{13}C 150.9 MHz, δ_{ref} 77.0 ppm. Съкращенията са като в таблица 3.11.1. [b)] Тези корелации са слаби. [c)] Тези корелации са изключително слаби. [d)] ^{1}H сигналът на H-3β се препокрива и само ^{3}J = 2.7 Hz може да бъде определена; [e)] взето от *HSQC*; [f)] не могат да се определят поради препокриване с други сигнали. [g),h)] сигналите с една и съща буква могат да бъдат разменени.

5. Неоаюгапирин А

Пълното отнасяне на сигналите на **5** е дадено в таблица 3.11.6.

Таблица 3.11.6. ^{1}H и ^{13}C NMR спектралните данни, както и ^{1}H-^{1}H *COSY* и *HMBC* корелациите за **5** [600.13 MHz (^{1}H) и 150.903 MHz (^{13}C)][a].

Атом	δ(^{13}C) ppm	DEPT	δ(^{1}H) ppm	мултиплетност (J, Hz)	^{1}H-^{1}H COSY	HMBC
1	22.6	CH_2	2.30 (α) 1.90 (β)	ddd(14.7,4.9,4.0) ddd(14.6,11.5,0.9)	1β, 2β, 10β, 3α[c] 1α, 10β	3,4 – [f]
2	71.0	CH	3.98 (β)	ddd(4.9,3.1,0.8)	1α, 3α	3[c],4[c],10[c],19[c]
3	70.1	CH	4.40 (α)	ddd(5.1,2.9,1.9)	1α[c],2β, OH	1[c],2[c],4[b]
4	65.9	C				
5	42.5	C				
6	68.2	CH	4.63 (β)	dd (11.9, 4.7)	7α, 7β	4,18,7[b],19,CH_3CO
7	33.3	CH_2	1.65[e](α,ax) 1.39 (β,eq)	m ddd(13.1,4.6,3.0)	6β, 7β, 6β, 7α, 8β	– [f] 5[b]+9[b],6[c],8[c]
8	35.4	CH	1.53(β)	dqd(12.8,6.6,3.1)	7β, 17	– [f]
9	41.2	C				
10	40.7	CH	1.98 (β)	dd (11.6,3.9)	1α, 1β	1,4,5+9,11[c],19[c],20[c]
11	86.4	CH	4.09 (α)	dd (11.0,5.7)	12α, 12β	8, 10, 20
12	33.3	CH_2	1.65[e] (α) 1.93[e] (β)	m m	11α,12β 11α,12α,13β[c]	– [f]
13	41.8	CH	2.85 (β)	br ddd(5.1,3.0,1.2)	12β[c],14β[c],16β[c]	11,15
14	32.6	CH_2	1.72[e](α) 2.15 (β)	m ddt(12.7,9.2,8.3)	14β, 15 14α,13β[c],15	– [f] 12[b],13[b],15[b]
15	68.3	CH_2	3.8765 3.8617	m, 2H	14α,14β	13,14[b],16
16	108.1	CH	5.63 (β)	d (5.1)	13β[c]	11,13,15,14[b]
17	16.4	CH_3	0.89	d (6.1)	8β	6[b],7[b],8,9
18	44.1	CH_2	2.88 (A) 3.09 (B)	d (4.3) d (4.3)	18B 18A	3[b], 4 4,10[c]
19	91.0	CH	6.76	s		2,4,6[b],18[c],1′
20	14.3	CH_3	1.19	s		8,9,10,11
1′C=O	166.0	C				
2′	128.7	C				
3′	138.7	CH	7.06	qq (7.1, 1.4)	4′, 5′[c]	1′[b],4′[b],5′[b]
4′	14.5	CH_3	1.80	dq (7.1, 1.2)	3′, 5′[c]	1′[b],2′,3′
5′	11.9	CH_3	1.87	qui[d] (1.3)	3′[c], 4′[c]	1′,2′,3′
CH_3CO	170.0	C				
CH_3CO	21.0	CH_3	1.79	s		CH_3CO

[a)] В разтвор на $CDCl_3$; δ_{ref} 7.26; ^{13}C 150.9 MHz, δ_{ref} 77.0 ppm. Съкращенията са като в таблица 3.11.1. [b)] Тези корелации са слаби. [c)] Тези корелации са изключително слаби. [d)] Привидна мултиплетност; [e)] взето от *COSY*; [f)] не могат да се определят поради препокриване с други сигнали. [g),h)] сигналите с една и съща буква могат да бъдат разменени.

6. Скутегалерин А

Измерването на спектрите на **Скутегалерин А** и разкриване на структурата му е направено от проф. Coll от вещество, изолирано от д-р Бозов. Отново **Скутегалерин А**, **6**, беше изолиран от д-р Бозов, спектрите на **6** бяха отново измерени, но в лабораторията по ЯМР на ИОХ към БАН, и те бяха интерпретирани от нас с цел разкриване на структурата му. В даден момент на интерпретацията разбрахме, че това е същото вещество, чиято структура е определена от проф. Coll. Пълното отнасяне на сигналите на **6** е дадено в таблица 3.11.7.

Таблица 3.11.7. ^{1}H и ^{13}C NMR спектралните данни, както и ^{1}H-^{1}H *COSY* и *HMBC* корелациите за **6** [600.13 MHz (^{1}H) и 150.903 MHz (^{13}C)][a].

Атом	δ(^{13}C) ppm	DEPT	δ(^{1}H) ppm	мултиплетност (J, Hz)	^{1}H–^{1}H COSY	HMBC
1	67.1	CH	4.38 (α)	m	2β,10β[c]	2[c],9[c]
2	69.3	CH	4.11 (β)	dt(5.1,2.7)	1α,3B	4[c]
3	30.9	CH_2	2.46 (α) 2.23[e] (β)	br d(14.4) m, препокр.	3B 3A,2β	1[b],2[b],4 4,10
4	60.1	C				
5	43.4	C				
6	67.8	CH	4.61 (β)	dd (11.9, 4.5)	7α,7β	4,5,7,19,$CH_3\underline{C}O$
7	32.5	CH_2	1.63[e](α,ax) 1.37 (β,eq)	m ddd(13.0,4.5,2.9)	6β,7β 6β,7α	– [f] 8[c],9[b]
8	35.6	CH	1.53(β)	m	17	няма
9	40.5	C				
10	51.7	CH	1.76 (β)	d (1.9)	1α[c]	2[b],4[b],5[b],9[b],19[b],20[b]
11	87.2	CH	4.09 (α)	dd (11.3,5.0)	12α,12β	8,20
12	33.6	CH_2	1.65[e] (α) 2.00 (β)	m br d(5.2)	11α,12β 11α,12α,13β	– [f]
13	41.6	CH	2.92 (β)	m	12β,14β[c],16β[c]	няма
14	32.7	CH_2	1.74[e](α) 2.21[e] (β)	m m, препокр.	14β,15A,15B 14α,13β[c],15A,15B	– [f] – [f]
15	68.9	CH_2	3.941 (A) 3.876 (B)	m, 2H	14α,14β,15B 14α,14β,15A	14[c]
16	108.4	CH	5.69 (β)	d (5.2)	13β[c]	11,13,14,15
17	16.0	CH_3	0.88	d (6.6)	8β	7,8,9
18	50.4	CH_2	2.51 (A) 3.00 (B)	d (4.3) d (4.3)	18B 18A	3[c],4[c] 4[c]
19	90.5	CH	6.67	s		4,6,1′
20	16.3	CH_3	1.21	s		8,9,10,11
1′C=O	166.4	C				
2′	128.8	C				
3′	138.8	CH	7.11	qq (7.0, 1.5)	4′, 5′[c]	1′[b],4′[b],5′[b]
4′	14.6	CH_3	1.81	dq (7.1, 1.1)	3′, 5′[c]	1′[c],2′,3′
5′	11.9	CH_3	1.89	br s	3′[c], 4′[c]	1′,2′,3′
$CH_3\underline{C}O$	170.0	C				
$\underline{CH_3}CO$	21.0	CH_3	1.80	s		$CH_3\underline{C}O$

[a)] В разтвор на $CDCl_3$; δ_{ref} 7.26; ^{13}C 150.9 MHz, δ_{ref} 77.0 ppm. Съкращенията са като в таблица 3.11.1. [b)] Тези корелации са слаби. [c)] Тези корелации са изключително слаби. [d)] Привидна мултиплетност; [e)] взето от *COSY*; [f]) не могат да се определят поради препокриване с други сигнали. [g),h)] сигналите с една и съща буква могат да бъдат разменени.

ОБОБЩЕНИЕ НА РЕЗУЛТАТИТЕ И ПРИНОСИ

1. Създадени са спектрални библиотеки от 966 ИЧ, 330 Раман, 100 ATR и 1086 УВ-Вид спектри, както и две библиотеки от 38 225 и 1000 напълно отнесени ^{13}C-ЯМР спектри. Също така са създадени и няколко библиотеки със спектри, измерени в други лаборатории.

2. Програмирани са седем метода за търсене в библиотеки от ИЧ спектри - три от тях са за търсене по ивици (пикове) и четири - за търсене по спектрална крива; предложеният от нас метод за търсене по скаларно произведение на пикови таблици не е описан в литературата. Методите са оптимизирани за идентификация на органични съединения.

3. Разработен и изследван е метод за качествен анализ на смеси по техните ИЧ или Раман спектри. Компонентите в сместа се откриват с помощта на статистически анализ и/или чрез визуална оценка от изследователя. Методът може да се използва също за идентификация на единично съединение по неговият ИЧ спектър.

4. Изследван е стандартен метод за анализ на смеси, чрез изваждане на ИЧ спектри и са предложени четири евристики, които подобряват идентификацията на компонентите на смеси.

5. Разработени са два метода - на най-близките съседи и използване на концепцията за максимална обща подструктура - за анализ на структурите на хит-списък, получен при търсене на спектъра на органично съединение в библиотека от ИЧ спектри. Прилагането на концепцията за МОП е оптимизирано по няколко параметъра и е предложен нов ранг за сортиране на подструктурите.

6. Разработен е метод за интерпретационно търсене в библиотеки от напълно отнесени ^{13}C-ЯМР спектри. Предложено е кодиране на ароматност и тавтомерност на връзките, както и схема за отнасяне на сигналите в получените подструктури. Подструктурите се подреждат по тяхната надеждност с вероятностна функция. Методът е тестван за интерпретация на спектрите на природни съединения.

7. За новосъздадената библиотека от 1000 напълно отнесени ^{13}C-ЯМР спектри е предложен ранг за сортиране на подструктурите, които са резултат от интерпретационно търсене.

8. Изследвана е връзката между спектралното подобие и структурното подобие за ИЧ и Раман спектри. Създадена е съвместна база от данни, съставена от ИЧ и Раман спектри на едни и същи съединения.

9. Отнесени са сигналите в ^{1}H- и ^{13}C-ЯМР спектри на шест органични съединения. Отнасянето на сигналите на органичните съединения е изключително важно за следващите интерпретации на спектрите на подобни по структура съединения, и затова отнасянето може да се причисли към задачите на органичния анализ.

ПУБЛИКАЦИИ, ВКЛЮЧЕНИ В ДИСЕРТАЦИЯТА

(общ импакт фактор 15.991)

D1. **P.N. Penchev**, A.N. Sohou and G.N. Andreev; *Description and Performance Analyses of an Infrared Library Search System*. **Spectroscopy Letters**, **29** (8), 1513-1522 (1996). **IF: 0.472 (7 пъти цитирана)**

D2. K. Varmuza, **P. Penchev**, F. Stancl, W. Werther; *Systematic Structure Elucidation of Organic Compounds by Mass Spectra Classification*. **Journal of Molecular Structure**, **408/409**, 91-96 (1997). **IF: 0.884 (3 пъти цитирана)**

D3. K. Varmuza, **P. Penchev**, H. Scsibrany. *Maximum Common Substructures of Organic compounds Exhibiting Similar Infrared Spectra*. **Journal of Chemical Information and Computer Sciences**, **38**, 420-427 (1998). **IF: 2.609 (31 пъти цитирана)**

D4. **P.N. Penchev**, N.T. Kochev and G.N. Andreev; *IRSS: A Programme System for Infrared Library Search*. **Comptes Rendus de l'Academie Bulgare des Sciences**, **51** (1-2), 67-70 (1998). **IF: 0.089 (4 пъти цитирана)**

D5. K. Varmuza, **P.N. Penchev**, H. Scsibrany; *Large and Frequently Occurring Substructures in Organic Compounds Obtained by Library Search of Infrared Spectra*. **Vibrational Spectroscopy**, **19**, 407-412 (1999). **IF: 0.848 (9 пъти цитирана)**

D6. **P.N. Penchev**, G.N. Andreev, K. Varmuza; Automatic *Classification of Infrared Spectra Using a Set of Improved Expert-based Features*. **Analytica Chimica Acta**, **388** (1-2), 145-159 (1999). **IF: 1.894 (25 пъти цитирана)**

D7. **Plamen N. Penchev**, Kurt Varmuza; *Characteristic substructures in sets of organic compounds with similar infrared spectra*. **Computers & Chemistry**, **25** 231–237 (2001). **IF: 1.632 (3 пъти цитирана)**

D8. Kurt Varmuza, Nikolay T. Kochev and **Plamen N. Penchev**; *Evaluation of Hitlists from IR Library Searches by the Concept of Maximum Common Substructures*. **Analytical Sciences**, **17**, i659-i662 (2001). **IF: 0.916 (3 пъти цитирана)**

D9. **P.N. Penchev**, V.L. Miteva, A.N. Sohou, N.T. Kochev, G.N. Andreev; *Implementation and Testing of Routine Procedure for Mixture Analysis by Search in Infrared Spectral Library*; **Bulgarian Chemical Communications**, **40** (4), 556-560 (2008). **(5 пъти цитирана)**

D10. **Plamen N. Penchev**, Klaus-Peter Schulz, and Morton E. Munk; *INFERCNMR: A ^{13}C NMR Interpretive Library Search System*. **Journal of Chemical Information and Modeling**, **52**, 1513–1528 (2012). **IF: 4.675 (3 пъти цитирана)**

D11. Petko I. Bozov, **Plamen N. Penchev** and Josep Coll; neo-*Clerodane Diterpenoids from Scutellaria galericulata*. **Natural Product Communications**. **9** (3), 347-350 (2014). **IF: 0.956 (1 път цитирана)**

D12. **Plamen N. Penchev**, Stefka R. Nachkova, Tonka A. Vasileva and Petko I. Bozov; *^{1}H and ^{13}C NMR analysis of neo-Clerodane Diterpenoid Scutecyprin*. **Natural Product Communications**. **9** (8), 1065-1068 (2014). **IF: 0.956 (1 път цитирана)**

D13. **P.N. Penchev**, O.K. Argirov and G.N. Andreev; *Mass Spectra Classification According to Substructures and Molecular Formula Using Artificial Neural Networks*. **Analytical Laboratory**, **3**, 29-33 (1994). **(1 път цитирана)**

D14. **Plamen N. Penchev**, Nikolay T. Kotchev, George N. Andreev; *Infrared spectra interpretation by means of computer*. **Traveaux Scientifiques d'Universite de Plovdiv**, **29** (5), 21-26 (2000). **(4 пъти цитирана)**

D15. D. Hristozov, **P. Penchev**, G. Andreev; *Searching in UV/Vis Spectral Library*. **Traveaux Scientifiques d'Universite de Plovdiv, 30** (5), 63-66 (2001). **(3 пъти цитирана)**

D16. **P.N. Penchev**, G.N. Andreev, K. Varmuza; *Computer-assisted structure elucidation of organic compounds by infrared spectroscopy*. **Asian Chemistry Letters**, **13** (3&4), 195-200 (2009).

D17. N.M. Stoyanov, **P.N. Penchev** and M.N. Marinov; *Synthesis and spectral study of 3-(3-thienylmethylene)-1H,3H-naphtho-[1,8-c,d]-pyran-1-on*. **Asian Chemistry Letters**, **15** (1 & 2), 45-50 (2011).

D18. J.S. Petrov and **P.N. Penchev**; *A Complete ^{1}H and ^{13}C NMR data assignment for N-Benzo[1,3]dioxol-5-ylmethyl-2-(2,2,2-trichloroacetyl amino) benzamide*. **Asian Chemistry Letters**, **14** (3-4), 273-278 (2010).

D19. **P.N. Penchev** and J.S. Petrov; A complete ^{1}H *and* ^{13}C NMR data assignment for *N-phenyl-2-[(trichloroacetyl)amino]benzamide*. **Asian Chemistry Letters**, **15** (1 & 2), 21-26 (2011).

D20. N. Kochev, **P. Penchev**, G. Andreev, K. Varmuza; *Improved Realisation of Maximum Common Substructure Concept for Structure Elucidation*. **Traveaux Scientifiques d'Universite de Plovdiv, 30** (5), 73-78 (2001).

D21. S. Tsoneva, S. Nachkova, **P. Penchev**; *ATR spectra database of organic compounds*. **Scientific Works: University of Ruse "Angel Kanchev"**, **52**, Issue 10.1, 38-40 (2013).

D22. S. Nachkova, S. Milenkova, P. Bozov, **P. Penchev**; *Interpretive search in a ^{13}C-NMR spectral library of plant compounds*. **Scientific Works: University of Ruse "Angel Kanchev"**, **52**, Issue 10.1, 47-51 (2013).

D23. **P. Penchev**, S. Tsoneva, Ts. Krusteva and S. Nachkova; *Spectral Libraries of Vibrational Spectra*. **Scientific Researches of the Union of Scientist in Bulgaria – Plovdiv, Series B. Natural Sciences and the Humanities. 16**, 79-84 (2014).

D24. S. Tsoneva, S. Nachkova, **P. Penchev**; *Joint Spectral Database of Infrared and Raman Spectra*. **Scientific Works: University of Ruse "Angel Kanchev"**, **52**, Issue 10.1, 47-51 (2013).

УЧАСТИЯ В НАУЧНИ ПРОЯВИ

(14 постера и една лекция на български и две на английски)

P1. **P.N. Penchev**, A.N. Sohou, G.N. Andreev. *Interpretation of Reduced Infrared Spectra with the Aid of Artificial Neural Networks.* Second National Conference of Chemistry, Plovdiv (Bulgaria), April 12-14 1995, R VI/32.

P2. **P.N. Penchev**, A.N. Sohou, G.N. Andreev. *User-made Infrared Spectra Library Search System.* Second Nat. Conf. of Chemistry, Plovdiv (Bulgaria) 12-14 April 1995, R VI/33.

P3. K. Varmuza, **P.N. Penchev**, H. Scsibrany. *Maximum Common Substructures of Organic Compounds Exhibiting Similar Infrared Spectra.* 12th CIC Workshop, Mannedorf 1997.

P4. K. Varmuza, **P.N. Penchev**, H. Scsibrany. *Large and Frequently Occurring Substructures in Organic Compounds Obtained by Library Search of Infrared Spectra.* 3rd International Symposium on Adv. Infrared and Raman Spectroscopy, Vienna (Austria) July 5-9, 1998.

P5. K. Varmuza, H. Scsibrany, **P.N. Penchev,** F. Ehrentreich. *Maximum Common Substructures of Organic Compounds Exhibiting Similar Infrared Spectra or Mass Spectra.* 5th Intern. Conference on Chemical Structures, Noordwijkerhout, The Netherlands, June 6-10 1999.

P6. **P.N. Penchev**, N.T. Kotchev, A. N. Sohou, G. N. Andreev. *Identification of Organic Compounds and Mixtures through a Search in Infrared Spectral Databases*. Scientific Conference 75 Aniversary of Anal. Chem. Dep., Chem. Fac. at Sofia Univ., Sofia, May, 1999.

P7. **P.N. Penchev**, N.T. Kotchev, G. N. Andreev. *Computer-assisted Structure Elucidation of Organic Compounds by Infrared Spectroscopy.* Ninth International Workshop on QSAR in Environmental Sciences (QSAR2000). Bourgas, Bulgaria, September 16 - 20, 2000.

P8. **P.N. Penchev**, N.T. Kotchev, G. N. Andreev. *k-Nearest Neighbour Classification of Infrared Spectra.* Ninth Intern. Workshop on Quant. Structure Activity Relationships in Environmental Sciences (QSAR2000). Bourgas, Bulgaria, September 16 - 20, 2000.

P9. S. Nachkova, S. Milenkova, P. Bozov, **P. Penchev**; *A ^{13}C NMR Interpretive Library Search System*. International Conference of Young Scientists – PLOVDIV'2013, 13-16 юни 2013 г. (конференция с международно участие).

P10. **P. Penchev**, S. Tsoneva, Ts. Krusteva, S. Nachkova; *Spectral Libraries of Vibrational Spectra*. Научна сесия „Дни на науката 2013" на СУБ, гр. Пловдив, 30 - 31 окт. 2013 г.

P11. S. Nachkova, S. Milenkova, P. Bozov, **P. Penchev**; *Interpretive Search in a ^{13}C-NMR Spectral Library of Compounds Isolated from Plants*. Научна конференция РУ & ДНТ & СУ' 2013, гр. Разград, Ваканционен СПА комплекс "Островче", 01-02.11.2013 г. (конференция с международно участие).

P12. S. Tsoneva, S. Nachkova and **P. Penchev**; *ATR Spectra Database of Organic Compounds*. Научна конференция РУ & ДНТ & СУ' 2013, гр. Разград, Ваканционен СПА комплекс "Островче", 01-02.11.2013 г. (конференция с международно участие).

P13. Стефка Начкова, Слава Цонева, **Пламен Пенчев**; *Потребителски спектрални библиотеки*. Семинар с международно участие на ПУ "П.Хилендарски", ACM2 & Thermo Scientific, на тема*: "Съвременни методи за анализ и контрол на храни и околна среда", гр.Пловдив, 21 май 2014 г., гр. Пловдив.*

P14. *Слава Цонева,* **П. Пенчев**, С. Начкова*;* Търсене по подобие в Раман спектрални библиотеки. *Конференция на Русенския университет, Разград, 29-30 октомври, 2014 г.*

L1. **П. Пенчев**, Разкриване на структурата на органични съединения с помощта на ИЧ спектри и компютри. Научна сесия, посветена на празника на културата и науката и деня на химията, 19 май 2000 г., Химически факултет.

L2. P. **Penchev**; Computer-Assisted Structure Elucidation of Organic Compounds by Infrared Spectroscopy (in English), 26 August 1999 at Center for Computational Quantum Chemistry at University of Georgia at Athens, USA.

L3. **Plamen Penchev**; *^{13}C NMR Interpretive Library Search*, 22 June 2007, 6th CHEMISTRY CONFERENCE of Chemical Faculty of University of Plovdiv, Plovdiv 20-22 June 2007. (**лекция на английски, а конференцията е с международно участие**).

Цитирания на статиите по дисертацията (103 броя)

D1. **P.N. Penchev**, A.N. Sohou and G.N. Andreev; **Spectrosc. Letters**, **29**, 1513-1522 (1996).

C1. W. Zhou, Sh. Xv, C. Liu & J. Zhang; **Appl. Spectrosc. Reviews**, 2016, **51**, 318–332.

C2. W. Zhou, L Xie, Y Ying; **Transact. Chin. Soc. Agricult. Engineering**, **29**, 285-292 (2013).

C3. W. Zhou, Y. Ying, L. Xie; **App. Spectrosc. Reviews**, **47**, 654–670, (2012).

C4. Li J.F., Fan B.T., Doucet J.P., Annick Panaye; **Appli. Spectros.**, **57**, 858-867 (2003).

C5. В.И. Вершинин, Б.Г. Дерендяев, К.С. Лебедев; *Компьютерная идентификация органических соединений*. (монография, 182 стр.) **Изд. «Наука»**, 2002, Москва.

C6. N.T. Kawai, J.A. Janni; **Spectroscopy**, **15**, 33-41 (2000).

C7. M.L. McKelvy, T.R. Britt, B.L. Davis, J.K. Gillie, F.B. Graves, L.A. Lentz; *Infrared spectroscopy*. **Anal. Chem.**, **70**, R119-R177 (1998).

D2. K. Varmuza, **P. Penchev**, F. Stancl, W. Werther; **J. Mol. Struct.**, **408/409**, 91-96 (1997).

C8. Karoly Héberger; Chemoinformatics—multivariate mathematical–statistical methods for data evaluation in K. Vékey, A. Telekes, A. Vertes (Eds) - **Medical Applications of Mass Spectrometry**, Elsevier (2008)

C9. Markus Meringer; *Mathematische Modelle fur die kombinatorische Chemie und die molekulare Strukturaufklarung*. **Dr. rer. nat. Dissertation**, Der Universitat Bayreuth, (2004), стр. 340.

C10. Kascheres C., Negri G., Ferreira M.M.C., Sabino L.C.; **J. Chem. Society-Perkin Transactions. 2** (12): 2237-2243 (2001).

D3. K. Varmuza, **P. Penchev**, H. Scsibrany. **J. Chem. Inf. Computer Sci.**, **38**, 420-427 (1998).

C11. W. Zhou, Sh. Xv, C. Liu & J. Zhang; **Appl. Spectrosc. Reviews**, 2016, **51**, 318–332.

C12. А.Н. Морозов, И.В. Кочиков, А.В. Новгородская, А.А. Сологуб, И.Л. Фуфурин; **Компьютерная оптика**, 2015, **39**, 614-621.

C13. Оксана Сергеевна Фирюлина; *Алгоритмы поиска максимальных независимых множеств графа и экспериментальная оценка их эффективности*. **Диссертация на соискание ученой степени кандидата физико-математических наук**. Санкт-Петербургский Государственный Университет, 2014.

C14. Farhan N. Rabah, Salah A. Aliesawi, Zahraa Z. Abdulkareem; *PCA and DWT with Resilient ANN based Organic Compounds Charts Recognition.* **International Journal of Computer Applications**, **88** (1), 22-27. (2014).

C15. W. Zhou, Y. Ying, L. Xie; **App. Spectrosc. Reviews**, **47**, 654–670, (2012).

C16. Mikhail Elyashberg, Antony Williams and Kirill Blinov; *Contemporary Computer-Assisted Approaches to Molecular Structure Elucidation. Chapter 1. General Principles of CASE Systems*. **Royal Society of Chemistry** (2012).

C17. Bruni Silvia, De Luca Eleonora, Guglielmi Vittoria, Federica Pozzi. **Appl. Spectrosc.**, **65**, 1017-1023 (2011).

C18. Cerruela Garcia Gonzalo; Luque Ruiz Irene; Angel Gomez-Nieto Miguel; **J. Chem. Inf. Modeling**. **51**, 1216-1232 (2011).

C19. Roberto Todeschini, Viviana Consonni; *Molecular Descriptors for Chemoinformatics*. **Wiley-VCH Verlag GmbH & Co.**, Weinheim (2009).

C20. Elyashberg ME, Williams A, Martin GE; **Progress in Nuclear Magnetic Resonance Spectroscopy, 53** (1-2), 1-104 (2008).

C21. Makarov, L.I.; *Metric properties of the functions of distances between molecular graphs.* **Journal of Structural Chemistry, 48** (2): 219-224 *(*2007*).*

C22. Л. И. Макаров; *Особые вершины взвешенного графа выборки*. **Автометрия**, № 5, т. **41**, с. 92-98 (2005).

C23. Manuel Urbano Cuadrado; *Desarrollo de un lims y una plataforma para la automatización de procesos analíticos continuos basados en la tecnología orientada a objetos. desarrollo y uso de métodos quimiométricos para el tratamiento de datos espectroscópicos*. **Tesis Doctoral**, pp. 349, Universidad de Córdoba (2005).

C24. Makarov LI; *Estimations of subgraph positions in molecular graphs and their common subgraph peculiarities*. **Journal of Structural Chemistry 46** (4): 738-743 (2005).

C25. Derendyaev B.G., Bogdanova T.F., Piottukh-Peletsky V.N., Makarov L.I.; **Anal. Chim. Acta. 509**, 209-216 (2004).

C26. Lingran Chen*; Substructure and Maximal Common Substructure Searching*. In: P. Bultinck, H. De Winter, W. Langenaeker, J.P. Tollenaere (Eds*.); Computational Medicinal Chemistry for Drug Discovery*. Marcel Dekker Inc, **Taylor & Francis Group LLC** (2004), стр. 525 в PDF.

C27. Derendyaev B.G., Bogdanova T.F., Makarov L.I., Piottukh-Peletsky V.N.,; **J. Struct. Chem. 44** (4), 581-586 (2003).

C28. Makarov L.I.; *Taxonomy algorithm for molecular graphs*. **MATCH-Communications in Mathematical and in Computer Chemistry. 49**, 171-178 (2003).

C29. Rucker C., Rucker G., Meringer M.; **J. Chem. Inf. Comp. Sci. 42** (3), 640-650 (2002).
C30. В.И. Вершинин, Б.Г. Дерендяев, К.С. Лебедев; *Компьютерная идентификация органических соединений*. (монография, 182 стр.) **Изд. «Наука»**, 2002, Москва.
C31. Дерендяев Б.Г., Пиоттух-Пелецкий В.Н., Чмутина К.С., Нехорошев С.А.; **Химия в интересах устойчивого развития**, **9** 405-416 (2001).
C32. Derendyaev B.G., Mashkov V.E., Piottukh-Peletskii V.N., Nekhoroshev, SA; *Computer-assisted verification of agreement between IR spectrum and hypothetical chemical structure*. **J. Struct. Chem., 42** (2), 261-270 (2001).
C33. Derendyaev B.G., Makarov L.I., Bogdanova T.F., Piottukh-Peletskii V.N.; **J. Struct. Chem., 42** (2), 271-280 (2001).
C34. Joffre R., Agren G.I., Gillon D., Bosatta, E; **OIKOS**. **93**, 451-458 (2001).
C35. Ehrentreich F.; *Three-step procedure for infrared spectrum interpretation*. **Anal. Chim. Acta. 427**, 233-244 (2001).
C36. Piottukh-Peletsky V.N., Korobeinicheva I.K., Bogdanova T.F., Molodtsov S.G., Derendyaev B.G.; *Exhaustive set of non-isomorphic sub-graphs and its application to chemical structure elucidation using a IR spectroscopy database*. **Anal. Chim. Acta. 409** (1-2), 181-195 (2000).
C37. M.E. Elyashberg, K.A. Blinov, E.R. Martirosian; *A new approach to computer-aided molecular structure elucidation: the expert system Structure Elucidator*. **Lab. Autom. Inf. Management, 34**, 15-30 (1999).
C38. Piottukh-Peletsky V.N., Derendyaev B.G.; *Which IR search system is better for selection of unknown structure analogues?* **Anal. Chim. Acta, 396** (1), 99-103 (1999).
C39. Ehrentreich F.; *Joined knowledge- and signal processing for infrared spectrum interpretation*. **Anal. Chim. Acta, 393**, 193-200 (1999).
C40. Mikhail E. Elyashberg; *Expert systems for structure elucidation of organic molecules by spectral methods*: **Russ. Chem. Rev**., 1999, **68** (7), 525-547.
C41. Elyashberg M.E.; *Expert systems for the determination of structures of organic molecules by spectral methods*. **Usp. Khim**., **68**, 579-604 (1999).
D4. **P.N. Penchev**, N.T. Kochev, G.N. Andreev; **Comptes Rendus de l'Academie Bulgare des Sciences**, **51**, 67-70 (1998).
C42. W. Zhou, Sh. Xv, C. Liu & J. Zhang; **Appl. Spectrosc. Reviews**, 2016, **51**, 318–332.
C43. W. Zhou, Y. Ying, L. Xie; **App. Spectrosc. Reviews**, **47**, 654–670, (2012).
C44. S. Dagnon, R. Tasheva, A. Stoilova, D. Christeva, and A. Edreva; Evaluation of Aroma in Oriental Tobaccos as Based On Valeric Acid Gas Chromatography. **Beiträge zur Tabakforschung International**, 2008, **23**, 115-120.
C45. Li J.F., Fan B.T., Doucet J.P., Annick Panaye; **Applied Spectroscopy, 57** (7): 858-867 (2003)
D5. K. Varmuza, **P.N. Penchev**, H. Scsibrany; **Vibrational Spectroscopy**, **19**, 407-412 (1999).
C46. А.Н. Морозов, И.В. Кочиков, А.В. Новгородская, А.А. Сологуб, И.Л. Фуфурин; **Компьютерная оптика**, 2015, **39**, 614-621.
C47. W. Zhou, Y. Ying, L. Xie; **App. Spectrosc. Reviews**, **47**, 654–670 (2012).
C48. W. Zhou, L. Xie, Y. Ying; **Transact. Chinese Soc. Agricult. Engineering**, **29**, 285-292 (2013).
C49. Kavak H., Esen R.; *Computer-assisted infrared spectra interpretation for amorphous silicon alloys*. **J. Quantit. Spectrosc. & Radiative Transf. 96**, 525-535 (2005).
C50. Derendyaev B.G., Piottukh-Peletskii V.N., Chmutina K.S., Zhbankov R.G., Korolevich M.V. **J. Applied Spectrosc.**, **70**, 615-627 (2003).

C51. Piottukh-Peletskii V.N., Chmutina K.S., Korolevich M.V.; **J. Struct. Chem. 44**, 763-770 (2003).
C52. Derendyaev B.G., Bogdanova T.F., Piottukh-Peletsky V.N., L.I. Mak; **Anal. Chim. Acta. 509**, 209-216 (2004).
C53. Joffre R., Agren G.I., Gillon D., Bosatta, E; **OIKOS. 93**, 451-458 (2001).
C54. Barbara Debska, Barbara Guzowska-Swider, and Daniel Cabrol-Bass; **J. Chem. Inf. Comp. Sci.**, **40**, 330-338 (2000).
D6. **Plamen N. Penchev**, George N. Andreev, Kurt Varmuza; **Analytica Chimica Acta**, **388**, 145-159 (1999).
C55. Wanhuai Zhou, Shoudong Xv, Congjiu Liu & Jianfeng Zhang; **Applied Spectroscopy Reviews**, 2016, **51**, 318–332.
C56. Farhan N. Rabah, Salah A. Aliesawi, Zahraa Z. Abdulkareem; **International Journal of Computer Applications**, **88**, 22-27. (2014).
C57. W. Zhou, Y. Ying, L. Xie; **Applied Spectroscopy Reviews**, **47**, 654–670, (2012).
C58. Gina Tiron, Steluța Gosav; **Romanian Reports in Physics**, **62**, (2), 405–413 (2010).
C59. K. Banas, A. Banas, H. O. Moser, M. Bahou, W. Li, P. Yang, M. Cholewa, S. K. Lim; **Anal. Chem., 82**, 3038–3044 (2010).
C60. A. Bagheri Garmarudi, M. Khanmohammadi, N. Khoddami, K. Shabani; **Intern. J. Math., Computat., Phy., Electr. Comput. Engineering**, 2010, **4**, 1367-1369.
C61. Mohammadreza Khanmohammadi, Amir Bagheri Garmarudi, Nafiseh Khoddami, Keyvan Shabani, Mohammadreza Khanlari; **Microchem. J.**, **95** 337–340 (2010).
C62. Khanmohammadi M., Garmarudi A.B., Ghasemi K.; **J. Chemom.**, **23**, 538-544 (2009).
C63. Gosav, S.; Praisler, M.; **Romanian J. Physics**, **54**, 929-935 (2009).
C64. Katherine Ubina Flores Rojas; *Determinación no Destructiva de Parámetros de Calidad de Frutas y Hortalizas Mediante Espectroscopía de Reflectancia en El Infrarrojo Cercano*. **Tesis Doctoral**, Universidad de Córdoba, (2009).
C65. Markus C. Hemmer; *Expert Systems in Chemistry Research*. **CRC Press, Taylor & Francis Group** (2008).
C66. Judge K, Brown CW, Hamel L; *Sensitivity of infrared spectra to chemical functional groups*. **Anal. Chem., 80**, 4186-4192 (2008).
C67. Karpushkin E., Bogomolov A., Zhukov Y.,; **Chem. Intell. Lab. Syst. 88,** 107-117 (2007).
C68. Gosav S, Praisler M, Dorohoi DO; **J. Mol. Struct., 834-836**, 188-194 Sp. Iss. (2007).
C69. Gillon, D, Bosatta, E, Gosav S, Praisler M, Dorohoi DO, Popa, G; **Talanta, 70**, 922-928 (2006).
C70. Gosav S, Praisler M, Van Bocxlaer J, Van Bocxlaer, J, Massart, DL; **Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy, 64** (5): 1110-1117 Sp. Iss. (2006).
C71. Zeev B. Alfassi, Zvi Boger, Yigal Ronen; *Statistical Treatment of Analytical Data*. **Blackwell Science**, CRC Press (2005).
C72. Markus Hemmer; *Radial Distribution Functions in Computational Chemistry – Theory and Application*. **Dr. rer. nat. Dissertation**, Friedrich-Alexander-Universität Erlangen-Nürnberg, (2004). стр 138.
C73. Markus Meringer; *Mathematische Modelle fur die kombinatorische Chemie und die molekulare Strukturaufklarung*. **Dr. rer. nat. Dissertation**, Der Universitat Bayreuth, (2004), стр. 336.
C74. Praisler M., Van Bocxlaer J., De Leenheer A., Massart, DL; **J. Chromat. A, 962**, 161-173 (2002).
C75. Schoonjans V., Massart D.L.; **J. Pharmac. Biomed. Anal., 26**, 225-239 (2001).

C76. Schoonjans V., Questier F., Guo Q., Van der Heyden, Y, Massart, D.L.; **J. Pharmac. Biomed. Anal., 24** (4), 613-627 (2001).

C77. B. Debska, B. Guzowska-Swider, D. Cabrol-Bass; **J. Chem. Inf. Comp. Sci.**, **40**, 330-338 (2000).

C78. M. Hemmer and J. Aires-de-Sousa, Structure-Spectra Correlations. In: J. Gasteiger and T. Engel (Eds.), **Chemoinformatics**. Wiley-VCH, Berlin 2003.

C79. M.C. Hemmer. Expert Systems. **In "Handbook of Vibrational Spectroscopy"**, Volume 3, Edited by J. M. Chalmers and P.R. Griffiths, *John Wiley & Sons, Ltd*, Volume 3, pp. 1962-1982 (2002).

D7. **Plamen N. Penchev**, Kurt Varmuza; **Comput. & Chem.**, **25,** 231–237 (2001)

C80. А.Н. Морозов, И.В. Кочиков, А.В. Новгородская, А.А. Сологуб, И.Л. Фуфурин; **Компьютерная оптика**, 2015, **39**, 614-621.

C81. Zeev B. Alfassi, Zvi Boger, Yigal Ronen; *Statistical Treatment of Analytical Data*. **Blackwell Science**, CRC Press (2005).

C82. Li J.F., Fan B.T., Doucet J.P., Annick Panaye; Spectral code index (SPECOIND): *A general infrared spectral database search method*. **Appl. Spectros., 57**, 858-867 (2003).

D8. Kurt Varmuza, Nikolay T. Kochev and **Plamen N. Penchev**; **Anal. Sci.**, **17**, i659-i662 (2001).

C83. W. Zhou, Sh. Xv, C. Liu & J. Zhang; **Appl. Spectrosc. Reviews**, 2016, **51**, 318–332.

C84. А.Н. Морозов, И.В. Кочиков, А.В. Новгородская, А.А. Сологуб, И.Л. Фуфурин; **Компьютерная оптика**, 2015, **39**, 614-621.

C85. W. Zhou, Y. Ying, L. Xie; **App. Spectrosc. Reviews**, **47**, 654–670, (2012).

D9. **P.N. Penchev**, V.L. Miteva, A.N. Sohou, N.T. Kochev, G.N. Andreev; **Bulg. Chem. Commun.**, **40**, 556-560 (2008).

C86. W. Zhou, Sh. Xv, C. Liu & J. Zhang; **Appl. Spectrosc. Reviews**, 2016, **51**, 318–332.

C87. W Zhou, L Xie, Y Ying; **Transactions of the Chinese Society of Agricultural Engineering**, **29**, 285-292 (2013).

C88. Y Fang, Y Ma, H Li, K Liang, S Wang, H Wang; **Optical Review, 20** (3), 259-265 (2013).

C89. Zhou, W., Xie, L., Ying, Y.; **American Soc. Agricult. Bio. Engineers Annual International Meeting**, **7**, pp. 5952-5958 (2012).

C90. W. Zhou, Y. Ying, L. Xie; **App. Spectrosc. Reviews**, **47**, 654–670, (2012).

D10. **Plamen N. Penchev**, Klaus-Peter Schulz, Morton E. Munk; **J. Chem. Inf. Model.**, **52**, 1513–1528 (2012).

C91. A. Mohamed, C.H. Nguyen, H. Mamitsuka; **Briefings in Bioinf.**, 2015, 1–13.

C92. Mikhail Elyashberg; **Trends in Analytical Chemistry, 69,** 88–97 (2015).

C93. Mark Edgar; **Annu. Rep. Prog. Chem., Sect. B: Org. Chem.**, **109**, 256-274 (2013).

D11. Petko I. Bozov, **Plamen N. Penchev**, Josep Coll; **Nat. Prod. Commun., 9**, 347-350 (2014).

C94. James R. Hanson; **Nat. Prod. Rep.**, 2015, DOI: 10.1039/c5np00087d

D12. Petko I. Bozov, **Plamen N. Penchev** and Josep Coll; neo-*Clerodane Diterpenoids from Scutellaria galericulata*. **Natural Product Communications. 9** (3), 347-350 (2014).

C95. James R. Hanson; **Nat. Prod. Rep.**, 2015, DOI: 10.1039/c5np00087d

D13. **P. Penchev**, O. Argirov, G. Andreev; **Anal. Laboratory**, **3** (1), 23-28 (1994).

C96. A. Eghbaldar, T. P. Forrest and D. Cabrol-Bass; **Anal. Chim. Acta**, **359**, 283-301 (1998), цитат [36].

D14. **Plamen N. Penchev**, Nikolay T. Kotchev, George N. Andreev; **Traveaux Scientifiques d'Universite de Plovdiv**, **29** (5), 21-26 (2000).

C97. Markus C. Hemmer; *Expert Systems in Chemistry Research*. **CRC Press, Taylor & Francis Group** (2008).

C98. Markus Hemmer*; Radial Distribution Functions in Computational Chemistry – Theory and Application*. **Dr. rer. nat. Dissertation**, Friedrich-Alexander-Universität Erlangen-Nürnberg, (2004). стр 137.

C99. M. Hemmer and J. Aires-de-Sousa, *Structure-Spectra Correlations*: in J. Gasteiger and T. Engel (Eds.), **Chemoinformatics**. Wiley-VCH, Berlin (2003).

C100. M.C. Hemmer. *Expert Systems*. **In "Handbook of Vibrational Spectroscopy"**, Volume 3, Edited by J. M. Chalmers and P.R. Griffiths, *John Wiley & Sons, Ltd*, Volume 3, pp. 1962-1982 (2002).

D15. D. Hristozov, **P. Penchev**, G. Andreev; **Traveaux Scientifiques d'Universite de Plovdiv, 30**, 63-66 (2001).

C101. S. Vishnu, R. Nidamanuri, R. Bremananth; **Geocarto International,** 1–20 (2012).

C102. W. Zhou, Y. Ying, L. Xie; **App. Spectrosc. Reviews**, **47**, 654–670, (2012).

C103. Boris L. Milman; *Reliability and Errors of Identification*, In: B.L. Milman (Ed.) **Chemical Identification and its Quality Assurance**, pp 63-113, Springer (2011), стр. 97,112 цитат [103].