ПЛОВДИВСКИ УНИВЕРСИТЕТ „ПАИСИЙ ХИЛЕНДАРСКИ“

ХИМИЧЕСКИ ФАКУЛТЕТ

КАТЕДРА АНАЛИТИЧНА ХИМИЯ И КОМПЮТЪРНА ХИМИЯ

СЛАВА ХРИСТОВА ЦОНЕВА

КОМПЮТЪРНА ИНТЕРПРЕТАЦИЯ НА ВИБРАЦИОННИ СПЕКТРИ

АВТОРЕФЕРАТ

ЗА ПОЛУЧАВАНЕ НА ОБРАЗОВАТЕЛНА И НАУЧНА СТЕПЕН

ДОКТОР

Научен ръководител

Проф. дхн. Пламен Н. Пенчев

Пловдив

2017

Дисертационният труд е обсъден и насрочен за защита от катедрен съвет на катедра Аналитична химия и компютърна химия на Химически факултет при Пловдивдски университет „Паисий Хилендарски“ на 15 .09.2017г.

Дисертацията съдържа 170 страници, 21 таблици, 31 фигури и 35 уравнения. Библиографската справка обхваща 247 статии, в т.ч. книги, енциклопедии, ръководства и Интернет страници. По дисертацията са отпечатани 5 публикации, от които 2 в българско списание с импакт фактор.

Материалите по защитата са на разположение на интересуващите се в отдел „Развитие на академичния състав и докторантури” към ПУ „Паисий Хилендарски” и Националния център за информация и документация към Министерството на образованието, младежта и науката.

Научно жури:

Защитата на дисертационния труд ще се състои на …….. г. от ……. ч. в …. …… на ПУ “Паисий Хилендарски“, ул. “Цар Асен” № 24, на заседание на Научно жури. Материалите по защитата са на разположение на интересуващите се в Централна библиотека на ПУ „Паисий Хилендарски“.

1. ВЪВЕДЕНИЕ

Прогресивното развитие на спектроскопията като наука е тясно обвързано с все по-бързите темпове, с които се разширява обхватът на приложение на компютърните системи. Изследователите целенасочено търсят ефективното сечение сред инструменталните методи за анализ в стремежите си да разкриват обусловени връзки между признаци и свойства на разнообразието от обекти в реалния свят. Начините, по които електромагнитното лъчение (ЕМЛ) въздейства върху дадена проба, ни дават възможности да бъдат забелязани характеристики в молекулната ѝ структура, като напр. присъствието на функционални групи и характеристични структурни фрагменти (ИЧ и Раман), обкръжението на Н и С атоми (ЯМР) или степента на спрежение (УВ-Вид).

Вибрационните спектрални методи са сред най-често използваните за получаване на евтина и надеждна спектрална информация, с които се работи широко в областта на класическата и модерната органична химия. В зависимост от поставените цели и очакваните резултати, експерименталните основи на един подобен анализ биха могли да се разпростират от прякото сравняване на спектрите и функционален анализ на неизвестно съединение до теоретично пресмятане на честоти и интензитети за набор от нормални трептения. Най-широко разпространение в практиката е получил т.нар. корелационен подход **[1]**, който се позовава в търсенето на пряка зависимост между положението, интензитета и ширината на спектралните ивици и породилите ги молекулни свойства.

Сложната задача на интерпретирането на вибрационен спектър на непознато съединение често изисква усвояването на един или няколко спектрални метода – ИЧ, МС, ЯМР, Хроматография, – които в съчетание с аналитичните методи да допринесат за повишаване на надеждността и ефикасността на получаваните резултати. „Създаването" на човек-експерт в област, която да е едновременно и тясно специализирана в една конкретна, и широко приложима в други научни сфери, е един доста трудо- и времеемък процес. Това е свързано с необходимостта от усвояване на огромни масиви от числа, изучаването на теоретичните основи на метода/-ите/, усвояването на една или няколко експериментални техники. Подобна неоправдано силова задача за един специалист, може спокойно да послужи като оправдание, и дори предпоставка за внедряването на отначало полуавтоматизирани, а в ерата на технологиите, и на компютърно-подпомогнати методи за интерпретация на спектрална информация.

Налагането на компютърното управление на спектрометрите от своя страна създава предпоставки за коренни промени в работата на химика, а именно: внедряване на методи на „изкуствен интелект", хемометричен подход, както и методите на формалната логика и теорията на информацията с цел извличане на нужното знание. Безспорно електроизчислителна техника значително превъзхожда човешкия мозък по бързина, точност, надеждност при установени условия, освен това е по-непредубедена, работи без „симпатии" или субективни ефекти. И нещо много важно – възможност за разрешаване на сложни спектроскопски задачи, свързани с идентифицирането на смеси (GC/IR/MS).

Бързото увеличаване на компютърната мощност създава възможност за извършване на изчисления на все по-високо ниво, което води до получаването на резултати, намиращи се в много по-добро съгласие с данните от експеримента.

2. ЦЕЛИ И ЗАДАЧИ

На основание направения литературен обзор беше формулирана ***ЦЕЛТА*** на настоящата дисертация: ***подобряване на някои от методите за търсене в библиотеки от вибрационни спектри, както и изследване на зависимостта между структурно и спектрално подобие за тези спектри***.

За постигането на целта бяха поставени следните ***ЗАДАЧИ***:

- Създаване на библиотека от ATR спектри;
- Допълване на библиотеката от Раман спектри с около 70 нови спектъра;
- Оптимизиране на толерансите при търсене по ивици на ИЧ/Раман спектри за идентификация;
- Проверка на библиотечното търсене на ATR спектри в библиотеки от ИЧ спектри на поглъщане, както и търсене в обратния ред;
- Изследване на зависимостта между структурно и спектрално подобие за ИЧ/Раман спектри при метод на търсене по пикове;
- Сравнение на зависимостта между структурно и спектрално подобие за ИЧ и Раман спектри;
- Създаване на нова методика за анализ на ИЧ и Раман спектри на смеси.

3. СПЕКТРАЛНИ БИБЛИОТЕКИ, СОФТУЕР И АЛГОРИТМИ

3.1. Спектрални библиотеки:

В лабораторията по молекулна спектроскопия са измерени допълнително над 200 нови библиотечни спектъра (в т.ч. ИЧ и Раман) и е създадена една изцяло нова библиотека, съставена от ATR спектри на органични съединения. Част от тях са използвани за обогатяване на БД, други – за реализиране на тестовите анализи към настоящия дисертационен труд.

За постигане на поставените по-горе задачи са проведени анализи с няколко различни спектрални библиотеки, които съдържат предимно спектри на органични съединения от различни класове и за регистрирането на които е използвана различна техника (Таблица 1).

Таблица 1. Спектрални библиотеки, с които са извършени анализите

Наименование на библиотека	Общ брой спектри в библиотеката	Описание и апаратура
IR	13484	• органични съединения • създадена от Chemical Concepts на апарат Bruker IFS 85 • разделителна способност 1.93 cm^{-1} в интервала 4000-400 cm^{-1} • адаптирана да работи с потребителска програма IRSS
IR01-IR06	911	• органични съединения • измерени в нашата лаборатория на апарати VERTEX 70 и Perkin-Elmer 1750 • разделителна способност 2 cm^{-1} в интервала 4000-400 cm^{-1} за VERTEX 70 и 4 cm^{-1} в интервала 4000-450 cm^{-1} за Perkin-Elmer • работи с потребителска програма IRSS
RAR	330	• органични съединения • измерени в нашата лаборатория и допълнени със спектрите на 70 нови съединения на апарат RAM II & RFS-100 • разделителна способност 2 cm^{-1} в интервала 4000-50 cm^{-1} • работи с потребителска програма IRSS

ATR	102	• органични съединения • нова библиотека от спектри на съединения, измерени в нашата лаборатория на апарат VERTEX 70 с MiRacle • разделителна способност 2 cm^{-1} в интервала 4500-600 cm^{-1} • работи с потребителска програма IRSS

Тези библиотеки са използвани за:

- реализирането на експерименталната проверка на алгоритмите за търсене по ивици в ИЧ и Раман спектри с цел определяне на оптималните стойности на толерансите по вълново число ($\Delta\nu$) и по сигнал ($\Delta A/\Delta I$);
- при проверка на библиотечното търсене с цел идентификация на 40 ATR спектъра в библиотека ИЧ спектри и обратно;
- при изследване зависимостта между структурното подобие и спектралното подобие по метода за търсене по ивици в ИЧ и Раман спектрални библиотеки;
- при метода на многопроменливата регресия за идентификация на компоненти на ИЧ и Раман спектри на бинерни смеси;
- за идентификация на неизвестно съединение.

Всички допълнително заснети спектри са на съединения клас химически чисти /"ч. з. а."/ от фирми производители като Merck, Fluka, Aldrich, Dr. Ehrenstorfer, а също и вещества, синтезирани от колеги от други катедри в Химическия факултет. Базовата линия се коригира автоматично със софтуера OPUS версия 6.5 с крива от втора степен, като при зареждането им в програмата IRSS се изглаждат с алгоритъма на Savitzky-Golay **[2]**. Спектрите и съответстващите им структури се съхраняват под формата на JCAMP-DX **[3]** и съответно в *.skc (MDL, Inc. **[4]**) файлове, като се експортират във формат, поддържан от IRSS.

3.2. Обучаващи и тестващи извадки.

Генерирането на обучаваща (learning set) и тестваща/валидираща (test/validation set) имат своята значима роля при установяване на закономерности между изследваните признаци, което е определящо при избора на „работещ" модел, който ще описва добре зависимостите между избраните величини.

За изпълнението на поставените задачи към настоящата дисертация са използвани представените в Таблица 1 спектрални библиотеки. Работата с тях и изборът на обучаващи и тестващи извадки е съобразена главно със състава им. При библиотечно търсене (БТ) по метода на работа с ивици е установено, че резултатите зависят главно от толерансите, които са предварително зададени при реализирането на сравненията, за $\Delta\nu$ и ΔA.

В голямата библиотека IR13484 са сравнени над 90 млн. двойки спектри, и същият брой двойки структури. Спектрите в цялата библиотека са разпределени на случаен принцип в 13 групи по 1000 спектъра. Една от тези извадки е използвана като „неизвестна" (тестваща извадка), а останалите 10 извадки за обучаваща, всички избрани на случаен принцип. При изборът на обучителна и тестваща извадка в RAR, както и в комбинираните библиотеки (IRRa, RaIR, IRRaAu – 185 спектра), бе решено всеки един спектър да бъде потърсен в цялата библиотека, но първият хит (идентичен на търсения спектър) да се премахва от хит-списъка (ХС), за да бъде пренебрегнат ефектът от идентификация на „неизвестното" съединение. Изборът на този подход за осъществяването на библиотечното търсене се обуславя от малката численост на тези библиотеките, което в определени ситуации би могло да компроментира получените резултати, тъй като са пренебрегнати фактори за изчерпателност и представителност.

3.3. Използван софтуер.

3.3.1. OPUS

Софтуерната програма е създадена и лицензирана от фирма Bruker, като се използва за регистриране на спектрална информация, както и обработка и съхранение на спектралните данни, дава възможности за БТ, коригиране на спектри и др.

Програмата е потребителски ориентирана (user friendly) и работи в средата на ОС Windows**[5]**.

3.3.2. Програмата SubMat

С тази програма са генерирани бинарни кодове за подструктурните дескриптори, които се използват за определяне степен на подобие между структури (Уравнение 3).

Програмата предоставя възможности за въвеждане на ограничения по отношение на общите елементи, употребата на изомерен генератор и систематичното генериране на подстуктури, елиминиране на прекалено екзотични подструктури.

Работим с набор от 500 структурни дескриптора, които са ни предоставени от Varmuza [вж. Табл.1., **[6]**].

3.3.3. Платформа с отворен код – SciLab

Програмният език Scilab е използван за реализирането на изчисления при работата с големи масиви от данни **[7]**.

Конзолата на Scilab работи в среда на Windows, open-sourse (Cecill license) е и потребителски ориентирана. Възможностите на този програмен език се разпростират в области като: линейна алгебра; линейна, квадратична и нелинейна оптимизация;л линейна матрична променлива оптимизация, статистическа обработка на данни и т.н.; предоставя възможности за графично изобразяване на данни (2D & 3D), както и някои типове симулации.

3.3.4. Програмата ISISDraw

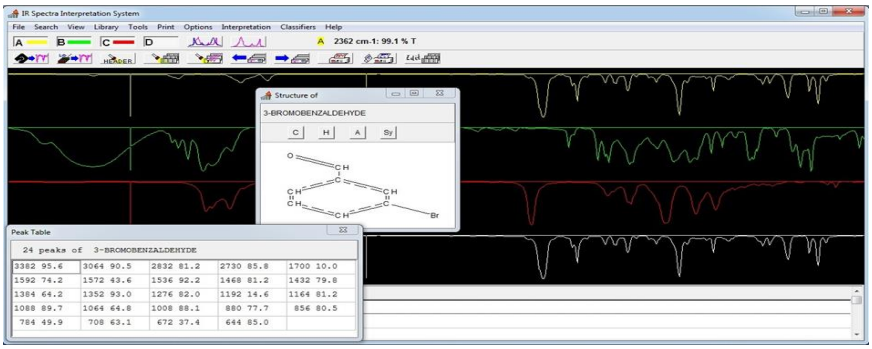
С програмата ISIS Draw са генерирани структурите на съединенията, присъстващи в библиотеката. Тази програма предоставя и възможности за експортиране на файловете във формат *.skc, подходящ за компилиране с програмата за БТ IRSS **[4]** .

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

Програмата IRSS е разработена проф.П.Пенчев **[8], [9]**. Кодът е написан и компилиран на Delphi 1 (16 битова програма) и се състои от над 17 700 реда оператори. Тази програма е потребителски ориентирана, която работи в среда на Windows. Настоящата версия е програмирана през 2000 г. и до този момент подобрявана с допълнителни алгоритми.

Разполага с разнообразни функционални възможности, като: реализиране на БТ на спектри на съединения по три метода – търсене по цяла спектрална крива, по пикове и по химично име; създаване и редактиране на потребителски библиотеки от спектри; преобразуване на спектрален файл от формат JCAMP-DX в библиотечен спектрален файл; интерактивен регресионен анализ на спектъра на смес; изваждане на спектри, с цел търсене на компонентите на дадена смес; а също и възможности за класифициране на ИЧ спектри по набор от подструктури с методите на ИНМ и ЛДА и създаване на kNN-класификатори за химични структури и тяхното прилагане върху резултати от БТ.

На Фигура 1 е представен основният екран на програмата IRSS.



Фигура 1. Основен екран на програмата IRSS с три спектъра, заредени в буферите на програмата и намерен като първи хит в хит-списъка на един от тях при потърсване в БД.

3.4. Използвани алгоритми за спектрално и структурно подобие в ИЧ и Раман спектрални библиотеки.

Изследването и оценяването на връзката между структурното и спектралното подобие е застъпено във всички методи за обработка и интерпретация на спектрална информация, където по подразбиране е водещият интерес за разкриване структурата на търсеното съединение **[10]**, **[11]**. Работата с вибрационни спектри на органични съединения предлага и достатъчно на брой възможности за разкриване на закономорности, но и често пъти затруднения, поради характерното изобилие от данни в самите спектри.

В част от нашите изследвания сме проучвали преносимостта на използваната от идея Varmuza и сътр. **[12]** за разкриването на връзките между спектралното и структурното подобие, но при работа с метода за търсене по пикове, съдържащи данни за интензитет и вълново число на спектралните ивици.

Програмата IRSS разполага с, описани в литературата алгоритми за изчисляване на спектрално подобието между библиотечните спектри и спектъра на съединението, което се определя **[13]**, **[14]** – търсене по пикова таблица и търсене по цяла спектрална крива.

3.4.1. Мерки за спектрално подобие с използване на цяла спектрална крива

Четирите основни алгоритъма за търсене по цяла спектрална крива, използвани като мерки за изчисляване на спектралното подобие: средно квадратично отклонение (с.к.о.), средно абсолютно отклонение (с.а.о.), скаларно произведение (с.п.) и коефициент на корелация (к.к.) **[15]**, са препрограмирани от проф. дхн П.Пенчев в софтуера на програмата IRSS за поддържане на и търсене в библиотеки от ИЧ, ATR и Раман спектри **[8]**.

3.4.2. Мерки за спектрално подобие с използване на пикови таблици.

При библиотечното търсене по метода на работа с пикови таблици се използват в практиката два основни алгоритъма за търсене по пикове – прав (forward) и обратен (reverse) **[11]**, **[14]**. Реализациите на тези алгоритми са заимствани от системата Sadtler **[14]** за тяхното приложение в програмата IRSS, като са извършени редица промени относно практическото им приложение.

Сравняването на непознатия с референтните спектри в БД включва разглеждането на всеки един от тях като математически множества **[15]**, пълното съвпадение между които е желан от теоретична гледна точка резултат, но в практиката се възприема като изключение **[13]**. Последното налага изискването от въвеждане на допуск (tolerance), който да задава интервални граници при търсене на ивиците на референтния спектър, както по абсцисата (вълновото число), така и по ординатата (интензитет на сигнала – абсорбцията/разсеяна светлина). За идентични се считат ивиците между сравняваните спектри, когато попаднат в един спектрален прозорец – правоъгълник с ширина $2\Delta\nu$ и височина $2\Delta A$, определени от потребителя, с център ивицата на референтния спектър. Величините $\Delta\nu$ и ΔA са оптимални толеранси за съвпадение на пиковете, за които е в сила $\Delta\nu \geq |\Delta\nu_k^R - \Delta\nu_l^U|$ и съответно $\Delta A \geq |\Delta A_k^R - \Delta A_l^U|$.

В програмата IRSS броят съвпаднали пикове се определя по т.нар Hungarian алгоритъм, който се използва за намиране на оптималното съвпадение при зададени критерии.

Използваната мярка за спектрално подобие се нарича хит качествен индекс (hit-list quality index, HQI) **[14]** и се изписва с трицифрено число ABC (виж **[15]**), всеки символ от което се изчислява по независим начини:

- **при правия алгоритъм HQI_F = ABC**
- **при обратния HQI_R = BAC**

За тези две мерки за спектрално подобие, обаче, съществува диференциране по отношение на използвания алгоритъм – прав (ABC) или обратен (BAC). Те са несиметрични и числените стойности на двете трансформации (ABC и BAC) се различават, когато се разменят местата на референтното и неизвестното съединение, тъй като зависят от броя ивици в пиковите таблици на двете. Това създава проблеми при изчисляване на спектралното подобие, SpSim, на всички двойки спектри k ≠ l, защото ще имаме неравенство от вида SpSim(k,l) ≠ SpSim(l,k).

Като по-независима е представената нова мярка за спектрално подобие, която е използвана при търсене по пикове – $SpSim_1$. Всички изчисления, реализирани с използването на тази мярка, връщат резултат, независещ пряко от пиковите таблици на сравняваните спектри. Причината е, че отнасянето на броя съвпаднали пикове K между непознатия спектър от M пика и референтния от N пика, е към общия брой пикове и за двата. Тази мярка е подходяща за установяване на спектралното подобие и е нормирана в интервала 0 – 1. Дава се със следния математически израз (Уравнение 1):

$$SpSim_1 = 2K / (M + N)$$

Уравнение 1

Очевидно е, че мярката $SpSim_1$ е симетрична по отношение на избора за последователността на сравнение между спектрите на неизвестното и референтното съединение.

Друга, използвана от нас, мярка подобие между спектрите, е скаларното произведение по пикови таблици, дефинирана с Уравнение 2. Тази мярка е нормирана в интервала 0 – 1, като с A^U_k и A^R_k са отбелязани съответно интензитетите на съвпадналите пикове между референтния и непознатия спектър, а в знаменателя стоят съответните норми на пиковите таблици (Уравнение 2).

$$SpSim_2 = \frac{\sum_k A_k^U A_k^R}{\|A^U\|\|A^R\|}$$

Уравнение 2

Тази мярка също е симетрична по отношение на избора на реда референтен и непознат спектър, като отчита промени на интензитетите на пиковете.

Със зависимостта: $HQI_i = 999*(1- SpSim_i)$, i = 1,2 се преобразува получената числена стойност за спектрално подобие в HQI.

При анализ на смеси предвид участието на пиковете на няколко вещества се очаква определено намаление на HQI_2.

3.4.3. Мерки за структурно подобие.

Оценяването на структурното подобие включва, от една страна, генерирането на т.нар. структурен пръстов отпечатък (fingerprint) на съединенията, участващи в търсенето, и от друга – изчисляване на индекс за определяне степента на съвпадение между структурите на неизвестното и библиотечните съединения **[15]**.

Като количествена оценка е използван индексът на Танимото при оценяване подобието между структурите **[16]**, **[6]**, уравнение (Уравнение 3).

$$Tan_{k,n} = \frac{\sum_m d_{k,m} \; and \; d_{n,m}}{\sum_m d_{k,m} \; or \; d_{n,m}}$$

Уравнение 3

Той е широко разпространен и много подходящ за осъществяване на сравнения между химични структури **[17]**, **[12]**. Определя се като големината на сечението е отнесено към големината на обединението между две структури (Уравнение 3), където обектите $d_{k,m}$ и $d_{n,m}$ са представени с бинарни вектори **[6]**.

За изчисляването структурия „fingerprint“ на съединенията, включени в спектралните библиотеки беше използван набор от 500 подструктури, дефинирани от проф. Вармуца **[12]** и предоставени ни от него като SDF файл. Генерирането на всеки един структурен „fingerprint“ изисква работа с програмата SubMat **[6]**.

При идентични структури числителят и знаменателят са равни на броя единици в съответния пръстов отпечатък, което прави индекса да е равен на единица. Той има стойност нула, при пълно несъответствие между подструктурите в едната и другата структура и обратно.

4. РЕЗУЛТАТИ И ДИСКУСИЯ

4.1. Експериментална проверка на алгоритмите за библиотечно търсене с цел идентификация на съединения по техните ИЧ и Раман спектри.

Описаните в т.3.4.2 мерки за спектрално подобие по метода за БТ по пикове, с които работи и програмата IRSS, зависят от редица параметри, главно отнасящи се до условията за измерване и обработка на спектрите **[18], [19]**. Тяхното влияние, обаче, може да се пренебрегне при спазване на стандартна методика **[20]**. Ето защо идентификационната способност на алгоритмите за търсене при работа с пикова таблица е оптимизирана по стойностите на ΔSig (сигнал, ΔA/ΔI) и Δν.

Прагът, с който са създадени пиковите таблици на спектрите, е избран да е равен на 0.03 a.u. Ефективността на използваните спектралните мерки се определя от броя на правилно идентифицираните съединения, т.е появата на идентичен спектър като първи хит в хит-списъка, получен при библотечното търсене.

Толерансите по ΔSig и Δν са променяни паралелно в следните интервали:

Δν = 3, 4, 5, … 40 cm^{-1} и ΔSig = 0.1, 0.2, ... 1.0 a.u.

При методите за търсене по пикове се наблюдава добра идентификационна способност в случаите с високо структурно подобие между съединенията, главно поради възможността за избор на толеранси по ΔSig и Δν от изследователя.

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

За провеждане на изследванията с ИЧ спектри бяха избрани тридесет (30) тестови спектъра от библиотеки IR01-IR03, за които съответстват идентични в библиотека IR13484. Те са разглеждани като спектри на неизвестни вещества и са подложени на БТ по пикови таблици със съответните мерки за спектрално подобие. Паралелно са извършвани промени на толерансите по вълново число ($\Delta\nu$) и абсорбция (ΔA) в указаните по-горе стойности, а резултатите се съхраняват в текстови формат.

Успоредно с това изследване е проверена и идентификационната способност на алгоритмите по метода за търсене по спектрала крива за същите 30 тестови съединение, при работа и с четирите мерки за спектрално подобие (виж т.3.4.1). Не винаги търсеното съединение се появява като първи хит в ХС. Направената проверка с тези съединения показва, че най-добри резултати са получени със с.п. и к.к., съответно най-лоши са резултатите при работа със с.а.о.

За проведените анализи при работа с пикови таблици бяха използвани три от методите за БТ – прав и обратен алгоритъм и скаларно произведение по пикове, с които са реализирани общо 34 200 (3 x 30 x 38 x 10) претърсвания в ИЧ библиотеката.

Резултатите са усреднени за всички 30 съединения по позицията на хита на идентифицираното съединение в ХС, поотделно за всеки един от използваните методи.

Резултат, даващ средна позиция 1.0 по даден метод за търсене означава, че за тези стойности на ΔA и Δν всички 30 съединения се появяват на първо място в хит-списъка. Съответно

стойност 1.03 означава, че точно едно съединение от тези 30 е второ в хит-списъка, докато останалите 29 са идентифицирани на първо място в хит-списъка, съответно по-високата стойност има двузначно значение.

С най-добра идентификационна способност работи правият метод за търсене по пикове – в широк обхват на толерансите той дава неизвестното съединение като първи хит. Съответно резултатите за обратния алгоритъм имат своето логическо обяснение в зависимостта за HQI (т.3.4.2), а при скаларното произведение по пикове от установена разлика за три от тестовите съединения в пиковети таблици на „непознатия“ и идентичния му спектър. Както и се очаква при увеличаване на Δν се губи селективността на метода, а при най-малката стойност на $\Delta\nu = 3$ cm^{-1} се намалява разпознавателната (идентифициращата) способност, защото нараства броят на несъвпадналите пикове.

За оптимални комбинации на ΔA и Δν са избрани:

- за право търсене: $\Delta A = 0.4 – 0.9$ a.u. и $\Delta\nu = 3 – 7$ cm^{-1}
- за обратно търсене: $\Delta A = 0.4 – 0.8$ a.u. и $\Delta\nu = 3 – 7$ cm^{-1}

За третия алгоритъм се вижда, че ако се пренебрегнат причината за резултатите, се получават следните оптимални интервали за толерансите:

- скаларно произведение $\Delta A = 0.4 – 1.0$ a.u. и $\Delta\nu = 4 – 7$ cm^{-1}

Ясно се вижда приблизителното съвпадение на обхватите, в които се намират оптималните стойности на толерансите за двете величини при правия и обратния алгоритъм. Това може да се възприеме и като 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 сместа. По-лошото представяне на третия алгоритъм за идентификация на чисти вещества не е проблем, тъй като той е въведен за търсене на спектри на смеси **[15]**.

4.1.2. Търсене по ивици (пикове) в Раман спектрите.

Приложимостта на алгоритмите за търсене по двата метода с цел установяване на спектралното подобие е проверена експериментално и при работа с Раман спектрална библиотека. Очаква се, че идентификацията на веществата посредством техните Раман спектри **[21]** е възможна и ще даде полезни резултати, тъй като в тях също е „кодирана“ информация за стуктурите.

За проверката на идетификационната способност на алгоритмите за БТ при работа с Раман спектри бяха допълнително заснети 50 спектъра на съединения, които да имат идентични в библиотеката RAR. Последните са реализирани на апарат RAM II (Bruker Optics), при по ниска мощност на лазера по отношение на съответстващите им библиотечни спектри.

Спектрите на „непознатите“ съединения бяха потърсени в библиотека, като беше обследвана както ефективността на идентификация на съединението, а така също и доколко по-лошото съотношение сигнал/шум (S/N), характерно за Раман спектрите ще се отрази върху резултатите.

В много голям процент от случаите (94-96 %) и за четирите мерки за спектрално подобие по спектрална крива успешно се идентифицират като първи хит търсеното съединение. Единствено в два от случаите – 2-Tiophen-3-ylmethylene-3H-benzo[de]isochromen-1-one и 3-(3-Chloro-benzylidene)-3H-benzo[de]isohromen-1-one 40/40-flated позицията на хита не е сред първите 200 хита. Съответно първото съединение се идентифицира единствено с използването на алгоритъма (к.к.) като 13-ти хит, а второто – 2-ри хит. Второто съединение се идентифицира също и с алгоритмите (с.к.о.) и (с.п.) съответно на 21-во и 22-ро място в хит-списъка.

Четири мерки за спектрално подобие при търсене по пикове са използвани за провеждането на експерименталната проверка – прав и обратен алгоритъм, скаларно произведение по пикове и симетричната мярка. Подобно на изслезването, проведено с ИЧ спектри

(т. 4.1.1) се пренебрегват параметрите, които влияят върху търсенето, с употребата на стандартната методика за измерване и обработване на Раман спектри.

И тук са проверени оптималните стойности на ΔI и Δν относно идентификационната способност на алгоритмите за търсене. Прагът при създаването на пиковите таблици е 0.03 a.i.u., а оценяването на ефективността се определя от броя на правилно идентифицираните съединения, което означава поява като първи хит в ХС.

За определяне на оптималните стойности толерансите за ΔI и Δν са променяни в същите интервали, споменати по-горе (т. 4.1).

Извършен е общ брой потърсвания в библиотеката от 76 000 (4 x 50 x 38 x 10), а резултатите са усреднени за всички 50 съединения по позицията на хита на идентифицираното съединение в ХС, поотделно за всеки един от използваните методи.

Стойност 1.02 по даден метод за търсене означава, че за тези стойности на ΔI и Δν точно едно съединение от тези 50 е второ в ХС, докато останилите 49 са идентифицирани на първо. За резултати като 1.04, 1.06 и 1.08 съществува двузначно обяснение. Липсата на средна позиция 1.0 при четирите алгоритъма се дължи най-вероятно на избора на спектри за непознати с по-лошо съотношение S/N.

При идентификация с използването на правия алгоритъм и симетричната мярка резултатите се припокриват за един и същи интервал от стойности. Обяснението за резултатите при работа с обратния алгоритъм е аналогично, както в т.4.1.1, а скаларното произведение по пикове дава най-добри резултати. Както и се очаква при увеличаване на Δν се губи селективността на метода, а при най-малката стойност на $\Delta\nu = 3\ cm^{-1}$ се намалява разпознавателната (идентифициращата) способност, защото нараства броят на несъвпадналите пикове.

За оптимални комбинации на ΔA и Δν са избрани:

- за право търсене: ΔI = 0.2 – 1.0 a.e. и $\Delta\nu = 4 - 7\ cm^{-1}$
- за обратно търсене: ΔI = 0.3 a.e. и 0.9 a.e. и $\Delta\nu = 4 - 7\ cm^{-1}$

За третия алгоритъм се вижда, че се получават следните оптимални интервали за неопределеностите:

- за скаларно произведение ΔI = 0.2 - 0.3 a.u. и $\Delta\nu = 4 - 15\ cm^{-1}$ & ΔI = 0.4 - 1.0 a.e. и $\Delta\nu = 4 - 11\ cm^{-1}$
- SimSearch: ΔI = 0.2 - 1.0 a.e. и $\Delta\nu = 4 - 7\ cm^{-1}$

Широкият интервал от стойности за най-добрата идентификация на Раман съединенията при търсене по пикове, използвайки скаларното произведение като мярка за подобие, се дължи вероятно на факта, че този алгоритъм работи с относителните интензитети на спектралните характеристики на селектираните съединения.

4.2. Оценка на библиотечното търсене на ATR спектри в библиотека от ИЧ спектри.

Заснемането на ИЧ спектър за дадено съединение в действителност представлява уникална и специфична за него информация **[18]**. Използването, обаче, на различни техники за получаването ѝ, предвид тяхното бързо развитие, изисква и търсене на възможности за неограниченото им използване за целите на библиотечното търсене **[18]**. Макар и все още най-масово да се работи с трансмисионни спектри в широката практика за реализиране на БТ, през последните години все повече се засилва интересът към работата със спектри, заснети с ATR техника. Сред характеристиките, които дават преднина на метода на пълното вътрешно отражение (ATR), са минимизирането на пробоподготовката, както и възможностите за извършване повърхностни анализи **[22]**, и намаляване на площта на контактната повърхност между пробата и облъчващото лъчение **[23]** и др.

Съществуват някои значими разлики между спектрите на съединения, заснети с двата типа техники, произхождащи от различните физически ефекти, които се наблюдават при

взаимодействието на ЕМЛ и веществата **[23]**. Така например, двата вида спектри се диференцират по ширина и относителни интензитети на спектралните ивици, което несъмнено води до чести проблеми при употребата им в библиотечното търсене.

Проведеното от нас изследване е насочено към сравнение на два различни метода за получаване на спектрална информация и доколко това влияе върху резултатите **[18]**. Вниманието ни беше привлечено от изследване на възможности за ефективно използване на „нови“ спектроскопски БД в рутинните анализи – безспорен „критичен“ момент при методите за идентификация на неизвестни съединения в колечествен и качествен аспект.

Използвали сме спектри, заснети с ИЧ техниката и идентични на тях по качествен състав, получени при използване на техниката с пълно вътрешно отражение с едноотражателен елемент (ATR) (Таблица 1).

Избрани са 40 тестови съединения за провеждането на библиотечното търсене на техните ATR спектри в ИЧ библиотека и обратно – ИЧ спектри в ATR библиотека.

Реализацията на библиотечното търсене е осъществено с програмата IRSS **[8]**, **[9]** с метода за търсене по цяла спектрална крива на неизвестно съединение в съответната библиотека със спектрална мярка за подобие коефициент на корелация, а за изчисляване на структурното подобие между структурите в генерирания хитсписък – индексът на Танимото (Уравнение 3).

Процедурата по търсене в библиотека на спектър на непознато съединение е широко използваната в рутинната практика **[10], [24]**. Тя включва сравняване на спектрите на неизвестни съединение с тези на присъстващите в библиотеката с цел да се установи спектрално подобие помежду им. Резултатът е генерирането на списък от спектри на съединения (хитсписък, ХС), подредени в низходящ ред по степен на подобие с търсения и с намалявне на HQI. Идеалният случай, е когато като първи хит в хит-списъка се появи референтен спектър на търсеното съединение. Както може да се види и от получените резултати, обаче, при реални ситуации, поради недостатъчна гъвкавост на използвания алгоритъм, значително влияние на техниката за регистриране на спектралните данни **[18]**, размерът на спектралната библиотека или друго, може да се появи спектърът на търсеното съединение в по-ниски позици в хит-списъка.

Получените резултатите от проведените анализи на търсене на ATR спектри в ИЧ библиотека, са представени в Таблица 2.

Позиция на хита в хит-списъка	Честота на поява.
1	28
2	3
4	3
5	1
7	1
8	1
10	1
12	1
21	1
Всички	40

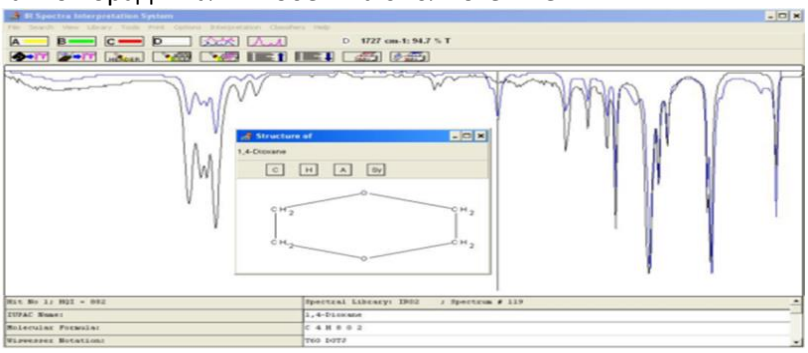
Таблица 2. Позиция в хит-списъка при точна идентификация на съединението и честота на поява при библиотечното търсене.

Получените резултати са задоволителни и показват, че изпълнението и на двете задачи е възможно, въпреки малката численост, както и различния брой спектри в двете библиотеки – ATR (102) и ИЧ (911). Безспорно влияние имат и споменатите по-горе различия в ширина и относителните интензитети на спектралните ивици, което при пълното вътрешно отражение зависи от дължината на вълната, освен това, кристалът, с който е оборудвана техниката ATR, е от ZnSe и поглъща ЕМЛ под 600 cm^{-1}.

В пет от случаите БТ показва зависимост от варирането на интензитетите и на ширините на спектралните ивици, довело до влошаването на резултатите и появата на идентичните референтни

спектри на по-задни позиции в хит-списъка. Наблюдават се и четири случая, при които нарастването на фона на спектъра с намаляване на вълновото число води до по-лоша идентификационна способност на метода, където пък трябва да се отчете влиянието на използвания алгоритъм за фоновата корекция на спектъра.

Получените резултати не могат да бъдат разглеждани като изчерпателни и/или достоверни, главно поради малкия обем на библиотеките.



Фигура 2. Сравнение на спектри, заснети с ИЧ (черната крива) и ATR (синята крива) техника.

Като пример е представен на Фигура 2 ATR спектър но 1,4-Диоксан, който е потърсен в ИЧ спектрална библиотека при използването на мярка за спектрално подобие коефициент на корелация. Въпреки появата на ивицата при 1724 cm^{-1}, показваща възможно разпадане на пробата, и поглъщането на ZnSe под 600 cm^{-1}, и за двата спектъра тези обстоятелства не оказват съществено влияние върху търсенето, което може да се отсъди по стойността на HQI = 888.

Изводът, който може да се направи тук е безспорната ползата от възможността за успешно идентифициране на ATR спектри в ИЧ библиотека и обратно.

4.3. Изследване на връзката между структурното и спектралното подобие в библиотеки от ИЧ и Раман спектри.

4.3.1. Оценяване на средното и кумулативното структурно подобие в библиотека от ИЧ спектри.

За да се оцени структурното подобие в библиотека от ИЧ спектри е използвана библиотеката IR 13484, за която условията за получаване на спектралната информация, както и процедурата по създаването на тестваща и обучителна извадки, са представени в т.3.

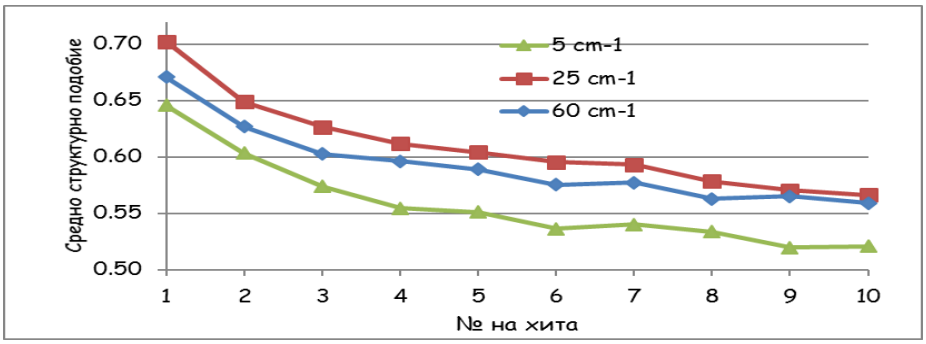
Степента на подобие между структурите на изведените в хит-списъка спектри и структурата на търсеното съединение е изчислена, като си използва индексът на Танимото (Уравнение 3) при потърсването на един от 1000-та ИЧ спектъра в библиотеката от 10 000.

За по-добра оценка на структурното подобие е избрано да бъде приложено усредняване на всички тези стойности на хитовете в хит-списъка по 1000-та търсения. Използвани са два типа усредняване на резултатите, които са: средно структурно подобие и средно структурно подобие чрез натрупване (кумулативно). Оценяването на средното структурно подобие, означаваме с TanA(h), където h е номерът на съответния хит, и се изчислява за всички стойности по първи хит в списъка от 1000 търсения, а кумулативното средно структурно подобие, съответно – е усредната стойност на TanA(h) по хитове от първи до този с номер h, бележи се с TanC(h). Може да се даде с обобщената зависимост (Уравнение 4):

$$TanC(N) = \frac{1}{N}\sum_{i=1}^{N} TanA_i(h)$$

Уравнение 4

При търсенето в Раман спектрална библиотека, TanA(h) се получава при усредняване по 330, появили се на първа позиция в хит-списъка.

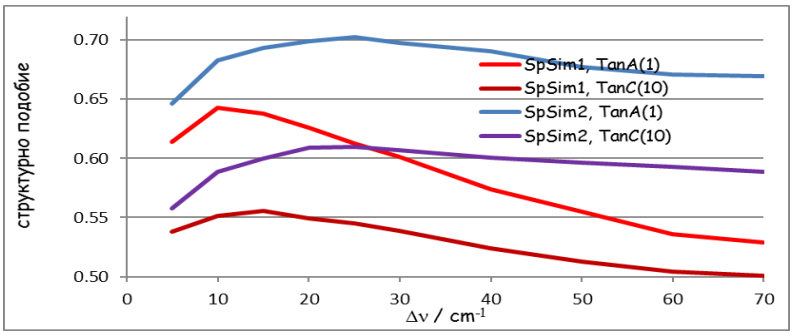


Фигура 3. Средното структурно подобие, TanA(h), като функция от номера на хита за търсене по пикове с (Уравнение 2) на ИЧ спектри при ΔA = 0.7 a.u. и три стойности на Δv.

На Фигура 3 е представената графична зависимост на средното структурно подобие, като функция от номера на хита. Вижда се, че съществува тенденция към намаляване на средното структурно подобие с повишаване номера на хита в ХС. За високите и твърде ниските стойности на на толеранса по вълново число, чувствителността спада, а оптимални резултати биха могли да се очакват при работата с $\Delta v = 25 cm^{-1}$. Това, от своя страна може да означава, че изборът на Δv е фактор, от който зависи средното структурно подобие на спектрално подобните съединения.

На Фигура 4 са представени зависимостите съответно на средното структурно подобие за първи хит и кумулативното средно структурно подобие „от първи до десети хит“ в IR13484. Графично е сравнено как се влияе структурното подобие при провеждането на анализите с двете различни мерки за спектрално подобие – $SpSim_1$ и $SpSim_2$. Могат да се направят следните изводи:

- ✓ структурното подобие при използване на $SpSim_2$ показва по-добри резултати, за което може да се отсъди от по-високите им чувствителности при графичното сравнение на резултатите.
- ✓ кривите за $SpSim_2$ намаляват по-полегато с нарастване на Δν, отколкото кривите за $SpSim_1$, което е следствие зависимостта на скаларното произведение по пикове от относителният интензитет на съвпадналите пикове, за разлика от $SpSim_1$ – броя съвпаднали пикове.
- ✓ добре може да се прецени и максималното структурно подобие, оценено с двете спектрални мерки – между 10 и 15 cm^{-1} за $SpSim_1$ и около 25 cm^{-1} за $SpSim_2$.

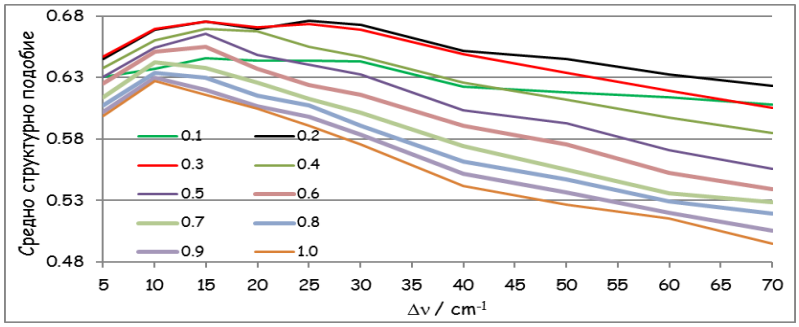


Фигура 4. Средно структурно подобие, TanA(1), и кумулативно средно структурно подобие, TanC(10), като функция от Δv за търсене по пикове на ИЧ спектри при ΔA = 0.7 a.u. и двете спектрални мерки от Уравнение 1 и Уравнение 2.

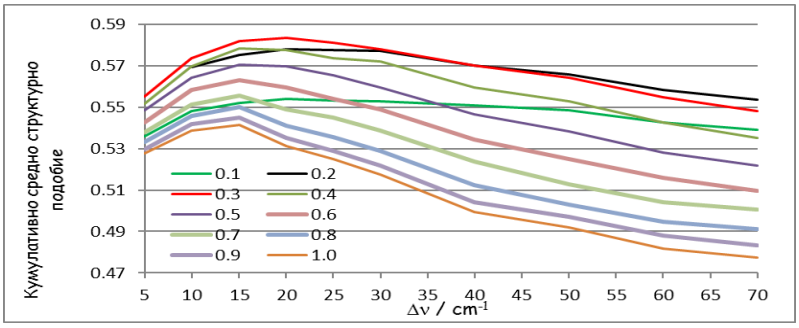
4.3.2. Изследване на връзката между структурното и спектрално подобие в библиотеки от ИЧ спектри спектри при работа с пикови таблици.

Използвана е идеята на Varmuza и съавт. **[12]**, приложена в ИЧ спектрални библотеки по метода на търсене по спектрална крива, който е заменен с метода за търсене по пикове.

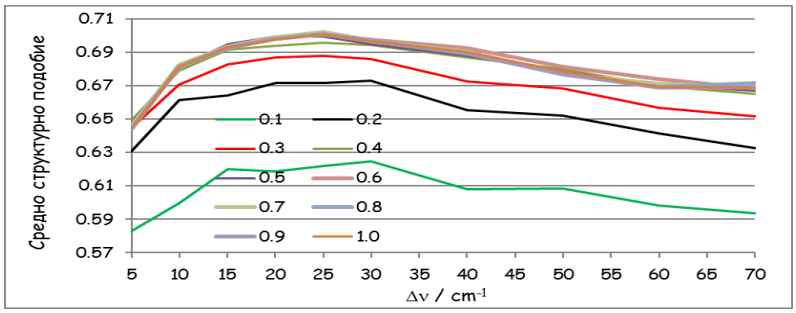
Провеждането на БТ по пикови таблици е съпроводено с определяне на оптималните стойности на толерансите на Δv и ΔA. При тези изследвания t_R и t_U са избрани в еднакви стойности – 0.03 a.u., а за интервалите при сравнение по вълновото число и интензитета на ивиците, са използвани следните интервални стойности: Δv = 5, 10, 15, 20, 25, 30, 40, 50, 60 и 70 cm^{-1}, и съответно за ΔA = 0.1, 0.2 ... 1.0 a.u.



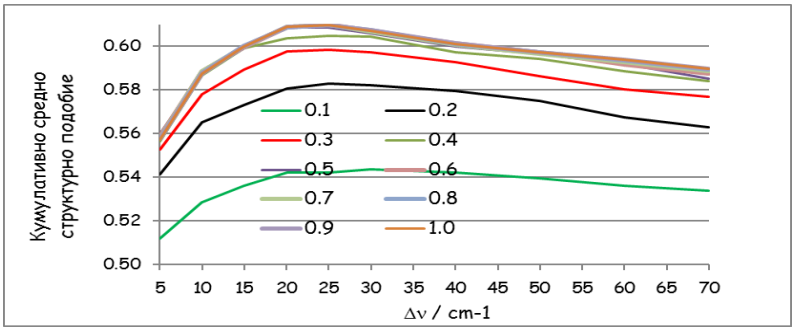
Фигура 5. Средно структурно подобие, TanA(1), като функция от Δv за търсене по пикове на ИЧ спектри при всички ΔA = 0.1, ... 1.0 a.u. и спектралната мярка за подобие от Уравнение 1.



Фигура 6. Кумулативно средно структурно подобие, TanC(10), като функция от Δv за търсене по пикове на ИЧ спектри при всички ΔA = 0.1, ... 1.0 a.u. и спектралната мярка за подобие от Уравнение 1.



Фигура 7. Средно структурно подобие, TanA(1), като функция от Δv за търсене по пикове на ИЧ спектри при всички ΔA = 0.1, ... 1.0 a.u. и спектралната мярка за подобие от Уравнение 2



Фигура 8. Кумулативно средно структурно подобие, TanC(10), като функция от Δv за търсене по пикове на ИЧ спектри при всички ΔA = 0.1, ... 1.0 a.u. и спектралната мярка за подобие от Уравнение 2.

От графиките на Фигура 5 и Фигура 6 могат да се проследят зависимостите на структурното подобие, като функция от толерансите по вълново число. От представените графични зависимости може да се установи какво е средното структурно подобие за първи хит и кумулативното средно структурно подобие за хитовете от 1-ви до 10-ти при използването на спектралната мярка за подобие $SpSim_1$ от уравнение (Уравнение 1). Съответно графиките, когато е използвана мярката $SpSim_2$ от уравнение (Уравнение 2) са на Фигура 7 и Фигура 8.

Получените разултати потвърждават първоначалната информация (Фигура 4) за по-добрата чувствителност на библиотечното търсене в ИЧ библиотека при работата с мярката $SpSim_2$ в сравнение със симетричната мярка – $SpSim_1$. Първата дава по-високи стойности за средното структурно подобие по първия хит, TanA(1): 0.702 за ΔA = 0.7 a.u. и Δv = 25 cm^{-1} за $SpSim_2$ спрямо

0.676 за ΔA = 0.2 a.u. и Δν = 25 cm^{-1} на $SpSim_1$. Същата зависимост се наблюдава и за кумулативното средно структурно подобие, TanC(10): 0.610 (0.60987) за ΔA = 0.6 a.u. и Δν = 25 cm^{-1} срещу 0.583 за ΔA = 0.3 a.u. и Δν = 20 cm^{-1}. При работата с $SpSim_2$ е много близка стойността на TanC(10) за интервалните допуски – ΔA = 0.7 a.u. и Δν=25 cm^{-1} – 0.60975, затова са възприети като оптимални стойности на толерансите ΔA = 0.7 a.u. и Δν= 25 cm^{-1} при работата с ИЧ спектри за тази мярка за спектрално подобие. За сравнение, резултатите, получени при идентификационното търсене в библиотеката на 30 тестови съединения показват следните оптимални толеранси при работата със спектрална мярка за подобие $SpSim_2$: ΔA = 0.4-1.0 a.u. и $\Delta\nu$.= 4-7 cm^{-1} при вариране на стойностите на толерансите в определени интервали (виж т. 4.1.1).

4.3.3. Разпределение на стойностите на структурното подобие.

Спектралната библиотека IR13484, която е използвана за провеждане на системните изследванията, се характеризира с голямо разнообразие от химични съединения, принадлежащи на различни класове, което, от своя страна, предполага известна повтаряемост на структурни или субструктурни фрагменти в тях. Това налага изискване за оценка на средното структурно подобие за всички съединения от библиотеката, а така също и на представителна извадка от нея.

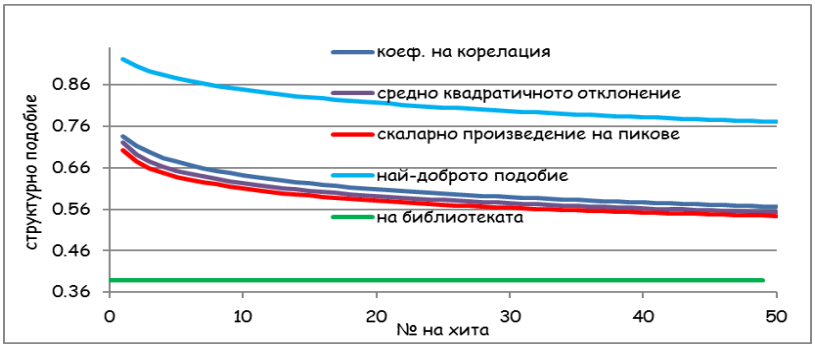
Представените резултати на Фигура 9 дават информация за структурното подобие в цялата спектрална библиотека, за 1-ви и съответно 1 – 10-ти хит в ХС от БТ и на най-доброто структурно

Фигура 9. Разпределение на структурното подобие. Хитовете са получени при търсене по скаларно произведение на пикове с ΔA = 0.7 a.u. и Δν = 25 cm^{-1}.

подобие. С*редното структурно подобие* в спектралната библиотека е изчислено като структурата на всяко 10 (десето) съединение от 1000-та спектъра в тестващата извадката е сравнена с всяка 10 (десета) от 10 000 структури на обучаващата извадка. Това прави изчисляване на общо 100 000 индекса на Танимото по уравнение (Уравнение 3). От друга страна, структурата на всяко от 1000-та „неизвестни“ съединения има своя най-подобна сред 10 000, които при библиотечното търсене, следва да се появят като първи хит в генерирания хитсписък от съединения. Тези 1000 числови стойности, изчислени с индекса на Танимото, съставляват извадка и в нея присъстват структури, които могат да се категоризират като *най-добро структурно подобие*. За съжаление не винаги първи хит, показващ спектрално подобие между сравняваните спектри, дава и подобни по структура съединения, както може да се види от направеното сравнение на Фигура 9. Генерирани са и 1000 индекса на Танимото, показващи средното структурно подобие между двойките спектри („неизвестно“ съединение – референтно съединение, като първи хит) и 1000 стойности на кумулативното средно структурно подобие: последните се получават при усредняване по хитове

4.3.4. Сравнение на търсенето по пикове с търсене по спектрална крива в ИЧ библиотека.

Библиотечното търсене, проведено по метода за търсене по пикове, дава резултати, показващи, от една страна, влиянието на изборът на толеранси по $\Delta\nu$ и ΔA, и от друга – на праг при създаването на пикови таблици при търсенето по подобие. За да се направи преценка на използвания подход за търсене в спектралната библотека, включително и по отношение на използваните мерки за спектрално подобие при оценка на структурното подобие, е направено сравнение, резултатите от което са изобразени графично на Фигура 10.



Фигура 10. Структурното подобие като функция от номера на хита. Търсене по пикове на ИЧ спектри е със спектралната мярка за подобие от Уравнение 2 и неопределености ΔA = 0.7 a.u. и Δν= 25 cm^{-1}.

Както се вижда от сравнените резултати между проведени изследвания при БТ по цяла спектрална крива и по пикове, търсенето по пикове с мярката за спектрално подобие *скаларно произведение* дават по-лоши резултати при прилагането на концепцията за максимална обща подструктура. Кривите на трите метода за търсене в спектралните библиотеки са представени със зависимостта на TanA(h) от h (номера на хита). Хоризонталната крива отчита средното структурно подобие на структурите в библиотеката, 0.389. Получаването на кривата за най-добро структурно подобие е осъществено на няколко стъпки. На първо място, между обучаващата и тестващата извадки е проведено търсене, при което за всяко от 1000 съединения от тестващата извадка е намерено най-добро структурно подобие с всичките 10 000 съединения от обучаващата извадка. На второ място, е извършено сортиране на всеки от хилядата списъка в низходящ ред, в резултат на което се получава един вид хит-списък, но не по спектрално, а по структурно подобие. Накрая, хилядата списъка са усреднени по нарастване на номера в хит-списъка.

Изводът за така представените криви, показващи функцията на разпределение на стойностите на структурното подобие при използването на съответните мерки за спектрално подобие, е, че при използването на трите спектрални мерки за подобие в действителност се отчита и структурното подобие на спектрите спрямо средното структурно подобие в библиотеката, макар и да не е най-доброто.

4.3.5. Изследване на връзката между структурното и спектрално подобие в библиотеки от Раман спектри при работа с пикови таблици.

Раман спектрите, въпреки слабата чувствителност на метода, се характеризират с някои предимства пред традиционните ИЧ спектри на поглъщане **[25]**.

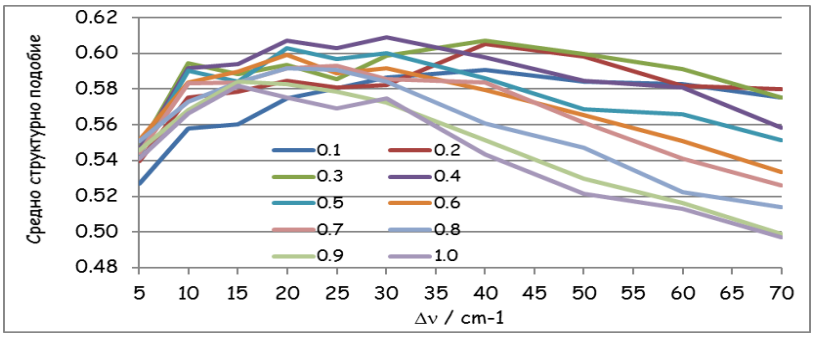
При тези изследвания е приложена гореописаната методика (т.4.3), по идеята на Varmuza и съавт. **[12]**, като е избрана за работа библиотека от Раман спектри. Този метод позволява да се направи количествена оценка на резултатите за търсене по подобие и намиране на оптимални стойности на толерансите по $\Delta\nu$ и ΔI.

За да се оцени структурното подобие в библиотека от Раман спектри е използвана библиотеката RAR от 330 спектъра, за която условията за регистриране на спектрална информация, както и процедурата по създаването на тестваща и обучителна извадки, са представени в т.3.

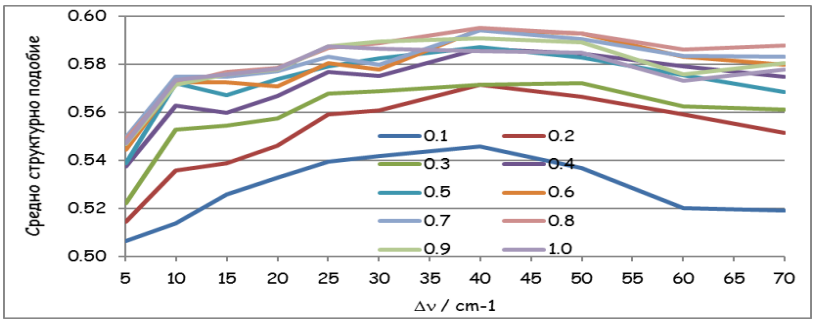
Независимо от малкия обем на Раман спектралната библиотека RAR, е проведено изследване за определяне на спектро-структурни корелации при работа с двете мерки за спектрално подобие от Уравнение 1 и Уравнение 2 и съответно индекс на Танимото (Уравнение 3) за оценяване на структурното подобие между спектрите. Последните са варирани в същите, посочени в т.4.3.2 толеранси, и са използвани същите стойности на праг t_R и t_U (0.03 a.u.).

Тъй като ИЧ и Раман спектрите се разглеждат в един и същи спектрален интервал в спектралните библиотеки, поддържани от софтеура IRSS, то ординатата на Раман спектрите стои интензитет на разсеяното лъчение, при сравнението с резултатите за ИЧ спектри ще отбелязваме с мерната единица – a.i.u. (arbitrary intensity units).

За сравнение, резултатите, получени при идентификационното търсене в библиотеката на 50 тестови съединения показват следните оптимални толеранси при работата със спектрална мярка за подобие $SpSim_2$: ΔI = 0.2-1.0 a.i.u. и Δν = 4-11 cm^{-1} при вариране на стойностите на толерансите в определените интервали.



Фигура 11. Средно структурно подобие, TanA(1), като функция от Δν за търсене по пикове на Раман спектри при всички ΔI = 0.1, ... 1.0 a.u. и спектралната мярка за подобие от Уравнение 1.



Фигура 12. Средно структурно подобие, TanA(1), като функция от Δν за търсене по пикове на Раман спектри при всички ΔI = 0.1, ... 1.0 a.u. и спектралната мярка за подобие от Уравнение 2.

На Фигура 11 и Фигура 12 представени зависимостите на средното структурно подобие от Δν и ΔI при работа с двете спектрални мерки за подобие (Уравнение 1 и Уравнение 2). Сравнението на тези криви с кривите от Фигура 5 и Фигура 7 показва, че те имат същото поведение с промяната на двата толеранса. При използването на спектрална мярка за подобие $SpSim_1$ кривата за ΔI = 0.1 a.u. е най-полегата, където едно от тълкуванията е, че и за Раман спектрите се наблюдава характеристичност по интензитет на ивиците на химичните групи, докато повечето изследователи считат, че при тези спектри тя не е толкова силно изявена както при ИЧ спектрите **[26]**.

Значителното криволичене на линиите на графиките на Фигура 11 и Фигура 12 в сравнение с тези от Фигура 5 и Фигура 7 се обвързва главно с малкия обем на спектралната библиотека RAR.

В сравнение с ИЧ библиотеката при търсенето в Раман спектрална библиотека по-добра се оказва спектралната мярка $SpSim_1$, при оценката на структурното подобие в библиотеката. Доказателство за това са получените резултати: 0.595 за ΔI = 0.8 a.u. и Δν = 40 cm^{-1} срещу 0.609 за ΔI = 0.4 a.u. и Δν = 30 cm^{-1}. Разликата между двете стойности, 0.014, е по-малка от разликата за ИЧ спектри, 0.026 (= 0.702 - 0.676), а освен това имаме по-малка извадка, така че по тези резултати не може да се съди коя от двете мерки за спектрално подобие е по-добра.

Сравняването по абсолютната стойност на структурното подобие в двете библиотеки – IR13484 и RAR, е трудно осъществимо, от една страна, поради значимите различия в състава и размера на двете.

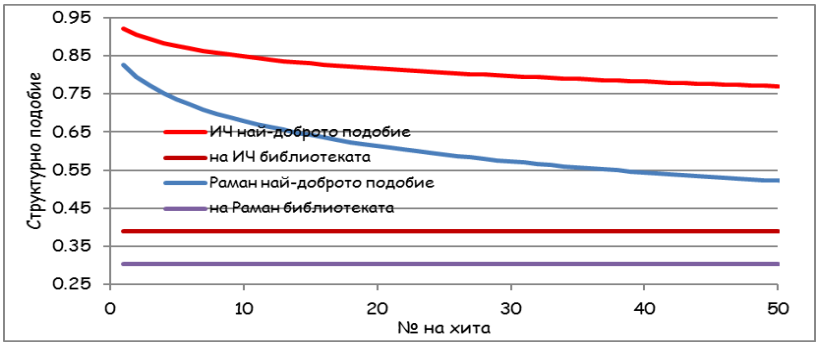
От графиките се вижда, че максималното средно структурно подобие и при двете мерки за оценяване на спектралното подобие (0.595 и 0.609), е близко до средната стойност на най-доброто структурно подобие – 0.8 и по високо от средното структурно подобие – 0.3 за библиотеката (Фигура 15).

Определените оптимални толаренси, при които се отчита максималното струкутрно подобие, за използваните две мерки за спектрално подобие по ΔI и Δν са в следните граници:

- $SpSim_1$ – ΔI=0.3-0.4 a.i.u. и Δν=20-40 cm^{-1}
- $SpSim_2$ – ΔI=0.7-0.9 a.i.u. и Δν=30-55 cm^{-1}

Може да се заключи, че чувствителността на методът намалява при по-ниските стойности по двата толеранса и по-ниска селективност, обратно, при високите им стойности – очевидно е, че при тези две ситуации има съществено намаляване на структурното подобие.

При използването на спектралната мярка $SpSim_2$ по-широкия интервал на толерансите може да се обясни с проява на по-добра селективност, доказателство за което е и струпването на кривите по ординатата (Фигура 12).



Фигура 13. Сравнение на структурното подобие в двете библиотеки.

Макар и максималното структурно подобие в ИЧ библиотека да е по-голямо, тази библиотека не представлява допълнена по структури Раман библиотека, въпреки че сечението на структурите между двете е 102 структури. Това може да се обоснове и с Фигура 13, където те са сравнени подобно на Фигура 10. Вижда се също, че и средното структурно подобие на ИЧ библиотеката е по-голямо от това на Раман библиотеката.

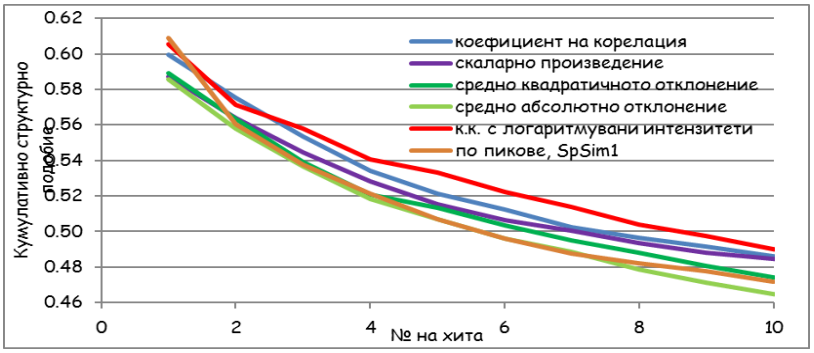
4.3.6. Сравнение на търсенето по пикове с търсене по спектрална крива в Раман библиотека.

Библиотечното търсене, проведено по метода за търсене по пикове в Раман спектрална библиотека, подобно на ИЧ, дава резултати също показващи, влиянието на изборът на толеранси по Δν и ΔI, както и на праг за пикови таблици. За да се направи преценка на използвания подход за търсене в спектралната библиотека, включително и по отношение на използваните мерки за спектрално подобие при оценка на структурното подобие, е направено сравнение, резултатите от което са представени на Фигура 14.

Както се вижда от сравнените резултати между проведени изследвания при БТ по цяла спектрална крива и по пикове, алгоритмите за търсене по пикове дават по-лоши резултати при прилагането на концепцията за максимална обща подструктура. Причината за това може да се търси и в по-малкия брой използвани структури в сравнение с обработвания с тези методи хит-списък. Последното се потвърждава и от сравнението на структурното подобие за структурите на хитовете със структурата на „неизвестното“ съединение.

Търсенето по пикове с по-добрата спектрална мярка за подобие, $SpSim_1$, и с допуските, които дават най-висока стойност на структурното подобие, ΔA = 0.4 a.u. и Δν = 30 cm^{-1}, е сравнено

със спектралните мерки, които използват цяла спектрална крива (виж **[15]**). Използвано е кумулативното структурно подобие, TanC(h) и резултатите са представени на Фигура 14.



Фигура 14. Кумулативното структурно подобие като функция от номера на хита. Търсене по пикове на Раман спектри е със спектралната мярка за подобие от Уравнение 1 и неопределености ΔA = 0.4 a.u. и Δν = 30 cm^{-1}.

Допълнително е сравнена и обработка на спектрите с операцията логаритмуване по Уравнение 5, в резултат на което се прави траснформация на стойностите на спектралните данни по ординатата от тип байт от 0 до 255 в интервала 0–1.

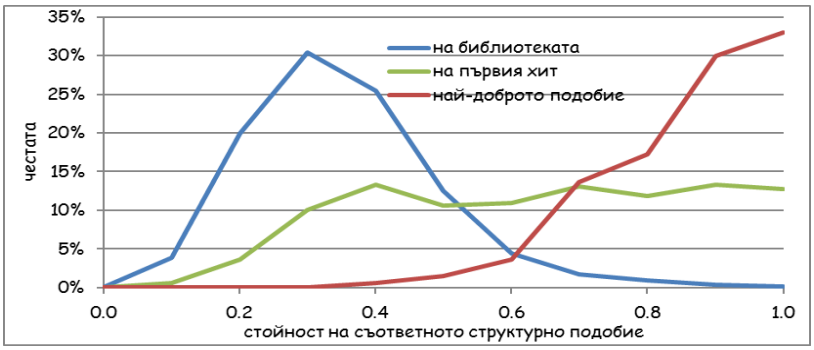
$$A_k log = log_2(A_k+1)/8$$

Уравнение 5

Този подход е избран и с идеята, че характерно за Раман спектрите е наличието на изключително интензивни ивици, което прави относителния интензитет на други ивици много малък. Резултатите показват, че търсенето по пикове с този алгоритъм за спектрално подобие при така избраните параметри е съизмеримо с другите спектрални мерки. Не може обаче да се твърди, че по-добрата чувствителност на последния, което е осезаемо след хит № 6 се дължи на появата в хит-списъка на нови подобни на „неизвестното“ съединения, които биха били изпуснати, заради по-високия интензитет на определени ивици. Ние считаме, че малкият обем на Раман спектралната библиотека не ни дава основание да направим подобно заключение за преносимост на използване статистически модел.

И тук плътността на разпределение на стойностите на структурното подобие е представено чрез хистограмите на Фигура 15. Библиотечното търсене е оценено с коефициента на корелация с метода за търсене по спектрални криви. Би могло да се констатира очевидната прилика на кривите от тази Фигура 15 с тези от Фигура 4, представена по-горе за ИЧ библиотеката.

Всички спектро-структурни корелации се основават търсенето на връзката между спектрални признаци и приписването на молекулно свойство, която е най-явна при търсенето по подобие в спектрални библиотеки и е в основата на метода за най-близките съседи и прилагането на концепцията за максимална обща подструктура при обработка на хит-списъците с резултати.



Фигура 15. Разпределение на структурното подобие за Раман библиотеката. Хитовете са получени при търсене по коефициент на корелация на Раман спектрални криви.

Някои аспекти от изследванията са ограничени на пръв поглед от липсата на големи спектрални библиотеки, но въпреки това сравнението на подобието на двойките спектри с подобието на двойките структури е възможно. Тук трябва да се отбележи, че комбинацията на

двойките спектри, а също и на двойките структури, не дават независими стойности, което е необходимо условие за проверка на редица статистически хипотези. Въпреки тези ограничения, ние считаме, че някои от изложените в тази част резултати са показателни и разкриват донякъде количеството информация, която се съдържа в ИЧ и Раман спектрите.

4.4. Изследване на връзката между структурното и спектралното подобие за ИЧ и Раман спектри на едни и същи съединения.

Двата метода – ИЧ и Раман, са взаимно допълващи се и дават големи възможности при разрешаването на разнообразни и сложни проблеми **[22]**. В тази връзка ориентацията на изследователите е насочена към търсене на все по-ефективни начини за комбинирано изследване на разнообразието от вещества, представляващи интерес, предлагайки различни подходи за комбинирано търсене в разнообразни библиотеки **[27]**, **[28]**.

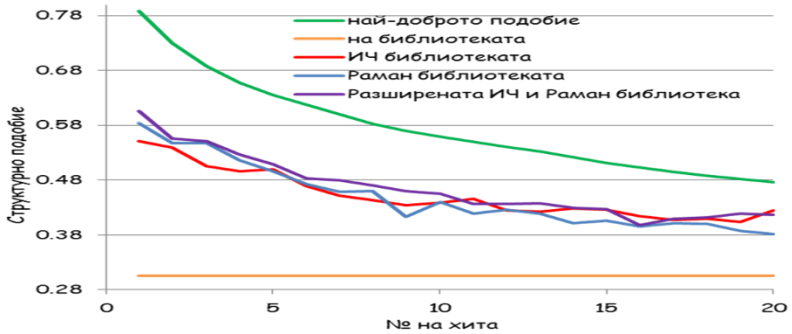
Проведено е изследване с цел разкриване на връзката между структурното и спектралното подобие в библиотеки от ИЧ и Раман спектри, а така също и в т.нар „разширена/комбинирана“ (augmented) библиотека.

Допълнително са саздадени три нови спектрални библиотеки – IRRa, RaIR и IRRaAu, всяка от които съдържа по 185 спектъра на идентични органични съедининия. Конкретно в спектралната библиотека IRRaAu всеки „спектър” съдържа паралелна информация за съответните ИЧ и Раман спектри. Първите 801 стойности във всеки спектър от нея съдържат информация за сигналите в даден ИЧ спектър, а останалите до 1602 – на Раман спектрите на съответните им съединения. Това може да се опише формално със следните уравнения:

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

Очевидно в новата библиотека, IRRaAu, „спектрите“ са предствени в размерност по абсцисата, различна за останалите две, а това от своя страна, не се поддържа от програмата IRSS и затова всички изчисления по тази точка с нея са извършени с програмата SciLab **[7]**.

На Фигура 16 са представени графичните зависимости на средно структурното подобие от номера на хита. Вижда се, че структурното подобие, усреднено за всички двойки структури 17020 (=185x184/2) е 0.305, което е близко до това за Раман спектралната библиотека (RAR – 0.303, а за ИЧ спектралната библиотека IR13484 то е 0.389).



Фигура 16. Сравнение на структурното подобие в трите библиотеки IRRa, RaIR и IRRaAu.

От кривите за трите библиотеки, представени на Фигура 16, не може да се прецени за коя от тях търсенето по подобие дава най-добри резултати, предвид и факта, че не може да се направи статистическо разграничение помежду им. Въпреки това, за така получените резултати, при предварително избраните условия, например до към четвърти хит Раман спектралната библиотека дава леко по-добри резултати от ИЧ библиотеката, но за следващите хитове търсенето в ИЧ библиотеката е малко по-добро от това в Раман библиотеката. Това леко „измества“ твърдението на Hippe за по-високата информативност на ИЧ в сравнение с Раман спектрите **[29]**.

Като следствие от графиката на Фигура 16 може да се заключи, че за почти всички хитове средното структурно подобие при БТ в съвместната библиотека е малко по-добро от това за ИЧ и

Раман библиотеките. Но ние считаме отново, че за по-сигурни изводи трябва да се използват по-големи спектрални библиотеки, за да се провери преносимостта на използвания статистически алгоритъм.

В **Таблица 3** са дадени коефициентите на корелация (Pearson correlation coefficients) на всички 17 020 двойки, показващи отношението между спектралното и структурното подобие в съответните библиотеки.

Спектрална библиотека	Коефициент на корелация
IRRa	0.513
RaIR	0.479
IRRaAu	0.564
IRRa / RaIR	0.573[1]

Таблица 3. Коефициент на корелация между структурното подобие на всички двойки структури и спектралното подобие на всички двойки спектри в съответната спектрална библиотека.

Несъмнена е статистическата значимост на коефициентите на корелация, представени в **Таблица 3.** Тяхната числена стойност показва доколко добре съответните спектри описват добре структурата на съединенията. Очевидно е, че най-добре отразяват структурното подобие спектрите на комбинираната библиотека, следвана от ИЧ и накрая от 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, което се очаква защото двата вида спектри отразяват различни аспекти от структурата на съединенията **[21]**. Нещо повече, няма точна линейна зависимост между тях и подробното разглеждане на подредбата на двата списъка на спектрално подобие показа, че те не са подредени по стойност по един и същ начин.

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

В настоящата част предлагаме една интересна алтернатива на комбинираното търсене, което е предложено в софтуера KnowItAll на фирмата за софтуер и спектрални библиотеки Sadtler **[27], [28]**. Създадената нова спектрална библиотека съдържа усреднените спектри на двойките ИЧ и Раман спектри в библиотеките IRRa и RaIR и е наречена от нас IRRaAv. В нея са включени характерните спектрални признаци за двата типа спектри, представени при едни и същи вълнови интервали, а по ординатата стойностите има са нормирани (0 – 1).

Възможни са и други аритметични комбинации по ординатите на двата вида спектри. Всички те са дадени с Уравнение 6-Уравнение 9.

усредняване: $A_k^{new} = (A_k^{IR} + A_k^{Ra})/2$ **Уравнение 6**

произведение: $A_k^{new} = (A_k^{IR} \times A_k^{Ra})$ **Уравнение 7**

изваждане: $A_k^{new} = (A_k^{IR} - A_k^{Ra})$ **Уравнение 8**

изваждане и събиране: $A_k^{new} = (A_k^{IR} + A_k^{Ra}) \times (A_k^{IR} - A_k^{Ra})$ **Уравнение 9**

По тези математически преобразувания бяха създадени четири библиотеки от комбинирани спектри: IRRaAv, IRRaPr, IRRaSu и IRRaSA. Размерът на тези библиотеки е малък, за да може се провери тяхната ефективност при търсене за идентификация, но е оценено търсенето по подобие, подобно на схемата, която е използвана за съставяне на данните в Таблица 3. Търсенето в спектрална библиотека е проведено с коефициент на корелация на спектрални криви (виж **[15]**). Четирите коефициента на корелация, заедно с два от Таблица 3 (за по-удобно сравнение) са представени в Таблица 4.

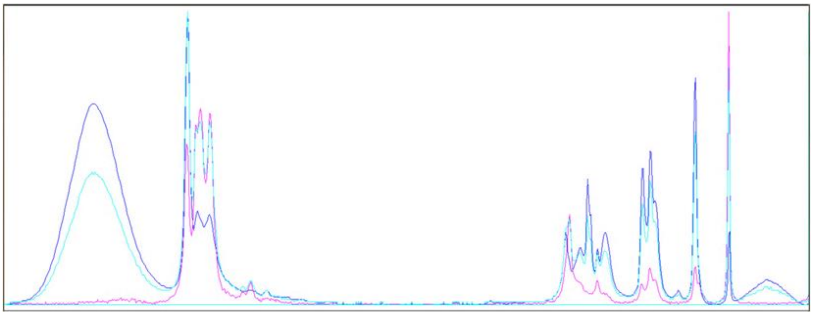
[1] *Това е коефициентът на корелация между спектралното подобие на всички двойки спектри в ИЧ библиотеката и всички двойки спектри в Раман библиотеката.*

Спектрална библиотека	Коефициент на корелация
IRRa	0.513
RaIR	0.479
IRRaAv	0.538
IRRaPr	0.444
IRRaSu	0.429
IRRaSA	0.413

Таблица 4. Коефициент на корелация между структурното подобие на всички двойки структури и спектралното подобие на всички двойки

Както се вижда от Таблица 4, спектро-структурната корелация и най-висока за „спектрите“ в усреднената библиотека, дори по отношение на оригиналните в ИЧ и Раман библиотеки. Произведението на спектри дава ниски стойности, което може да се дължи на присъствие на съединения с център на симетрия и 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.

ИЧ и Раман спектрите на няколко съединения, присъстващи и в двете библиотеки са много подходящ пример за изясняване на преобразуванията по Уравнение 6, макар и потърсването им в съответната библиотека да извежда в хит-списъка най-подобните в селектирана БД. На Фигура 17 са дадени ИЧ и Раман спектрите на пропан-2-ол, както и усреднения „спектър“. Вижда се, че ивицата за трептението ν(O-H) липсва в Раман спектъра, но благодарение на усредняването с ИЧ спектъра тя се появява в средния „спектър“. Опитът ни показва, че Раман спектърът на един наситен алкохол прилича много на Раман спектъра на съответния алкан. Други ивици, които не се появяват в Раман спектрите или са с много нисък интензитет, са тези на валентните трептения на амино-групата и карбонилната група.



Фигура 17. ИЧ и Раман спектрите на пропан-2-ол и техния среден спектър[2].

Първите пет хита от резултатите при търсене по коефициент на корелация на съответните спектри на пропан-2-ол в библиотека са дадени в Таблица 5. Вижда се ясно от химичните имена на хитовете, че търсенето на ИЧ спектър на пропан-2-ол дава първи пет хита, много подобни на неговата структура.

[2] *На Фигура 17 лилавият спектър е на от ИЧ библиотека, розовият – от Раман и синият от усреднената библиотека. По ординатата на спектъра се очита интензитетът на спектрите, а по абсциса – вълново число.*

Хит #	Спектър в IRRA	Спектър в RAIR	Спектър в IRRAAV
1	2-Бутанол	2-Бутанол	2-Бутанол
2	1-Пропанол	2-Метил-3-хептанон	1-Пропанол
3	Етанол	Бутанал	Етанол
4	n-Бутанол	4-Хептанон	n-Бутанол
5	Цитронелол	4-Метилпентан-2-он	Цитронелол

Таблица 5. Петте първи хита при търсене на спектрите на пропан-2-ол.

Изчисленията с индекса на Танимото потвърждават това наблюдение (Таблица 6) – в са дадени стойностите на спектралното подобие, HQI_4, и индексът на Танимото за хитовете, както и коефициентите на корелация за всеки един вид спектри.

	ИЧ спектър		Раман спектър		Среден спектър	
Хит #	HQI_4	Tan	HQI_4	Tan	HQI_4	Tan
1	954	0.842	961	0.842	948	0.842
2	918	0.79	941	0.333	916	0.79
3	913	0.723	934	0.423	911	0.723
4	900	0.682	932	0.366	907	0.682
5	897	0.518	932	0.423	906	0.518
средно	916.4	0.711	940	0.4774	917.6	0.711
R	0.817		0.897		0.730	

Таблица 6. Сравняване на спектралното и структурното подибие между хитовете в хит-списъка, при търсене в различни библиотеки спектри на пропан-2-ол[3].

За хитовете, получени с ИЧ и усреднените спектри, се наблюдава еднакви по средно структурно подобие за първите пет хита в списъка коефициент на корелация между спектралното и структурното подобие. Въпреки по-лошото средно структурно подобие за Раман спектрите и високия коефициент на корелация, в сравнение със средните „спектри“, то не е подредено в намаляващ ред с нарастване на номерата на хитовете .

Тази реализация на изкуствена спектрална библиотека има един недостатък, че потребителят не може да запише ръчно пиковете от ИЧ и Раман спектрите в прозореца за търсене по пикове, защото данните които има при две отделни разпечатки на ИЧ и Раман спектрите не съответстват на интензитета на пиковете на усреднения спектър. Но ако се заредят в буферите на програмата двата вида спектри и се изчисли средния спектър, то може с процедурата за намиране на пикове да се състави пикова таблица на средния „спектър“, която да бъде потърсена в библиотеката IRRaAv. Търсенето по пикове в тази библиотека, подобно на търсенето за идентификация с нея, не е изследвано от нас.

От Таблица 3 и Таблица 4 се вижда, че разширените ИЧ - Раман спектри показват по-добра корелация със структурите на съединенията, отколкото усреднените спектри: коефициентите на корелация са съответно 0.564 и 0.538. Това показва, че за други хемометрични приложения използването на разширени спектри е по-добра алтернатива, отколкото усредняването: и действително, повечето изследователи използват този подход при компютърна интерпретация на мултиспектрална информация – вижте **[30]**, **[31]**, **[32]**.

В заключение ще отбележим, че търсенето по подобие с пикове на ИЧ спектри дава най-добри резултати за $\Delta\nu$ в интервала 15-30 cm^{-1}, които са съизмерими с полуширината на повечето характеристични интервали на химичните групи **[33]**, **[34], [35]**. Докато резултатите за търсене за идентификация показват, че оптималните стойности на $\Delta\nu$ са в интервала 3-7 cm^{-1}.

По аналогия оптималните стойности, получени при БТ с цел идентификация и по подобие в Раман библиотеката са, както следва: 4-15 cm^{-1} и съответно 30-55 cm^{-1}.

[3] *Означения: спектрално подобие (HQI_4), структурното подобие (Tan) и коефициент на корелация R за първите пет хита при търсене в библиотека спектри.*

Оптималните стойности на толерансите по вълново число и абсорбция, определени за двете библиотеки – ИЧ и Раман – при двата типа БТ са доста различни, като при търсенето по подобие търсенето по подобие по-широки в сравнение с определенията при идентификацията.

4.6. Идентификация на компонентите в Раман и ИЧ спектрални смеси с използване на многопроменлива линейна регресия и на Алгоритъм за изваждане на спектри.

4.6.1. Методика на изследването

Възприемането на ИЧ и Раман техниките като методи за получаване на информация, които удовлетворяват спектроскопските закони, дава възможности за качествена идентификация на чисти вещества, както и тяхното количествено определяне при дадени условия **[36]**, **[37]**. Истинско предизвикателство пред спектроскопистите е разкодирането на получаваната информация и буквално търсенето на „пик зад пика“, което е особено осезаемо при работата със смеси. Многокомпонентният анализ е широко приложим при едновременното определяне на компонентите на дадена проба, а разнообразието от хемометрични алгоритми, съществуващи в практиката, се базират на възприемането на спектъра на дадена смес, като алгебрична сума от чистите спектри на компонентите, присъстващи в нея. Едно от изискванията за прилагането му е броят на измерванията да е по-голям или равен от броя неизвестни съставки в сместа (K≥N). За представянето но сумарният спектър ($M_{1,K}$) като линейна комбинация от спектрите на компонентите, влизащи в нейния състав – $S_{N,K}$, може да бъде изведено следното матрично уравнение (Уравнение 10):

$$M_{1,K} = C_{1,N}.S_{N,K},$$

Уравнение 10

В него с $C_{1,N}$ са обозначени концентрациите на N-та компонента в проба 1 и могат да бъдат намерени, където рангът на матрицата $S_{N,K}$ е равен на N. Индексите N и K дават информация за размерността на съответните матрици.

Придържайки се към постояни и неизменни условията на измерването на спектрите, за да бъдат намерени коефициентите в Уравнение 10, трябва да бъдат извършени допълнителни математически преобразувания, съобразени с правилата за пресмятане при работата с матрици. Умножава се горното матрично уравнение отдясно с обобщената обратна матрица на $S_{N,K} - S_{K,N}^{T}\left(S_{N,K}S_{K,N}^{T}\right)^{-1}$ и се получава (Уравнение 11):

$$M_{1,K}s_{K,N}^{T}\left(S_{N,K}S_{K,N}^{T}\right)^{-1} = C_{1,N}\left(S_{N,K}S_{K,N}^{T}\right)\left(S_{N,K}S_{K,N}^{T}\right)^{-1} = C_{1,N}E_{N,N} = C_{1,N},$$

Уравнение 11

където $E_{N,N}$ е единичната матрица и следва, че (Уравнение 12):

$$C_{1,N} = M_{1,K}S_{K,N}^{T}\left(S_{N,K}S_{K,N}^{T}\right)^{-1}$$

Уравнение 12

Тъй като няма ограничения по отношение на максималния брой наблюдения в многопроменливата линейна регресия, решение в този вариант е валидно при споменатото по-горе условие (K≥N), известно още като „преопределена“ система **[11]**. Тази процедура има същият ефект върху прецизността на резултатите, какъвто има увеличаването на броя измервания.

Вариацията в концентрационните нива се определя по Уравнение 13, което дава информация за увеличението на експерименталната грешка в резултата, а то, от своя страна, представлява доказателство за зависимостта на избора на броя спектри при многокомпонентния анализ за вибрационни спектри ($S_{N,K}$).

$$V(c) = s_e^2(K'K)^{-1}$$

Уравнение 13

$$V(c_1) = C11 * s_e^2$$
……
$$V(c_n) = C1n * s_e^2$$

Уравнение 14

където коефициентите C11 –C1n (Уравнение 14) са диагоналните елементи от матрицата $\left(S_{N,K}S_{K,N}^T\right)^{-1}$. Съответно доверителните интервали за псевдоконцентрациите се изчислява по зависимостта от Уравнение 15:

$$c_n \pm t_{k-n}^{1-\alpha}\sqrt{v(c_n)}$$

Уравнение 15

където n е броят измервания (спектрални признаци), k – брой хитове в хит-списъка от библиотечното търсене, а α може да заема стоиностите – 0.01 или 0.05, в зависимост от статистическата вероятност, с която се работи при едно измерване (99% или 95%).

Целият обем от данни при обработката на спектрална информация за бинерните смеси, с които работим, е получен в резултат на работата с програмата IRSS, която позволява реализирането както на регресионен метод с последователно включване на хитове и нормален регресионен метод, както и алгоритъм за изваждане на спектри.

Използваният от нас регресионен метод за многокомпонентен анализ на смеси от ИЧ и Раман спектри, се основава на сравнението на всеки един спектър от хитсписъкът последователно със спектъра на сместа за установяване на идентични спектрални признаци. Работили сме с обратен алгоритъм за търсене на бинерната смес в библиотека от спектри. Изчисляването на вектора $C_{1,N}$ (безразмерен) всъщност не може да се възприема като оценка на количествения състав на присъстващите компоненти в сместа, главно поради две причини: библиотечните спектри са заснети с различно и непретеглено количество вещество за твърдите проби, както и с различна и неизмерена дебелина на слоя за течните вещества. Освен това библиотечните спектри и този на непознатото съединение (или сместа) са нормирани в интервала 0 - 1 a.u. Работим с безразмерни величини, т.е. техните стойности нямат физически смисъл, поради което са наречени от нас псевдоконцентрации и имат смисъл на регресионни коефициенти. Участват в описание на линейната зависимост в многопроменливата линейна регресия. От своя страна, статистически отличимите от нула стойности за $C_{1,N}$ са използвани като критерии за възприемане на коефициентите като приемливи резултати. В използваният от нас софтуер IRSS се работи с 95% и 99% доверителен интервал на числените стойности на „псевдо-концентрациите“, както е описано Massart et al., **[25]** Уравнение 15.

Описаната от Somberg процедура за изваждане на спектри включва етапи на БТ, изваждане на спектър, където едно от основните изисквания е всички компоненти на сместа, да присъстват в използваната БД **[38]**.

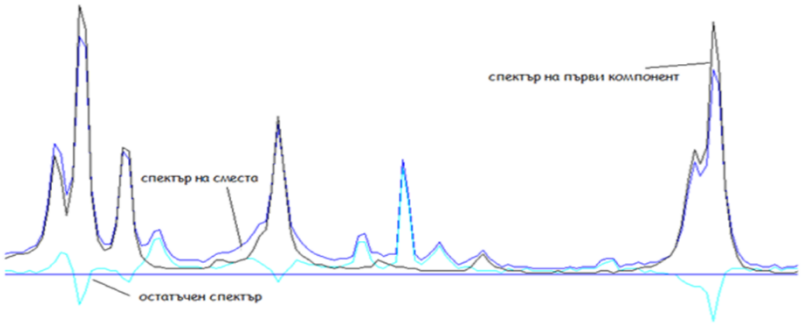
Използваната от нас процедура за изваждане на спектри следва непосредствено получаването на информация за регресионните коефициенти на спектрите на библиотечни съединения в хит-списъка, където първият от тези хитове със статистически различима от нула стойности се възприема като един от най-вероятните компоненти на сместа. Последният може да е сред първите хитове в генерирания от програмата хитсписък, но това не е задължително условие. Използвайки Уравнение 16, където $S_{F,K}$ е f-тия ред в матрицата $S_{N,K}$, f е реално число, така че ‘f*$S_{F,K}$’ премахва спектралните ивици на този компонент от спектъра на сместа, а като резултат се получава остатъчен спектър – $R_{1,K}$. Изборът на подходящ коефициент за изваждане в Уравнение 16, от една страна, може да се съобрази предложената от Somberg **[38]** идея и да бъде възприет за 1.0 по числена стойност, където, като довод се изтъква, че когато компонентите в сместа са в различни обемни части и въпреки ефектът от нормализация на спектрите, спектралните ивици на компонента с по-висока концентрация ще надхвъля тази на другия(те), в резултат на което и първият ще се появи

сред представителите с най-високо спектрално подобие в хит-списъка. Както и Somberg **[38]** подчертава, изборът на коректна числената стойност за коефициента има отношение към получаването на т.нар. “wings” или само съблюдаване на достигане на базовата линия Фигура 18. От друга страна, коефициентът f може да се получи в условие на експерименталната постановка, където потребителят наблюдава паралелно с процедурата на изваждането как се открояват спектърът на остатъка от сместа, „разрешавайки“ спектралните ивици на известния вече компонент да „изчезнат“ от остатъчния спектър. Подобна практика е реализирана от проф.П.Пенчев и сътр. и материалът е публикуван в ВСС през 2008 **[39]**, където преимуществено коефициентът f за използваните бинерни смеси заема стойности, близки до 1.0. Съществува и трета възможност за избор на числова стойност на коефициента на изваждане, а именно присвояването на стойността на $c_{1,F}$, получена от многопроменливата линейна регресия.

$$R_{1,K} = M_{1,K} - f\ S_{F,K}$$

Уравнение 16

Изборът за най-подходящ коефициент зависи от няколко фактора, които намират своето логично обяснение в получените резултати. От една страна, коефициентът за изваждане ще е най-ефективен, когато е избран да бъде 1.0, в случай че въпреки структурното подобие, спектралните ивици на компонентите отразяват добре селективността на метода по отношение на об.ч за характеристичните ивици. От друга – използването на регресионен коефициент или експериментално определения, за които, както може да се проследи от получените резултати, често са с близки стойности, следва да бъдат по-ниски или по-високи от 1.0.



Фигура 18. Етап от алгоритъма за изваждане на спектри с получаването на „крила“ в остатъчния спектър.

Процедурата по изваждане се съпровожда от т.нар. „окастряне“ (Фигура 18) в спектъра на негативните спектрални ивици и нормиране по ординатни стойност (0-1), след което отново се провежда БТ по метода на търсене по пикове. Алгоритъмът, по който се осъществява търсенето е същият – обратен, главно поради вероятността от все пак присъствие на ивици и на двете (или повече, ако к≥3) съставки в спектъра, а откриването на другата съставка в сместа става чрез провеждане на последваща регресия на остатъчния спектър. Първият представител със статистически различим от нула коефициент за $c_{1,S}$ се възприема като втория компонент в спектралната смес. Този цикъл от итеративни процедури може да продължи и за търсене на трета, четвърта и т.н. съставки, но опитът ни показва, че с този метод повече от три компоненти е трудно да бъдат идентифицирани.

Описан по този начин спектралният анализ на смеси, на пръв поглед може да изглежда ясно разбираем и открит, но трябва да се има предвид, че анализът дори и на смеси, чиито компоненти се открояват с напълно различими спектрални образи, могат да дадат незначими или дори грешни резултатите. Ясно е, че съществуват значими вътрешни и външни условия, които оказват влияние върху получените резултати при опитите за идентифициране на спектри, касаещи се главно до условията при които са били регистрирани спектрите. За съжаление в литературата не са дадени достатъчно ясни и аргументирани препоръки за големината на коефициента при изважане, а само се споменава за „нулиране“ по една до две спектрални ивици между смесения и идентифицирания

спектър. При достатъчно близки спектри на компонентите на сместа и в зависимост от степента на изваждане, това може да доведе до загуба на втория компонент, като например или да не се появи изобща в хит-списъка, или да не е сред първите хитове.

Може да бъдат определени няколко причини за получаването на незадоволителни резултати, в следствие на гореспоменатото:

- ✓ Остатъчният спектър е прекалено шумен и се наблюдават т.нар. крила “wings” причинени от различните ширини на ивиците в спектрите на сместа и библиотечния
- ✓ спектроскопистът обикновено избира една ивица и се стреми да я нулира при изваждане на спектрите. Няма гаранция, че избраната ивица за изваждане на се съдържа в спектъра на втория компонент на сместа или припокрива друга негова ивица. В този случай се губи съществена информация за остатъчния спектър.

Ясен критерий за прекратяване на изваждането не се използва. Колкото по-малък е коефициента на уравнението, толкова по-голяма е вероятността повече спектрални ивици от първия спектър да останат в остатъчния и обратно – колкото по-голям е, толкова по-малко ивици на втория компонент се запазват. Присъствието на крила пречи на търсенето, но е неизбежно.

Нашият опит показва, че за целите на процедурата по анализ на спектрални смеси не би следвало само да се изпълняват стъпките на алгоритъма и да се следят оптимални стойности, получени за параметрите, но така също и да бъде направена добра оценка по един добре конструиран креативен подход за разрешаването на проблема. Без съмнение т.нар. “educated user” би реализирал многофакторно БТ при използването на различни предписания за реализирането му. Естественият ход на анализа, който би се следвал от него, е: обследване на различни спектрални интервали, приритетно място сред които е за „фингърпринт“ региона; отличаване на спектралните ивици, които могат да бъдат отнесени за вероятните компоненти на сместа; вариация на двата спектрални интервала (по вълново число и абсорбция), а така също и „threshold“, който се определя при извеждането на пиковата таблица (по подразбиране различен за различните смеси, заради разликите във фоновото поглъщане). Съществено място заемат и влияят върху резултатите алгоритъма за спектрално подобие и броят на използваните хитове в списъка.

За преодоляването на подобни проблеми са формулирани три евристики от проф.Пенчев и сътр. **[39]**, някои от които намират своето логично обяснение в получените от нас резултати.

С помощта на този тривиален метод може да се използва регресионният анализ, на първо място за намиране на първия компонент, ако той не се появи при първи хит в търсенето на спектъра на сместа, и на второ – за идентифициране на втория компонент, ако и той не се появи като първи хит в търсенето на остатъчния спектър.

4.6.2. Резултати от идентификация на ИЧ и Раман спектрални смеси:

Проведени са тестове с ИЧ и Раман (42 ИЧ и 61 Раман) спектри на смеси за различни класове органични съединения. Получената за тях спектрална информация е резултат от предварително реализирани пробоподготовки (в кювети за Раман и капилярен слой с KBr за ИЧ) и регистриране на спектрите им от проф.П.Пенчев, С.Цонева и В.Митева. Смесите са приготвени в няколко различни, предварително избрани обемни съотношения на компонентите и са потърсени в съответните библиотеки RAR от 330 спектъра и IR01 - IR06 от общо 911 спектъра. Целта е проверка на алгоритъма за успешна идентификация на един и/или двата компонента от бинерната смес с използването на метода на многопроменливата линейна регресия и алгоритъм за изваждане на спектри.

Съгласно предварително изложени препоръки **[39]**, представените от нас резултати за идентифициране на Раман и ИЧ спектрални смеси са съобразени със следните условия:

- ✓ прагът („threshold“) в пиковите таблици на библиотечните спектри е избран да бъде 0.03, а използваният в спектралните смеси е препоръчително – 0.01 или съобразен с нивото на шума. Имат се предвид главно ниските концентрационни нива на някои компоненти, а от друга страна, наличието в някои ИЧ спектри на ротационни ивици на водните пари изисква по-висока

стойност, за да бъдат редуцирани от пиковата таблица. По отношение на Раман спектрите, поради повсеместното високо ниво на шума, се избира праг в интервала 0.02-0.04.

✓ по отношение на търсенето на остатъчният спектър и поради ниското ниво (S/N) при него, изборът попада върху широк интервал от стойности за праг (0.02-0.09). Раман спектрите по подразбиране са с по-лошо отношение S/N, следователно е абсолютно нормално при тях да се избира по-висок праг в сравнение с ИЧ. В зависимост от коефициента на изваждане, много често изборът на праг граничи със значително по-висока стойност от посочените. Към последното имат отношение и обемните части на сместа, с които участват компонентите и съответно степента на припокриване на ивиците в характеристичните интервали.

Опитът показва, че оптималният толеранс по вълново число за търсене на спектрални смеси, $\Delta\nu$, е по-висок в сравнение с използвания за търсене на чисти вещества. Избраните от нас толеранси по вълново число, съответно абсорбция/интензитет са: $\Delta\nu$=12 cm^{-1} и $\Delta A/\Delta I$=1.0 a.u./a.i.u. Последното е определено при максимална стойност, което означава, че интензитетът няма отношение при определяне на съвпадението по пикове.

За определянето на регресионните коефициенти само първите 40 хита в хит-списъка са използвани, въпреки че потребителят може да определи по негова преценка друга желана стойност.

Спектралният интервал, с който се работи при изчисляване на регресията е 1300-600 cm^{-1} (във fingerprint региона), но въпреки това при липсата на идентифициран компонент са използвани и по-широки интервали в следния ред – 1800-600 cm^{-1}, 1800-500 cm^{-1} и 3700-500 cm^{-1}, където отново е спазено условието изборът на съединение, което е най-напред в хит-списъка и със коефициент статистически различим от нула.

Използван е алгоритъмът за обратно търсене по пикова таблица при библиотечното търсене за разкриване на най-подобните референтни спектри на неизвестния. При регресията, обаче, анализът използва цялата спектрална крива в избрания спектрален интервал. Алгоритъмът за обратно търсене с използването на пикови таблици е специално създаден за работа със смеси, но в случай че остатъчният спектър е със много лошо съотношение S/N (както при Раман спектрите), то потребителят има право да избере и друг алгоритъм за търсене **[40]** по цяла спектрална крива.

- **Основните стъпки, които се следват при цялостната методика, са следните:**
  1) Зареждане на спектъра на сместа и създаване на пикова таблица при избора на подходящ праг (t_U)
  2) Провеждане на БТ по пикове (ивици) с използването на обратен алгоритъм при толеранси по вълново число и сигнал съответно – $\Delta\nu = 12\ cm^{-1}$ и ΔI = 1.0 a.u. (a.i.u.) в съответната спектрална библиотека и получаване на хитсписък от съединения, които показват най-високо спектрално подобие със спектъра на сместа.
  3) Провеждане на регресионен анализ (Таблица 7) със съединенията от хит-списъка по спектрална крива в предварително избран спектрален интервал (1300-600cm^{-1}), където е възможна и корекция, в зависимост от степента на удовлетвореността от получените резултати. Определяне на първи компонент в дадената смес.

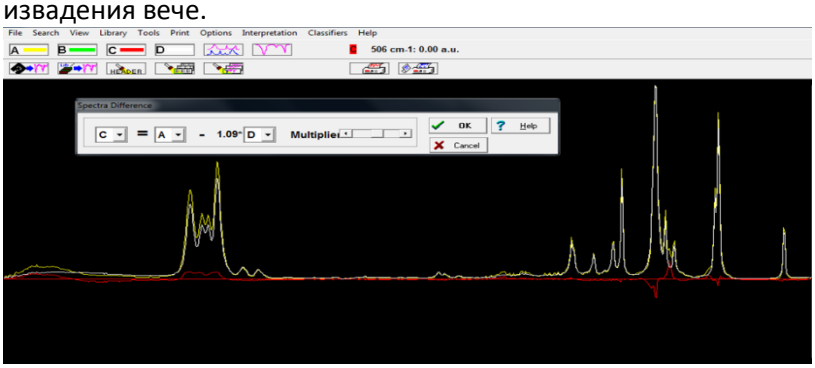
mixi14
Number of Spectra: 40
Statistical Significance: 95%

Hit's identification	mean	+- DC	
1,4-Dioxane	1.044	0.041	
Dodecane	-1.067	1.973	***
1,1,1-Trichloroethane	0.014	0.045	***
Pyrrolidine	0.016	0.180	***
Decyl aldehyde	-0.532	0.571	***
Trioctylamine	0.073	0.773	***
2-Oxo-cyclohexanecarboxylic acid methyl ester	-0.024	0.052	***
3-Decanone	0.156	0.415	***
2,3-Dimethyl-1-hexen	0.079	0.536	***
3-Eicosanone	-0.386	0.685	***

Таблица 7. Резултат от регресионния анализ на смес от 1,4-Диоксан и Тетрахидрофуран

4) Изваждане на спектъра на идентифицирания като първи компонент от спектъра на сместа с подходящ коефициент, съобразно използваните три подхода (Фигура 19)

При изваждането на идентифицирания като първи компонент в сместа получаваме като резултат т.нар. „остатъчен спектър", за който се създава нова пикова таблица с праг по-висок, заради очевидно по-високото ниво на шум в него. Последното е като резултат от вероятността за присъствие на ивици както на спектъра на другия компонент, така и на остатъчни ивици от извадения вече.



Фигура 19. Изваждане на спектри.

5) Създаване на пикова таблица на остатъчния спектър с подходящ праг (има шум)
6) Провежда се регресионен анализ и с остатъчния спектър, за да се провери вероятността той да бъде идентифициран.

При търсене на остатъчния спектър в библиотеката при зададени за експеримента условия не винаги се идентифицира чрез регресията втория компонент на сместа, затова в тези ситуации се пристъпва към разширяване на спектралния интервал. За да се избегне ефектът на различно „смесване" на псевдо-концентрациите на двата спектъра в изчисленията, което се наблюдава при наличието на повече от един спектър за едно съединение и пречи на търсенето, се изтрива хита на съединението с по-малък HQI в хит-списъка, преди да се извърши регресионният анализ.

Проведени са три типа изследвания за разкриване индивидуалните компоненти на смесите, съобразно подходите при избора на коефициент f за изваждане: 1.0, от регресията и коефициентът от експертното изваждане на спектрите от оператора. Проверката и на трите процедури носи своето разумно научно-приложно потвърждение и обяснение на получените резултатите. Експериментална проверка на част от резултатите е проведена при съблюдаване на условия, при които независим оператор (С.Начкова) провежда анализите, без да има информация за компонентите на смесите, т.е. т.нар. "blind experiment". В първата серия от изчисления, изхождайки от първите три евристики, операторът цели коректното изваждане на спектрите на генерираните чрез регресията съединения от този на сместа и определя неговата стойност, съобразявайки главно изследваните спектрални характеристики. При третата серия, за провеждането на анализите се взема предвид и се нагласява фактора за изваждане да бъде 1.0, а при втората серия – f се възприема от таблицата с резултати получена от многопроменливата регресия. Разбира се, единствено първата серия експерименти би следвало да се отнесе към групата на т.нар. "blind experiment", тъй като останалите две не покриват това условие напълно, защото изборът на фактор за изваждане се възприема или като константа, или получен от проведената регресия.

Общият брой смеси, с които бяха проведени изследванията, възлизат на 103 – 61, от които на Раман и съответно 42 на ИЧ спектри , – като съществува сечение между двата метода от общо 40 смеси. Всички смеси са пробоподготвени в различни обемни части и някои от тях в различни времеви периоди, което позволява да се проследи възпроизводимостта на получаваните резултати резултатите. За някои може да се сравнят резултатите от проведените изследвания, от една страна, по техника на заснемане и получаване на спектралната информация, и от друга – по използваните коефициенти за изваждане по Уравнение 16.

За оценяване на получените резултати от анализа на всички проведени анализи за двата метода (ИЧ и Раман) в Таблица 8 са представени 20 от използваните от нас смеси и коефициент на изваждане, получен от регресията. Някои от тях са доста сложни при разкриването на компонентите им с т.нар. blind experiment. Такива са например ИЧ или Раман спектрите на смесите: хексан&циклохексан, 1-деканол&1-нонанол, 1-деканол&1-октанол, 2-метил-1-фенипропан-1-он&4'-метилпропиофенон, Бензилацетон& Бутирофенон, 2-Етилхексан-1-ол&2-етилхексан-1,3-диол, Циклопентанон&Бензилацетон.

Първата (blind) и втората серия (регрес.коеф) анализи в повечето случаи дават сравними резултати и вдействителност по-добри от тези във третата серия (коеф. 1.0). Откроява се значимостта на спектроскопската преценка при избора на коректни фактори за изваждане и по-ниската им числена стойност от съответно определените от регресията, от където и присъствието на остатъчни ивици в спектъра след поцесът на изваждане на първия компонент. Въпреки това, е установена коректна идентификация в немалко случаи и на втория компонент в сместа, а използвайки обратен алгоритъм за търсене отчита именно този факт, поради което и често се появява във втория хитсписък и вече извадения спектър.

От друга страна, неуспехите при третата серия анализи показват, че когато двата компонента имат общи спектрални ивици с висок интензитет, поради тяхната адитивност, последващото нормиране спектъра **[41]** на сместа ги свежда до линейна комбинация от спектрите на компонентите с коефициент значително различим от 1.0.

Би следвало да се очаква по-слаба чувствителност на алгоритъма за идентифициране на компонентите с по-ниска концентрация, които, както ще се види и от резултатите, понякога липсват в списъка. Това, обаче, не е недостатък на регресионния метод, а на алгоритъма за търсене на спектъра на сместа **[19]**. Трябва да се отбележи, че по подразбиране по-слабата чувствителност на Раман метода в сравнение с ИЧ спектроскопия и особено в случаите на несиметрично заместени молекули, интензитетът на спектралните ивици се влияе и предварителното оптимизиране на условията на заснемане, непосредствено преди извършването на серията от измервания, а така също и изборът на мощността на лазера, с който се облъчва пробата, което се дава като допълнителна характеристика на спектъра и задължително ще се отразява при сравняване на идентични спектри с различни mW.

Анализът на резултатите, получени при работа с ИЧ спектрални смеси, показва, че в някои случаи, подобно на Somberg **[38]** коректното идентифициране на компонентите на сместа е факт, но в редица случаи се натъкваме или на грешен резултат, или на липсата на такъв. В някои от случаите това зависи от нивото на структурно подобие между компонентите в сместа, породило и високото спектрално подобие, но в други случаи, независимо от тези очаквания за спектро-структурни корелации, причините могат да бъдат резултат от допълнителни фактори, като например: припокриване на спектрални ивици, вибрационно взаимодействие, електронни ефекти в молекулите, образуване на асоциати и влошавани качеството на спектрите, поради загуба на селективност на ивиците и т.н **[33]**. Въпреки тези и други факти относно вероятностите за грешки или пречения при интерпретирането на спектралната информация, несъмнена е силата на регресионния анализ с тази си цел.

Таблица 8. Идентификация на компоненти на смеси на ИЧ и Раман спектри. В първата колона са дадени концентрациите на компонентите в смесите (обемни части), а регресионните коефициенти, отбелязани като f_1 и f_2; (error) – когато друг компонент е открит като компонент на сместа [4].

a. 2`-метилацетофенон (A) и 3`-метилацетофенон (B)

A:B	ИЧ				Раман			
v/v	*1ви компонент*	*2ри компонент*	f_1	f_2	*1ви компонент*	*2ри компонент*	f_1	f_2
1:1	B	A	0.50	1.07	A	B	0.50	0.94
1:4	B	A	0.81	0.77	B	A	0.91	0.81
1:9	B	A	0.88	0.48	B	A	0.85	0.35
4:1	A	B	0.71	0.66	A	B	0.78	0.86
9:1	A	B	0.83	0.77	A	B	0.84	0.47

b. 1,4-диоксан (A) и тетрахидрофуран (B)

A:B	ИЧ				Раман			
v/v	*1ви компонент*	*2ри компонент*	f_1	f_2	*1ви компонент*	*2ри компонент*	f_1	f_2
1:1	A	B	1.05	0.73	A	B	0.65	0.97
1:4	A	B	0.92	1.02	B	A	0.77	0.91
1:9	B	A	1.00	0.84	B	A	0.40	0.58
4:1	A	B[1)5]	1.08	0.74	A	B	0.85	0.37
9:1	A	–[2)6]	0.99	–	A	B	0.90	0.46

c. 3-хептанон (A) и 4-хептанон (B)

A:B	ИЧ				Раман			
v/v	*1ви компонент*	*2ри компонент*	f_1	f_2	*1ви компонент*	*2ри компонент*	f_1	f_2
1:1	B	A	0.35	0.40	B	A	0.38	0.93
1:4	B	A	0.59	0.31	B	A	0.77	1.44
1:9	B	A	0.71	0.38	B	A[1)]	0.98	4.04
4:1	A	B	0.49	0.36	A	B[1)]	0.95	1.67
9:1	A	error [7]	0.63	1.61	A	B[1)]	0.92	1.30

d. 1-нонанол (A) и 5-нонанол (B)

A:B	ИЧ				Раман			
v/v	*1ви компонент*	*2ри компонент*	f_1	f_2	*1ви компонент*	*2ри компонент*	f_1	f_2
1:1	A	B	0.16	0.36	–[1)]	–	–	–
1:4	B	error[1)]	0.47	0.74	B	error[1)]	0.43	–
1:9	B	error[1)]	0.57	0.72	B[1)]	error[1)]	0.44	1.12
4:1	A	B	0.23	0.09	B[1)]	–	0.15	–
9:1	A	error[1)]	0.22	0.60	error[1)]	error[1)]	0.62	0.48

Както може да се проследи в Таблица 8, дори за значително структурно подобни компоненти като 3-хептанон и 4-хептанон са идентифицирани успешно при анализирането им в смес. Както се и очаква, значими проблеми се появяват при някои от смесите в об.ч. 1:9 и 9:1. От друга страна, по-лоши са резултатите за смесите от 1-нонанол и 5-нонанол, което може да се обясни с присъствието на спектър на 1-деканол в съответната библиотека, поради незначимо малката

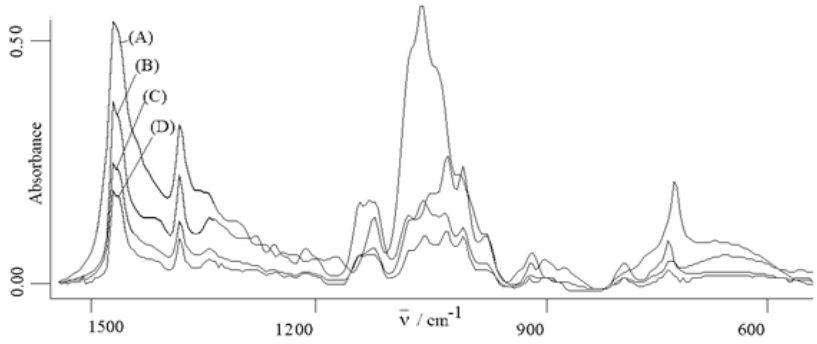
[4] Означения: A:B (v/v) – обемни части на компонент A и компонент B; 1-ви и 2-ри намерен – намиране на компонента преди, съответно след изваждането; f_1 – коефициент на изважнаде; f_2 – регресионен коефициент след изваждането.

[5] 1) – *оригиналните спектрални интервали за регресионните изчисления са разширени*

[6] 2) – *компонентите не са разкрити, т.е. регресионните коефициенти не састатистически разграничими*

[7] error – идентифициран е компонент, неприсъстващ в сместа

разлика между спектрите на 1-нонанол и 1-деканол. Появяват се и някои други първични алкохоли в хит-списъка, чиито спектри са много подобни на 1-деканол.



Фигура 20. Спектри на сместа (1:1 v/v) (A) от 1-нонанол и 5-нонанол, остатъчен спектър (B), 1-нонанол (C) и 5-нонанол (D)

Едни от най-значимите характеристични спектрални ивици, присъстващи в различни класове съединения, са тези на O-H и C=O валентно трептене. Отличима разлика между ИЧ и Раман спектри на съединения е именно присъствието на ивица на поглъщане (разсейване) на ЕМЛ, отговаряща за присъствие на O-H функционална група в структурата на определяемото вещество. Много отличима е липсата на тази ивица в Раман спектрите на наситени алкохоли, поради което те значително си приличат по спектрални характеристики с олефинови или наситени алкани със същата алифатна част. Ивицата на C=O валентното трептене в Раман спектрите е значително по-ниска по интензитет. Така тези две основни спектрални особености са причините за грешки, когато спектралните интервали, използвани при регресията се уширяват.

Следва да се подчертае, че конкретно за смес от 1-нонанол и 5-нонанол в 1:1 об.ч. Двете съединения, които са от клас алкохоли и помежду си представляват функционални изомери, на практика могат да се охарактеризират с високо структурно подобие. Търсенето по пикове за ИЧ смес дава като резултат 1-нонанол като 6-ти хит в спесъка от референтни спектри, а 5-нонанол – 10-ти хит. Съответнопървите пет хита са на 1-хексанол, 1-деканол (два пъти, поради повторение в две от библиотеките), додекан и октакозан. Остатъчният спектър е получен по Уравнение 16, като е използван регресионният коефициент, чиято числена стойност е f=0.16. На Фигура 20 са представени спектрите на сместа и съответните компоненти в спектрален интервал от 1500-600 cm^{-1}, а така също и на остатъчния спектър. Могат да се направят следните изводи:

- концентрационното отношение от 1:1 об.ч. не означава равни псевдоконцентрации; в спектралния интервал 1300-600 cm^{-1} спектрите на отделните съединения са доста различни;
- компонентите имат припокриващи се ивици около 1465 и 1380 cm^{-1} като следствие от присъствието на общи структурни фрагменти – -CH3 и -CH2; основната разлика между спектрите е за C-OH валентно трептене, отговарящо за ивици на присъствия на първичен, съответно вторичен алкохол, а така също и ρ(CH2), 724 и 732 cm^{-1};
- спектралните ивици на 1-нонанола ν(C-OH) = 1058 cm^{-1} и ρ(CH2) = 724 cm^{-1}, изчезват в значителна степен следизваждането и не присъстват в остатъчния спектър.

Едни от основните недостатъци на спектралната идентификация на съставките на смеси се дължи на две основни причини – изпускане на спектъра на компонента с най-ниска концентрация и подобни спектри в хит-списъка, трудноразличими сматематическа процедура от вида. Първият и досега е трудно разрешим, въпреки многостранните опити, а по отношение на втората причина – хитсписъкът се избира така, че избраните спектри да са най-подобни натърсения, което поподразбиране ги прави и подобни помежду им.

Като обобщение на получените резултати може да се каже, че ИЧ спектрите дават по-добри резултати от регресионния анализ при разпознаване на първия компонент и съответно за

разпознаването на втори компенент при работа с коефициент на изваждане, получен от регресията в сравнение със смесите на Раман спектри.

4.7. Резултати от идентификация на неизвстно съединение по ИЧ

Предизвикателствата, свързани с разкриване на възможни подструктури и/или структурни фрагменти, при интерпретиране на спектрална информация е породено от два основни фактора: висока степен на припокриване на спектралните ивици и отместване на характеристичните ивици в една или друга степен **[42]**. Припокриването на спектралните ивици е най-значително във „фингърпринт" региона, където се „отпечатва" структурният скелет, докато преместването на спектралните ивици към по-големите или по-малките честоти е в резултат от допълнителни фактори, които влияят при регистрирането на спектър **[18], [33]**. Тези обстоятелства пораждат едновременно необходимостта от реализирането на изобилие от познания за спектралните характеристики на функционалните групи.

Примерът, който ще бъде представен е за непознато съединение, представляващо продукт от проведена екстракция от доц.П.Бозов из растителен обект. С възможностите, които предлага програмата IRSS за БТ, беше успешно идентифицирано съединението, макар че получените резултати показват пълно несъответствие на първоначалните очаквания. Отнесени са спектралните ивици в ИЧ спектър за напълно непознатото съединение, потърсено при работа с няколко алгоритъма за оценка на спектралното подобие както по метода за търсене по спектрална крива, така и по пикова таблица.

Спектърът на предполагаемото извлечено съединение е заснет на апарат VERTEX 70 FT-IR (Bruker Optics) за заснемането на ИЧ, а за Раман спектъра модулът RAM II.

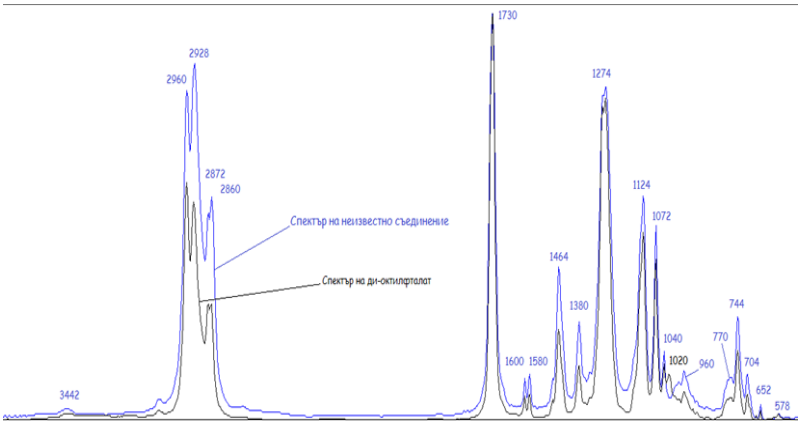
- ИЧ и Раман – работни условия:
  - ✓ Спектрален интервал – 4000-400 cm^{-1}/ 4000-50 cm^{-1}
  - ✓ Разделителна способност 2 cm^{-1} с 25 сканирания
  - ✓ За Раман спектъра е използван източник – Лазер при работна дължина на вълната в NIR (1074nm)

По идея на проф.П.Пенчев бе проведено библиотечното търсене при работа с библиотеки IR13484 и съответно RAR330. Може да се проследи, че в библиотечния спектър на съединението има една излишна ивица (1020 cm^{-1}), за която се предполага се, че е от онечистване и не пречи на идентификацията. За съжаление спектърът на търсеното съединение не присъства в Раман библиотеката, съответно не може да бъде идентифицирано с програмата IRSS. В случай бе извършено сравнение със спектър от атласа на Schrader[8].

Редица изследователи и специалисти подкрепят мнението, че предизвикателството е извеждането на причинно-следствените научни постулати не само за известни, но и за неизвестни обекти **[43]**, **[12]**, **[44]**.

Спектърът на неизвестното вещество беше потърсен в библиотека IR13484 и напълно идентифициран при висока степен на спектрално подобие с намерения библиотечен спектър на ди-октилфталата Фигура 21. Вижда се, съвпадението по вълново число и относителни интензитети на спектралните ивици при сравнението на двата спектъра, както в характеристичния, така и във фингърпринт региона. Силно интензивна е ивицата на валентното C=O трептене при 1730 cm^{-1}, а така също и около 1274 cm^{-1}, дължаща се на ν_{as} трептене на CO-O- в ди-бутилфталата. Присъствието на ивица в библиотечния спектър около 1017 cm^{-1} се дължи най-вероятно на онечистване и/или ефект на разтворителя.

[8] ИЧ/Раман Атлас на органични съединения, B.Schrader u W.Meier.



Фигура 21. Насложени ИЧ спектри на неизвестното съединение и библиотечния спектър на ди-октилфталат, потърсен по пикове.

На Фигура 21 може да се проследи и направи сръвнение до колко добре съвпадат спектрите на неизвестното съединение и на намереният като най-подобен библиотечен спектър.

За библиотечното търсене по цяла спектрална крива са представени резултати при търсене с четирите алгоритъма (т.3.4.1). С изключение на средно абсолютното отклонение, останалите алгоритми позиционират като най-подобен спектърът на ди-октилфталата от БД (Таблица 9).

Таблица 9. Резултати от идентификация по метода на търсене по спектрална крива

	Спектрална крива							
	с.к.о.	HQI	с.а.о.	HQI	с.п.	HQI	к.к.	HQI
ИЧ спектър	1	932	2	962	1	966	1	981
	ди-октилфталат		I-во – 2-оксоизобутилов естер на декановата киселина		ди-октилфталат		ди-октилфталат	

Извършената успешна идентификация на непознато съединение по негов ИЧ спектър е доказателство за потенциалните възможности на методите за БТ, за голямото им приложение за практиката и че в действителност са полезни при идентификация на спектри на непознати съединения.

5. ОБОБЩЕНИЕ НА РЕЗУЛТАТИТЕ И ПРИНОСИ:

- Създадена е библиотека от 102 ATR спектри, която е полезна и представлява ценно допълнение към наличната БД от абсорбционни спектри;
- Допълнена е библиотеката от Раман спектри с около 70 нови спектъра;
- Направената експериментална проверка на алгоритмите за търсене по ивици в ИЧ библиотека с цел идентификация на непознато съединение, показва, че най-добри резултати са получени при работата с правия алгоритъм. Оптималните толеранси по ΔA и Δν се намират в интервалите: Δν=3-7 cm^{-1} и ΔA=0.4-0.9 a.u.;
- Направената експериментална проверка на алгоритмите за търсене по ивици в Раман библиотека с цел идентификация на непознато съединение, показва, че най-добри резултати са получени при работата с правия алгоритъм. Оптималните толеранси по ΔI и Δν се намират в интервалите: Δν=4-11 cm^{-1} и ΔI=0.2-1.0 a.i.u. и Δν=4-15 cm^{-1} и ΔI=0.2-0.3 a.i.u.;
- Успешна идентификация при библиотечно търсене на ATR спектри в библиотеки от ИЧ спектри на поглъщане и обратно с мярка за спектрално подобие коефициента на корелация по метода за търсене по спектрална крива. Въпреки различията между двата вида спектри, тяхната идентификация е осъществима задача и успешна за проведените анализи;
- При изследване на зависимостта между структурно и спектрално подобие за ИЧ спектри с метода на търсене по пикове се установява, че скаларното произведение по пикове ($SpSim_2$)

дава по-висока оценка на структурното подобие от симетричната мярка ($SpSim_1$), докато за Раман спектрите е обратно.

- ИЧ спектри:
  - ✓ Всички криви от четирите фигури имат по-малки стойности на структурното подобие при крайните интервали на неопределеността по абсцисата: $\Delta\nu$ = 5-15 cm^{-1} и $\Delta\nu$ = 30-70 cm^{-1}. При малки стойности на $\Delta\nu$ отпадат подобни съединения от хит-списъка (спада чувствителността), а при големи се губи селективността на използваната мярка за спектрално подобие;
  - ✓ По-високата селективност на мярката $SpSim_2$ спрямо $SpSim_1$ се наблюдава също и при високи стойности на $\Delta\nu$. Това личи от кривите на средното структурно подобие за $SpSim_2$, които са по-полегати, отколкото кривите за мярката $SpSim_1$;
  - ✓ Сравняването на резултатите от търсене по пикове с търсене по спектрална крива показват, че във втория случай резултатите са по-добри, където най-високо структурно подобие се получава при библиотечно търсене с коефициент на корелация, макар и да не съвпада с най-доброто структурно подобие в библиотеката.
- Раман библиотека
  - ✓ Подобно за ИЧ и при Раман спектрите резултатите показват, че всички криви имат по-малки стойности на структурното подобие при крайните интервали на неопределеността по абсцисата: $\Delta\nu$ < 15 cm^{-1} и $\Delta\nu$ > 50-55 cm^{-1}. При малки стойности на $\Delta\nu$ отпадат подобни съединения от хит-списъка (спада чувствителността), а при големи се губи селективността при спектралното сравнение;
  - ✓ По-висока е оценката на структурното подобие при работата със симетричната мярка за оценяване на спектралното подобие ($SpSim_1$);
  - ✓ Раман библиотеките дават списък със съединения структурно подобни на неизвестното и могат да се използват при разкриване на структурите;
  - ✓ В сравнение с библиотечното търсене по спектрална крива и при Раман библиотеките коефициента на корелация с логаритмувани интензитети, използван като спектрална мярка за подобие, дава по-добри резултати за структурното подобие, което най-вероятно се дължи на високата вариация на интензитета на характеристичните групови честоти в Раман спектрите в сравнение с ИЧ.
- При работата с „разширената“ библиотека стойността на коефициента на корелация между структурното и спектралното подобие в библиотеката дава най-високи резултати.
- При работата с комбинирани спектрални библиотеки „спектрите“ в усреднената библиотека по-добре съответстват на структурата на съединенията, отколкото спектрите в оригиналните ИЧ и Раман библиотеки;
- Създадена е нова методика за анализ на ИЧ и Раман спектри на бинерни смеси при работа с многопроменлива линейна регресия и алгоритъм за изваждане с три вида коефициенти
- Най-добри резултати са получени при работата с коефициент на изваждане, получен от регресионния анализ и за ИЧ, и за Раман спектрите на смеси.
- Идентификацията на съединение по негов ИЧ спектър е принос в подкрепата на идентификационната способност на библиотечното търсене и практическата му насоченост.

6. ИЗПОЛЗВАНА ЛИТЕРАТУРА**:**

[1] Mikhail E. Elyashberg, “Infrared Spectra Interpretation by the Characteristic Frequency Approach.” Chichester, pp. 1307–1313, 1998.

[2] A. Savitzky and M. J. E. Golay, “Smoothing and Differentiation of Data by Simplified Least Squares Procedures.,” *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964.

[3] J. G. Grasselli, “International Union of Pure Jcamp-Dx , a Standard Format for Exchange of Infrared Spectra,” *Pure and Applied Chemistry*, vol. 63, no. 12, pp. 1781–1792, 1991.
[4] “ISIS ™ /Draw Tutorial,” *Isis*. MDL Information Systems, Inc, pp. 1–82.
[5] “Quick Reference Guide,” *OPUS. Spectroscopy software*. pp. 1–44, 2007.
[6] K. Varmuza, W. Demuth, M. Karlovits, and H. Scsibrany, “Binary Substructure Descriptors for Organic Compounds,” vol. 78, no. 2, pp. 141–149, 2005.
[7] “INTRODUCTION TO SCILAB,” *The sciLab consortium*, vol. 23, no. 1. p. 87, 2010.
[8] P. N. Penchev, N. T. Kochev, and G. N. Andreev, “IRSS: A Programme System for Infrared Library Search,” *Computes rendus-academie bulgare des sciences*, vol. 51, pp. 67–70, 1998.
[9] P. N. Penchev, A. N. Sohou, and G. N. Andreev, “Description and Performance Analysis of an Infrared Library Search System,” *Spectroscopy Letters*, vol. 29, no. 8, pp. 1513–1522, Dec. 1996.
[10] J. T. Clerc, “Automated spectra interpretation and library search systems,” in *Computer-Enhanced Analytical Spectroscopy*, H. L. C. Meuzelaar and T. L. Isenhour, Eds. New York and London: Plenum Press, 1987, pp. 145–161.
[11] W. O. George and H. A. Willis, Eds., *Computer Methods in UV, Visible and IR Spectroscopy*, Royal Soci. London, 1990.
[12] K. Varmuza, M. Karlovits, and W. Demuth, “Spectral similarity versus structural similarity: infrared spectroscopy,” *Analytica Chimica Acta*, vol. 490, no. 1–2, pp. 313–324, Aug. 2003.
[13] S. R. Lowry, D. A. Huppler, and C. R. Anderson, “Data base development and search algorithms for automated infrared spectral identification,” *Journal of Chemical Information and Modeling*, vol. 25, no. 3, pp. 235–241, Aug. 1985.
[14] Sadtler Research Labs, “The Sadtler IR Search Software Manual.” Division of Bio-Rad Laboratories, 1988.
[15] Пламен Николов Пенчев, “Компютърна интерпретация на молекулни спектри с цел разкриване на структурата на органични съединения; Дисертация за получаване на научна степен дхн,” Пловдивски университет “Паисий Хилендарски,” 2016.
[16] K. Varmuza and P. Filtzmozer, *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC Press. Taylor &Francis group, 2009.
[17] K. Baumann and J. T. Clerc, “Computer-assisted IR spectra prediction - linked similarity searches for structures and spectra,” *Analytica Chimica Acta*, vol. 348, no. 1–3, pp. 327–343, 1997.
[18] R. J. Rosenthal and S. R. Lowry, “Effects of sampling methodologies on FT-IR data base searching,” *Mikrochimica Acta*, vol. 89, no. 1–6, pp. 291–302, Jan. 1986.
[19] J. T. Clerc, E. Pretsch, and M. Ziircher, “Performance Analysis of Infrared Library Search Systems,” pp. 217–242, 1987.
[20] J. T. Clerc, E. Pretsch, and M. Zurcher, “Performance analysis of infrared library search systems,” *Mikrochimica Acta*, vol. 89, no. 1–6, pp. 217–242, Jan. 1986.
[21] B. Schrader, *Infrared and Raman Spectroscopy, Method and Applications*, vol. 11, no. 2. 1996.
[22] P. J. Larkin, *“IR and Raman Spectroscopy - Principles and Spectral Interpretation.”* 2011.
[23] B. H. Stuart, *Infrared Spectroscopy: Fundamentals and Applications*, vol. 8. Chichester, UK: John Wiley & Sons, Ltd, 2004.
[24] J. Zupan, “The Library Search in Analytical Chemistry,” vol. 550, pp. 466–472, 1982.
[25] E. Smith and G. Dent, *Modern Raman Spectroscopy - A Practical Approach*. Chichester, UK: John Wiley & Sons, Ltd, 2004.
[26] Z. Hippe, “Problem-solving methods in computer-aided organic structure determination,” *Journal of Chemical Information and Modeling*, vol. 25, no. 3, pp. 344–350, Aug. 1985.
[27] G. M. Banik, M. Scandone, R. Tuzynski, and D. Kernan, “A New Approach to Simultaneous Raman and IR Spectral Searches A new system for multitechnique spectral searching is described that utilizes analysis of,” 2005.

[28] G. M. Banik, D. Ph, T. Abshear, and K. Nedwed, “Multi-Technique Spectral Searching in KnowItAll,” in *Technical Note*, 2005, vol. 11, pp. 1–11.
[29] J. Zupan (Ed.), *Computer-supported Spectroscopic Data Bases*. Chichester, UK: Ellis Horwood, 1986.
[30] S. S. Williams, R. B. Lam, and T. L. Isenhour, “Search system for infrared and mass spectra by factor analysis and eigenvector projection,” *Analytical Chemistry*, vol. 55, no. 7, pp. 1117–1121, Jun. 1983.
[31] C. Klawun and C. L. Wilkins, “Joint Neural Network Interpretation of Infrared and Mass Spectra,” *Journal of Chemical Information and Computer Sciences*, vol. 36, no. 2, pp. 249–257, Jan. 1996.
[32] M. E. Munk, M. S. Madison, and E. W. Robb, “The Neural Network as a Tool for Multispectral Interpretation,” *Journal of Chemical Information and Computer Sciences*, vol. 36, no. 2, pp. 231–238, Jan. 1996.
[33] Г. Андреев, *Молекулна спектроскопия*. Изд. на ПУ —П. Хилендарски, 2010.
[34] R. M. Silverstein, F. X. Webster, and D. J. Kiemle, “Spectrometric identification of organic compounds, 7th edition,” *Journal of Molecular Structure*, vol. 30, no. 2, pp. 424–425, Feb. 2005.
[35] E. Pretsch, P. Bühlmann, and C. Affolter, *Structure Determination of Organic Compounds: Tables of Spectral Data*. 2013.
[36] L.-L. YU and L.-M. SHAO, “Qualitative Analyses of Open-Path Fourier Transform Spectra,” *Chinese Journal of Analytical Chemistry*, vol. 43, no. 2, pp. 226–232, Feb. 2015.
[37] O. O. Ilchenko, Y. V. Pilgun, A. S. Reynt, and A. M. Kutsyk, “NNLS and MCR-ALS Decomposition of Raman and FTIR Spectra of Multicomponent Liquid Solutions,” *Ukrainian Journal of Physics*, vol. 61, no. 6, pp. 519–522, Jun. 2016.
[38] H. Somberg, “Qualitative mixture analysis by Use of an IR Library search system,” 1988.
[39] P. N. Penchev, V. L. Miteva, A. N. Sohou, N. T. Kochev, and G. N. Andreev, “Implementation and testing of routine procedure for mixture analysis by search in infrared spectral library,” vol. 40, no. 4, pp. 1–5, 2008.
[40] N. Kochev, P. Penchev, G. Andreev, and K. Varmuza, “Improved Realisation of Maximum Common Substructure Concept for Structure Elucidation,” *Traveaux Scientifiques d’Universite de Plovdiv*, vol. 30(5), pp. 73–78, May 2001.
[41] P. Lampen, R. J. Lancashire, R. S. McDonald, P. S. McIntyre, D. N. Rutledge, and A. N. Davies, *A Generic JCAMP -DX Standard File Format*. 2002.
[42] E. Karpushkin, A. Bogomolov, Y. Zhukov, and M. Boruta, “New system for computer-aided infrared and Raman spectrum interpretation,” *Chemometrics and Intelligent Laboratory Systems*, vol. 88, no. 1, pp. 107–117, Aug. 2007.
[43] H. J. Luinge, “Automated interpretation of vibrational spectra,” *Vibrational Spectroscopy*, vol. 1, no. 1, pp. 3–18, Dec. 1990.
[44] H. J. Luinge, J. A. de Koeijer, J. H. van der Maas, J. M. Chalmers, and P. J. Tayler, “Application of an automated interpretation system for infrared spectra. Part II. Characterization of aromatic sulphone copolymers,” *Vibrational Spectroscopy*, vol. 4, no. 3, pp. 301–308, 1993.

7. ПУБЛИКАЦИИ:

1. Публикация в Bulgarian Chemical Communications – P. N. Penchev, S. H. Tsoneva, S. R. Nachkova; Spectral similarity versus structural similarity: Raman spectra. Volume 49 Special issue G (pp.61 –64) 2017
2. Публикация в Bulgarian Chemical Communications, RG Journal Impact: 0.23 – S. H. Tsoneva, S. R. Nachkova, G.N.Andreev, P. N. Penchev; Identification of mixture components by multiple linear regression and substraction of reference spectra: searching infrared and raman libraries. Volume 49 Special issue D (pp.19 –24), 2017

3. Публикация в Научни трудове на Русенския Университет - 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).
4. 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).
5. Публикация в Научни трудове на Русенския Университет - S.Tsoneva, S.Nachkova, P.Penchev; Joint Spectral Database of Infrared and Raman Spectra. Scientific works: : University of Ruse “Angel Kanchev”, 54, Issue 10.1, 36-40 (2014).

8. УЧАСТИЯ В КОНФЕРЕНЦИИ:

1. Научна конференция с международно участие 2013, РУ“Ангел Кънчев“, Филиал – Разград, гр.Разград;S. Tsoneva, S. Nachkova and P. Penchev; ATR spectra database of organic compounds (постер).

2. Scientific Researches of the Union of Scientist in Bulgaria – Plovdiv, (2013); P. Penchev, S. Tsoneva, Ts. Krusteva and S. Nachkova; Spectral Libraries of Vibrational Spectra (постер).

3. Семинар, проведен съвместно от ПУ“П.Хилендарски“, АСМ2 и Thermo scientific 2014; С.Начкова, С.Цоневаи П.Пенчев; Потребителски спектрални библиотеки (ИЧ, ATR, Раман, УВ-Вид и ^{13}С-ЯМР спектри), (постер).

4. Научна конференция с международно участие 2014, РУ“Ангел Кънчев“, Филиал – Разград, гр.Разград;С.Цонева, П.Пенчев и С.Начкова; Търсене по подобие в Раман спектрални библиотеки (постер).

5. Научна конференция с международно участие 2015, РУ“Ангел Кънчев“, Филиал – Разград, гр.Разград;S. Tsoneva, S. Nachkova and P. Penchev; Joint spectral database of Infrared and Raman spectra (постер).

6. 10th Chemistry Conference with international participation 9-11 October 2016 of University of Plovdiv “Paisii Hilendarski”, Plovdiv, Bulgaria; P. N. Penchev, S. H. Tsoneva, S. R. Nachkova, Spectral similarity versus vtructural similarity: Raman spectra (постер).

9. БЛАГОДАРНОСТИ:

- Благодаря на научният ми ръководител проф.дхн П.Пенчев
- Благодаря на колегите си, за тяхната отзивчивост и подкрепа, търпение и топло отношение и изключително ценни професионални съвети и насоки
- Благодарности на проектите НИ13-ХФ-006 и НИ15-ХФ-001, със съдействието на които посетих няколко конференции и семинари.