ПЛОВДИВСКИ УНИВЕРСИТЕТ ”ПАИСИЙ ХИЛЕНДАРСКИ”
ХИМИЧЕСКИ ФАКУЛТЕТ
КАТЕДРА АНАЛИТИЧНА ХИМИЯ И КОМПЮТЪРНА ХИМИЯ

СТЕФКА РУМЕНОВА НАЧКОВА

КОМПЮТЪРНИ МЕТОДИ ЗА ИНТЕРПРЕТАЦИЯ НА ^{13}C ЯМР СПЕКТРИ НА ПРИРОДНИ СЪЕДИНЕНИЯ

АВТОРЕФЕРАТ

НА ДИСЕРТАЦИОНЕН ТРУД ЗА ПРИСЪЖДАНЕ НА
ОБРАЗОВАТЕЛНАТА И НАУЧНА СТЕПЕН *ДОКТОР*

Област на висше образование: 4. Природни науки, математика и информатика; професионално направление: 4.2 Химични науки; докторска програма: Аналитична химия.

Научен ръководител:
Проф. д.н. ПЛАМЕН Н. ПЕНЧЕВ

ПЛОВДИВ 2018

Дисертацията е поместена в обем от 158 страници и съдържа 33 таблици, 31 фигури и 194 цитирани литературни източника.

Дисертационният труд е обсъден на заседание на катедра Аналитична химия и компютърна химия към Химическия факултет на ПУ "Хилендарски" на 13.02.2018 г. с взето решение за разкриване на процедура за защита.

Номерата на таблиците и фигурите, както и на цитираните източници съответства на номерацията в дисертационния труд.

СЪКРАЩЕНИЯ

ИС – информационно съдържание (на подструктурата)

ОИ – Обучителна извадка

ТИ – Тестваща извадка

ВИ – Валидираща извадка

кNN (k-Nearest Neighbours) – Метод на най-близките к съсада

MAD (Mean Absolute Deviation) – Средното абсолютно отклонение

I. ВЪВЕДЕНИЕ

Проблемът за разкриване структурата на съединенията заема важно място в съвременната химия. Често е необходимо обединяването на различни подходи и методи в процеса на определяне на неизвестна структура. Тези подходи могат да бъдат свързани, както с извършването на химични реакции за доказване, така и изцяло да се отнасят до обработката на спектрални данни. Като първа стъпка от процеса на разкриване на структурата обикновено се разглежда задачата за сравнение между спектъра/ите на непознатото съединение и спектралните данни на набор от известни съединения с цел разпознаване на структурата, ако тя се намира измежду тях, т.н. идентификация на структурата.

Широко приложение за разкриване структурата на органични молекули намира ядрено магнитния резонанс на въглеродни ядра, ^{13}C ЯМР, тъй като спектърът силно зависи от въглеродния скелет на съединенията. За разлика от протонния ЯМР, ^{13}C ЯМР обхваща широк интервал химични отмествания от около 250 ppm, които слабо се припокриват. Допълнително предимство е опростеният вид на протоннодекуплирания спектър, при който съответства по един сигнал на всеки въглероден атом или група магнитно еквивалентни въглеродни ядра. Наред с това, усъвършенстването на двумерните техники за регистриране на спектри, 2D ЯМР, през последните две десетилетия, позволява извличането на информация не само за топологичната свързаност на атомите, но и за стереоизомерията и конформационната гъвкавост на молекулата като цяло, от съществено значение при изследване на сложни природни съединения и биомолекули [1].

Интерпретацията на спектралните данни е времеемка и нелесна задача, дори за опитен спектроскопист, особено когато става въпрос за молекули, съдържащи голям брой атоми. Това е предпоставка за появата през 60-те години на миналия век и следващо бурно развитие на компютърните методи за автоматично разкриване на структурата на неизвестни съединения, едновременно с развитието на инструменталните техники [2].

Създадените комплексни системи за компютърно-подпомогнато разкриване на структурата (КПРС) имат за цел автоматична и пълна идентификация на неизвестното съединение. Често това не е възможно, не само поради несъвършенството на системите, а и по обективни причини, например, не винаги цялата информация за структурата се съдържа в спектъра [4]. Но дори получаването на частична информация от спектъра под формата на отделни структурни фрагменти, улеснява значително процеса на интерпретация и съкращава времето за анализ.

Библиотечното търсене, като част от компютърните методи за разкриване на структурата, е основна тема на настоящата дисертация.

II. ЦЕЛИ И ЗАДАЧИ

Предимствата на библиотечното търсене изобщо и в частност интерпретационното търсене, като част от комплексната работа за разкриване структурата на органични съединения, заедно с недостатъчното количество специализирани библиотеки със спектри на природни съединения обосноваха формулирането на основната цел на дисертационния труд:

Изследване приложението на интерпретационното библиотечно търсене за разкриване структурата на природни съединения по техните ^{13}C ЯМР спектри.

Наличието на спектрална колекция от природни съединения е необходимо за изпълнение на поставената цел, ето защо съставянето на такава библиотека по литературни данни присъства на първо място в списъка от задачи:

1. Съставяне на библиотека от ^{13}C ЯМР спектри на фитосъединения.
2. Оценка на работата на известен алгоритъм за интерпретационно библиотечно търсене и неговата функцията на надеждност за търсене на природни фитосъединения в неспециализирана библиотека органични съединения.
3. Проверка за преносимост на функцията на надеждност, създадена за неспециализирана библиотека, за търсене в библиотека от природни съединения.
4. Създаване на функция на надеждност за оценка резултатите от библиотечно търсене в създадената библиотека от фитосъединения.
5. Проверка на преносимостта на вероятностната функция, съставена за библиотеката от фитосъединения, при търсене в други библиотеки.
6. Тестване на възможностите за търсене по подобие на фитосъединения в различни библиотеки.

III. ИНТЕРПРЕТАЦИОННО БИБЛИОТЕЧНО ТЪРСЕНЕ

Методът за интерпретационно библиотечно търсене, описан от П. Пенчев и съавтори е реализиран в Windows-базирана програма с потребителски интерфейс INFERCNMR [86] и включва три обособени етапа: (1) извличане на подструктури от референтните структури в библиотеката; (2) оценка на вероятността всяка от тези подструктури да присъства в неизвестната структура и (3) отнасяне на сигнали от спектъра на неизвестното съединение към всеки от въглеродните атоми в подструктурата.

Оценката на надеждността на всяка от генерираните подструктури с програмата INFERCNMR се извършва от вероятностна функция, съставена за работната библиотека, посредством невронни мрежи с обратно разпространение на грешката (back propagation ANN). Вероятността е свързана с броя на верните и неверни подструктури, получени при дадена стойност q на изхода от невронната мрежа, така дискретната функция на надеждност P% се изчислява за i от 0.000 до 1.000 със стъпка 0.001, общо 1001 стойности. Надеждността, оценена от P% и съответстваща на q за дадената подструктура, се приписва на всяка от генерираните подструктури.

Проверката за приложимост на посочената функция на надеждност за търсене на спектри на природни съединения, както и съставянето на нови функции за библиотека от природни съединения, са направени с извадки от подструктури, получени по описания метод за интерпретационно библиотечно търсене с програмата INFERCNMR [86].

За тестване на ефективността на оригиналната функция на надеждност, както и на създадените нови такива, в настоящата дисертация са използвани два от статистическите критерии за оценка на ефективността на бинарни класификатори – точност на класификация за верните подструктури, наричана също прецизност (*precision, P*), и чувствителност (*sensitivity* или още *recall, R*) при прагове на надеждността 90%, 95% и 99%.

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad \text{(5 и 6)}$$

FP (False Positive) - *неверни, погрешно класифицирани като верни;*
TP (True Positive) - *верни, правилно класифицирани като верни;*
FN (False Negative) - *верни, погрешно класифицирани като неверни.*

IV. СПЕКТРАЛНИ БИБЛИОТЕКИ

За тестване на ефективността на интерпретационно библиотечно търсене, както и за сравнение на работата на оригиналната функция на надеждност с други функции, извлечени по различен начин, са използвани три спектрални библиотеки LAST, LAST1000 и PHYCHEM.

И трите библиотеки поддържат ароматни и тавтомерни връзки, които са обозначени в таблицата на свързаност [179]. Тавтомерните връзки са определени според дефиницията на IUPAC.

1. Библиотека LAST

Библиотека LAST е съставена от 38 225 уникални, напълно отнесени спектри, подбрани от няколко колекции (Merck, CIS, Sadtler и др.), съобразно

изискванията на системата за автоматично разкриване на структурата SESAMI. На всеки атом съответства еднозначно отнесено химично отместване, както и всеки спектър съдържа най-малко 6 отделни сигнала.

Интервалната молекулна формула за библиотеката е $C_{6\text{-}77}\,H_{0\text{-}148}\,N_{0\text{-}15}\,O_{0\text{-}38}\,F_{0\text{-}15}\,S_{0\text{-}8}\,Cl_{0\text{-}12}\,Br_{0\text{-}8}\,I_{0\text{-}4}$, а средната $C_{14.4}\,H_{18.5}\,N_{0.9}\,O_{2.6}\,F_{0.1}\,S_{0.1}\,Cl_{0.2}\,Br_{0.1}$

Библиотеката съдържа приблизително 7 000 съединения, чиито таблици на свързаност не са уникални, в това число влизат различни спектри на идентични съединения или спектри на стереоизомери.

2. Библиотека LAST1000

За проверка на работата на функцията на надеждност, съставена за библиотека LAST при търсене в библиотека с по-малък размер, както и за тестване на създадените от нас функции на надеждност за библиотека PHYCHEM с друга библиотека с подобен размер, но различен състав, бе съставена библиотека LAST1000.

Библиотеката съдържа 1000 съединения от библиотека LAST, подбрани по молекулна формула, така че да се доближават максимално по размер и елементен състав до тези във PHYCHEM. Селекцията е направена по отношение на броя атоми въглерод, кислород и азот в молекулната формула, според честотата на поява за дадена формула в библиотека PHYCHEM. Поради липсата на достатъчно на брой съединения от LAST, точно съвпадащи със съединения на PHYCHEM, са използвани съединения от LAST, съвпадащи с други формули от PHYCHEM (за които вече са намерени еквиваленти) или такива, различаващи се от тези във PHYCHEM с минимален брой от посочените елементи.

Средната и интервална молекулни формули $C_{22.4}\,H_{30.6}\,N_{0.23}\,O_{5.4}\,S_{0.1}\,Cl_{0.1}$ и $C_{8\text{-}77}\,H_{7\text{-}148}\,N_{0\text{-}6}\,O_{0\text{-}21}\,F_{0\text{-}6}\,S_{0\text{-}3}\,Cl_{0\text{-}4}\,Br_{0\text{-}4}\,I_{0\text{-}1}$, показват по-висок брой на въглеродните, водородните и кислородни атоми, но малко по-ниско средно съдържание на азот за избраните съединения в сравнение с LAST, същевременно близко по състав съдържание на тези елементи спрямо PHYCHEM.

3. Библиотека PHYCHEM

За целите на настоящата дисертация бе създадена спектрална колекция от хиляда протоннодекуплирани ^{13}C ЯМР спектъра на природни съединения, извлечени от растения. Използвани са спектри и структури на съединения, публикувани в списанието Phytochemistry в периода 2002-2006 г. Колекцията се състои от файлове в стандартен MOL формат, съдържащи 2D координатите на структурите на съединенията и сигналите на въглеродните атоми във ^{13}C ЯМР спектъра, изразени в милионни части (ppm).

Структурите на съединенията са начертани с програмния пакет ISISDraw 2.4 и са представени чрез таблица на свързаност, която показва вида на атомите

и типа на връзките между тях, а стойността на всеки ^{13}C сигнал е въведена в полето свойства на съответния атом, като реално число с точност до стотни или десети в зависимост от точността на химичните отмествания, докладвани в съответната публикация.

MOL файловете са преобразувани във формата на спектрална библиотека PHYCHEM, която се поддържа от програмата INFERCNMR. Всеки библиотечен запис съдържа химичното име и структурата на съединението, както и неговия ^{13}C ЯМР спектър, изразен като химични отмествания и мултиплетности на сигналите. За по-голямата част от спектрите е отбелязан видът на разтворителя, ако той е посочен в използвания източник. За структурите за които не е посочено номенклатурно наименование в статията, то е генерирано автоматично от структурата с модула ChemDrow Ultra 7.0 на пакета ChemOffice 2002

Елементния състав и размера на структурите в библиотеката, оценени от средната и интервална молекулна формула $C_{23.0}$ $H_{32.7}$ $N_{0.2}$ $O_{5.7}$ и C_{8-66} H_{7-292} N_{0-4} O_{0-23} S_{0-2} Br_{0-3} Mg_{0-1} Ca_{0-1} показват по-високо средно съдържание на въглерод и кислород в тях, спрямо библиотека LAST, т.е. структурите са по-големи.

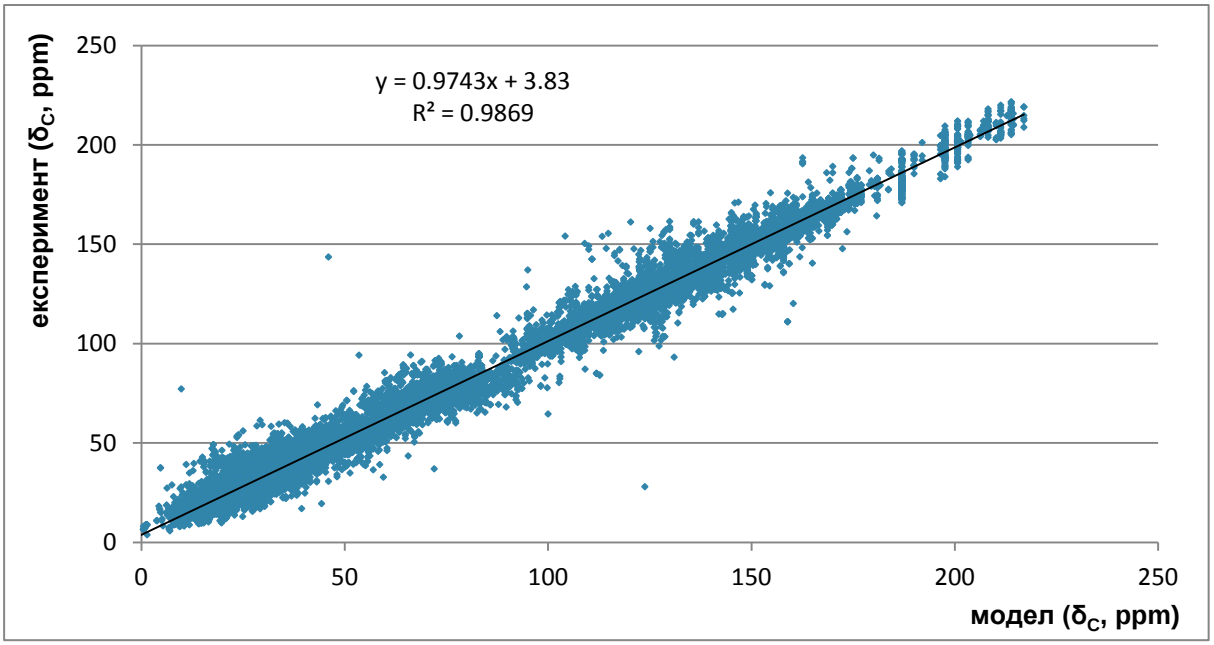
V. РЕЗУЛТАТИ И ДИСКУСИЯ

Всички спектри в създадената от нас библиотека PHYCHEM бяха предсказани с модула ChemDrow Ultra 7.0, работещ с адитивни схеми, с цел проверка за грешки. В резултат бяха открити няколко грешни структури, а някои от сигналите, за които бе изчислена голяма разлика с предсказаната стойност, бяха разгледани обстойно, а допуснатите грешки коригирани.

Средното абсолютно отклонение (MAD) и стандартното отклонение на разликите за цялата библиотека от общо 22 990 сигнала са оценени съответно на 4.5 и 6.2 ppm. Тези стойности са по-високи от публикуваните в литературата при използване на емпирични методи за предсказване на ^{13}C ЯМР спектри, в това число адитивни схеми, HOSE библиотечно търсене и невронни мрежи (NN), като често MAD е около и под 2 ppm [182].

На фигура 6 е представено регресионното уравнение на сигналите, като функция на експериментално регистрираните стойности от предсказаните за всички 22 990 сигнала на библиотека PHYCHEM. Наклонът на правата е по-нисък от гореспоменатите за други методи, а коефициентът на детерминираност (определеност) R^2 0.9869 е значително по-лош в сравнение с 0.9975 за HOSE и 0.9970 за NN съответно. Вероятно отчетената разлика се дължи на специфики на моделите, поддържани от конкретното приложение за предсказване на спектри. Същевременно, възможно обяснение за този резултат е наличието на необичайни структури в библиотеката, включително по-големи по размер от тестваните от

Elyashberg и съавтори, 80% от които съдържат по-малко от 20 въглеродни атома [182]. В библиотеката PHYCHEM приблизително 50% от структурите съдържат между 20 и 30 въглеродни атома, а само 30% по-малко от 20 атома.

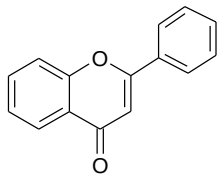
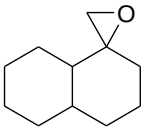
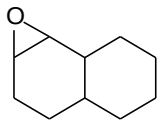
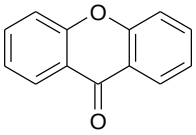
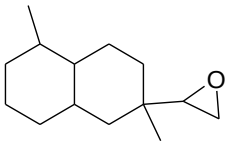
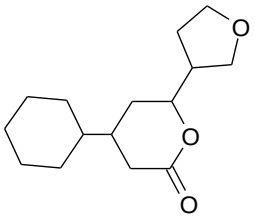
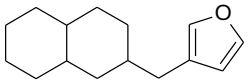
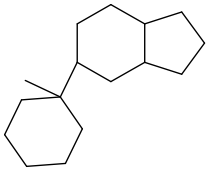
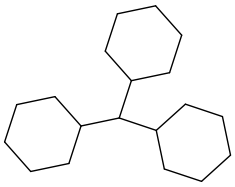
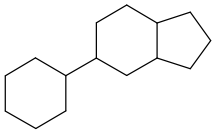
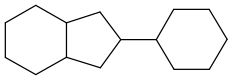
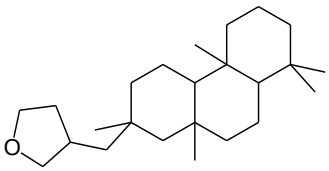
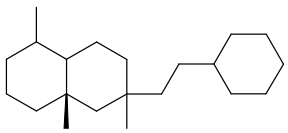
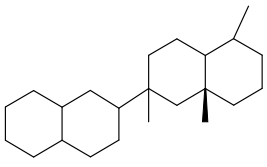
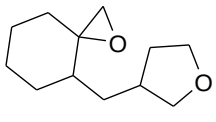
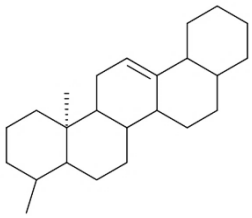
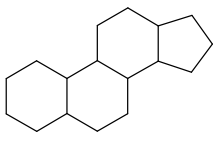
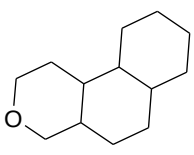


Фигура 6 Линейна регресия между предсказаните с адитивни модели и експериментално измерените стойности на химичните отмествания за всички сигнали в библиотека PHYCHEM.

За характеризиране на състава на библиотеката и нейните специфични особености, структурите на библиотечните съединения бяха анализирани по отношение на разнообразието от структурни фрагменти от които са изградени. Най-често срещаните фрагменти в библиотеката бяха извлечени с програмата Scaffold Hopper v19 (Tripod Development). Присъствието в библиотека PHYCHEM на най-разпространените уникални фрагменти (които не са част от други по-големи фрагменти), съдържащи повече от 10 тежки атома, е показано в таблица 6.

Прегледът на подструктурите в таблицата показва, че най-разпространените класове съединения в библиотеката са флавоноидите, поликетиди, сесквитерпени, срещат се също стероиди, циклични структури с тройни, пет и шестчленни пръстени и др. Най-високата честота бе отчетена за фрагмент с размер 17 тежки атома, който се съдържа в 5% от съединенията в библиотеката. Направено бе сравнение със спектралната библиотека от фитосъединения NAPROC-13, съдържаща повече от 20 000 структури [89]. Фрагментите с висока честота в библиотека PHYCHEM, бяха потърсени в NAPROC-13. Четири от тях имат честота на поява по-висока от 5%, а три от подструктурите не са намерени в библиотеката (Таблица 6, получер шрифт).

Таблица 6 Относителна честота на поява в проценти на най-големите от структурните фрагменти, срещащи се в библиотека PHYCHEM в 14 или повече библиотечни структури. С получер шрифт – относителна честота на поява или брой за всяка от подструктурите в свободната библиотека от природни съединения NAPROC-13, съдържаща повече от 20 000 съединения и техните въглеродни спектри.

5.0% **3.6%** 	1.8% **1.1%** 	1.5% **0.7%** 
2.5% **2.2%** 	1.4% **0.2%** 	1.5% **–** 
2.4% **1.2%** 	2.3% **7.5%** 	1.5% **–** 
2.4% **8.3%** 	1.4% **1** 	1.7% **0.2%** 
2.3% **–** 	2.2% **4.7 %** 	1.6% **0.3%** 
1.7% **4%** 	1.5% **6.2%** 	1.5% **0.7%** 

Yongye и съавтори сравняват пет структурни библиотеки от природни съединения [184]. Подобно на PHYCHEM често срещани в тях са скелетите на флавони, флавоноиди и стероиди, а средният процент на библиотечните съединения, които съдържат тези фрагменти, е съизмерим с този за PHYCHEM. Авторите въвеждат и няколко количествени показателя за сравнение на структурни библиотеки, между които са отношенията на общият брой уникални фрагменти N спрямо размера на библиотеката M, а също и броя на синглетите (фрагменти, съдържащи се само в една библиотечна структура, Nsing), спрямо N. Библиотеките посочени в статията показват стойности за N/M в интервала 0.18-0.49 и Nsing/N 0.40-0.74. Получените резултати за библиотека PHYCHEM, N/M=0.47 и Nsing/N=0.38, са съизмерими с цитираните.

За тестване на ефективността на интерпретационното библиотечно търсене на фитосъединения в неспециализирана библиотека от органични съединения, както и в библиотека от природни съединения са съставени три серии от съединения, чиито структури са публикувани в списание Phytochemistry (2001-2002 г., томове 57-60). Всяка от тях съдържа 100 съединения, които не присъстват в спектралната библиотека PHYCHEM и са потърсени поотделно в двете библиотеки. Списъците от подструктури, получени като резултат при търсене на 100-те съединения в дадена библиотека, са обединени в общ списък за всяка от трите серии, по този начин формирайки три извадки от подструктури, наречени обучителна (ОИ), тестваща (ТИ) и валидираща (ВИ) извадка. За съставяне на обучителната и тестваща извадка, съединения от една и съща статия са разпределяни в различни серии, с цел да се осигури максимално разнообразие на извадките. При съставяне на валидиращата извадка съединенията са избирани по едно от статия и без значение от размера на хитсписъка, включително такива за които търсенето не дава нито една подструктура. Това позволява да бъде отчетена ефективността на търсене изобщо, не само способността за разграничаване на верните от неверните подструктури от функцията на надеждност. За 8 от стоте съединения в серията, с която е получена ВИ, не е генерирана нито една подструктура.

Проверката за преносимост на функцията за библиотека, подобна по състав, но с различен размер, е направена с библиотека LAST1000.

Броят на съединенията в трите библиотеки, както и броят на подструктурите в извадките, използвани за извличане на функция и тестването ѝ, са представени в таблица 7.

При търсенето в библиотека LAST1000 само 98 от ОИ, 96 от ТИ, и 85 от съединенията във ВИ са дали повече от една подструктура, а при LAST само за четири съединения от ВИ не е предсказана нито една подструктура.

Таблица 7 Състав на извадките от подструктури, получени при интерпретационно търсене на три серии от природни съединения в три библиотеки. **(в) –** брой верни подструктури в извадката.

Библиотека	размер	ОИ	ОИ (в)	ТИ	ТИ (в)	ВИ	ВИ (в)
PHYCHEM	1 000	15 244	6 732	13 508	5 852	8 638	3 763
LAST1000	1 000	12 176	5 030	10 578	4 499	5 850	2 157
LAST	38 225	169 086	58 056	146 720	53 493	90 340	13 508
LAST	38 225	169 036*	62 851*				

** извадката от съединения е използвана при тестване на оригиналната функция на надеждност на библиотека LAST и се различава от използваната ОИ за другите библиотеки.*

1. Интерпретационно търсене на спектри на природни съединения в обща библиотека от органични съединения

В таблица 8 са представени стойности за прецизностите при прагове 90%, 95% и 99% вероятност за четирите извадки от подструктури (Таблица 7), получени при търсене на спектри на природни съединения в библиотека LAST. Сравнението с праговите стойности за първата извадка ОИ* показва, че при предсказването на вероятността подструктурата да бъде вярна, оригиналната функция на надеждност на тази библиотека работи коректно.

Таблица 8 Оценка на ефективността на оригиналната функция на надеждност на библиотека LAST за интерпретационно търсене на природни съединения с ОИ, ТИ и ВИ (по 100 съединения) при прагове – 90.0%, 95.0% и 99.0% вероятност.

	Прецизност, %		
извадка	**90.0**	**95.0**	**99**
ОИ*	93.2	95.8	98.7
ОИ	*81.1*	*87.1*	*94.0*
ТИ	89.5	95.7	99.1
ВИ	89.9	95.9	99.4

** със 100 спектъра от списание Phytochemistry (2002 г., том 58-59)*

Съставянето на ОИ, ТИ и ВИ по-късно, позволи работата на вероятностната функция да бъде проверена с по-голям брой природни съединения. От таблицата се вижда, че направените заключения за вероятностната функция се потвърждават за две от извадките, докато при едната от тях прецизността е по-ниска и за трите прагови вероятности. Бе установено, че неверните подструктури с приписана надеждност над 99.0% са получени при търсене на 5 от 100-те съединения в списъка. Вероятно това се дължи на неспособност на функцията да оценява коректно някои видове специфични подструктури.

Така представените резултати не носят информация за чувствителността на метода, т.е. не показват каква част от всички верни подструктури в извадката, са били разпознати като верни с вероятност над даден праг. За първата извадка, при избраните три прагови вероятности са получени чувствителности съответно 59.9%, 42.4% и 20.9%, но за останалите три е изчислена по-ниска чувствителност в интервала 11%-15% за праг 99% (Таблица 9).

Таблица 9 Чувствителност на метода за интерпретационно библиотечно търсене при търсене на природни съединения в библиотека LAST и нейната оригинална функция на надеждност за три прагови вероятности – 90.0%, 95.0% и 99.0%.

	Чувствителност, %		
Извадка	**90.0**	**95.0**	**99.0**
ОИ*	59.9	42.4	20.9
ОИ	53.5	36.2	12.9
ТИ	49.6	33.7	11.5
ВИ	48.0	32.9	14.4

Тези чувствителности са по-ниски в сравнение с получените при комплексната оценка (*leave-one-out* кръстосано валидиране с голяма валидираща извадка от 12 740 съединения), направена при създаването на вероятностната функция на библиотека LAST [86]. За трите прага вероятностните функции при различните толеранси дават средно 58.2%, 47.7%, 32.1% чувствителност. Тези стойности са получени като претеглено средно спрямо общия брой на верните подструктури, генерирани за всеки отделен толеранс.

Получените резултати потвърждават приложимостта на интерпретационното търсене в неспециализирана библиотека от органични съединения за разкриване структурата на природни съединения. Предсказаните подструктури разкриват средно 30% от структурата на потърсеното съединение, изразена като брой тежки атоми.

Детайлното разглеждане на получените резултати допълнително позволява да бъдат формулирани няколко извода. Първо, функцията на надеждност оценява коректно извлечените при търсене подструктури, с изключение на някои специфични случаи, при които за отделни съединения голям брой от генерираните неверни подструктури се оценяват с висока надеждност. Чувствителността на метода за търсене на фитосъединения в неспециализирана библиотека от органични съединения е почти два пъти по-ниска от намерената при кръстосано валидиране с голяма извадка от органични съединения. А това означава, че въпреки големия брой верни подструктури, генерирани при търсенето, често в резултатите отсъстват надеждни

подструктури, които биха могли да бъдат използвани като вход за структурен генератор. По-големи подструктури се извличат от референтни съединения със структура, близка до неизвестната като цяло, а не само на отделни изолирани части от нея, макар че последните често се оценяват с по-висока надеждност поради по-доброто съвпадение на сигналите им. Вероятно за извличане на подструктури, разкриващи голяма част от структурата на непознатото съединение, по-ефективно би било търсене в библиотека от природни съединения.

2. Интерпретационно търсене в библиотека от природни съединения PHYCHEM и библиотека LAST1000

Възможностите за извличане на подструктури с метода на интерпретационно библиотечно търсене и техните характеристики при търсене на природни съединения в библиотека, различна от оригиналната, но с нейната функция на надеждност, бяха тествани с библиотеките PHYCHEM и LAST1000.

Изчислената прецизност за трите прагови надеждности с подструктурите от ОИ, ТИ и ВИ, получени при търсене на трите серии природни съединения в библиотеките, са показани на таблица 13. За всички прагови вероятности са получени значително по-ниски стойности на оценената вероятност (прецизност) и при двете библиотеки, т.е. вероятностната функция не работи коректно. В съответствие с очакваното вероятностната функция, създадена за библиотека LAST, не е преносима за друга библиотека, различна по състав от оригиналната.

Таблица 13 Оценка на преносимостта на оригиналната функцията на надеждност на библиотека LAST за три прагови вероятности – 90.0%, 95.0% и 99.0%..

а) Библиотека PHYCHEM

	Прецизност, %		
извадка	**90.0**	**95.0**	**99.0**
ОИ	55.9	69.4	73.8
ТИ	54.6	67.9	69.1
ВИ	55.4	69.4	73.6

б) Библиотека LAST1000

	Прецизност, %		
извадка	**90.0**	**95.0**	**99.0**
ОИ	53.4	58.1	63.9
ТИ	54.3	67.1	73.6
ВИ	48.6	61.7	69.4

3. Създаване и тестване на функция на надеждност за оценка на резултатите от търсене в библиотека PHYCHEM

Вероятностната функция, оценяваща верността на предсказаните структури при интерпретационното търсене може да бъде съставена по различен начин, посредством прилагане на линейни и нелинейни хемометрични методи с предварително разпределение на класовете, като линеен дискриминантен анализ, невронни мрежи и др. Стандартен подход за съставяне на бинарен класификатор е използването на многопроменливи регресионни методи от тип логистична регресия.

3.1 Избор на подструктурни параметри

Съставянето на функция на надеждност с висока ефективност на разграничаване на верните от неверните подструктури изисква внимателно изследване на зависимостите между входните подструктурни параметри, както и тяхната свързаност с целевата променлива, указваща верността на подструктурата (0/1). Подходящ метод за оценка на величината на такава зависимост, но също и нейната посока е корелационния анализ на параметрите.

За признаци, които не са разпределени нормално, каквито са повечето от подструктурните параметри, описващи резултатите от търсенето, параметричните мерки, като коефициентът на корелация на Пиърсън, не оценяват коректно зависимостта между променливите. Коефициентът на рангова корелация на Спиърман, *Rank-order correlation coefficient (*r_s), от друга страна, борави с ранговоскалираните променливи и позволява да бъде оценено наличието на по-обща монотонна зависимост между тях, включително праволинейна. В частния случай на извличане на функция на надеждност от единичен критерий, съставен от основните подструктурни параметри, тя ще зависи само от начина по който са подредени верните и неверни подструктури според този критерий, но не и от големината на критерия, ето защо подходящи за отсяване на разграничаващата способност на основните и комбинираните подструктурни параметри са мерки, които боравят с ранговете, вместо със суровите стойности на величините. За оценка на разграничаващата способност на параметрите чрез корелацията им с бинарната променлива, показваща верността на подструктурата, използвахме ранговобисериалния коефициент, Rank biserial correlation coefficient (r_{rb}) и отношението на Фишер (F) на ранговете:

$$F = \frac{(M_1 - M_0)^2}{s_1^2 + s_0^2} \quad (7)$$

M_1 и M_0 – *среден ранг на параметъра за групата на верните и неверните подструктури;*

s_1 *и* s_0 – *стандартни отклонения на ранкираните параметри в двете групи*

$$r_{rb} = \frac{2(R_1 - R_0)}{n} \quad (8)$$

R_1 и R_0 – *среден ранг на параметъра за групата на верните и неверните подструктури*

n – *броят на всички подструктури*

Критерият за значимост z на коефициента r_{rb} за голяма извадка клони към нормално разпределение и има стойност 2.58 при ниво на значимост α 0.01. Получените стойности за коефициента и неговия критерий z, както и F са представени в таблица 14.

Таблица 14 Характеристики на основните параметри на подструктурите в обучителната и тестващата извадка, получени при търсене в библиотека PHYCHEM.

	ОИ			**ТИ**		
параметър	**r_{rb}**	**|z|**	**F**	**r_{rb}**	**|z|**	**F**
sLO	0.61	65.20	0.73	0.60	59.59	0.73
sSO	0.67	71.33	0.49	0.65	64.71	0.49
sFV	-0.29	30.92	0.13	-0.29	28.70	0.13
RMSD	-0.30	32.37	0.12	-0.27	27.38	0.15
MaxDev	-0.26	27.82	0.09	-0.23	22.55	0.12
sNA	-0.10	10.26	0.01	-0.10	10.03	0.01
sSR	-0.03	2.96	**0.00**	-0.05	5.42	**0.00**
sHS	**0.02**	**2.50**	**0.00**	**0.00**	**0.20**	**0.00**
sHSC	**0.14**	**14.49**	**0.02**	**-0.14**	**10.71**	**0.03**
sIHS	-0.14	15.21	0.03	-0.14	13.55	0.03
sIHSC	-0.15	15.64	0.02	-0.11	11.25	0.03

r_{rb} – *рангово бисериален коефициент на корелация;* **z** *– критерий за значимост на коефициента* r_{rb}, z (2.58, α =0.01); **F** – *отношението на Фишер на рангово скалираните параметри.*

Разграничаващата способност на основните параметри, оценена от бисериалния корелационен коефициент и отношението на Фишер, бе сравнена с получената за комбинираните параметри, използвани в оригиналния метод. Резултатите показват, че не само sSO/sSR, а и останалите два параметъра имат по-ниски коефициенти от променливите от които са съставени поотделно. Макар че тези нормировки имат ясен химически/спектроскопски смисъл, за подобряване на разграничаващата способност бяха проверени различни комбинации от основни параметри (Таблица 15).

Параметъра sNA бе включван в числителя на комбинираните променливи, въпреки обратната корелацията с целевата променлива, тъй като едно от предимствата на функцията на надеждност би било преимуществено класифициране с висока вероятност на по-големите по-размер, и съответно по-

полезни, от верните подструктури. В действителност получените комбинирани параметри с sNA в числителя дадоха по-добри резултати.

Получена бе значителна разлика на корелацията между двата най-значими параметъра sSO и sLO при двете библиотеки PHYCHEM и LAST. Корелацията на основните параметри с целевата променлива в библиотека LAST показа, че връзката между присъствието на подструктурата в библиотеката (sLO) и верността на подструктурата е по-слаба в голямата библиотека, вероятно резултат от по-ниския дял на съединения със специфична структура, подобна на структурата на потърсените природни съединения.

Таблица 15 Характеристики на комбинирани параметри, извлечени от основните параметри на подструктурите от обучителната и тестващата извадка, ОИ и ТИ, получени при търсене в библиотека PHYCHEM.

	ОИ			ТИ		
параметър	**r_{rb}**	**|z|**	**F**	**r_{rb}**	**|z|**	**F**
sSO/sLO	-0.37	39.22	0.25	-0.11	37.30	0.26
sSO/sSR	-0.42	45.09	0.32	-0.37	41.13	0.30
sFV/sNC	-0.25	26.87	0.10	-0.25	24.45	0.10
sSO*sLO	0.65	69.49	0.87	0.64	63.80	0.81
sSO*sLO/sFV	0.66	70.37	0.96	0.65	64.71	0.90
sSO*sLO*sNA/sFV	0.69	73.15	1.09	0.67	66.82	1.00
1/(sFV+sRMSD)	0.33	34.63	0.17	0.32	31.72	0.16
1/(sFV+MaxDev)	0.38	40.29	0.23	0.35	35.11	0.20
sLO/(MaxDev+1)	0.64	68.32	0.89	0.62	61.62	0.79
sLO/sFV	0.63	67.08	0.85	0.62	61.59	0.79
sLO*sNA	0.64	68.28	0.90	0.62	61.87	0.81
sLO/(MaxDev+1)/sFV	0.66	70.60	0.98	0.64	64.26	0.89
sLO*sNA/(MaxDev+1)/sFV	0.70	74.22	1.15	0.67	67.05	1.02
sSO*sLO*sNA/sFV/(MaxDev+1)	0.72	76.38	1.25	0.69	69.31	1.12
sSO*sLO*sNA*sIHS/sFV/(MaxDev+1)	**0.75**	**79.92**	**1.47**	**0.72**	**73.10**	**1.34**
sSO*sLO*sNA/sIHS/sFV/(MaxDev+1)	0.68	71.80	1.03	0.66	65.52	0.94

r_{rb} – *рангово бисериален коефициент на корелация;* **z** *– критерий за значимост на коефициента* r_{rb}, z (2.58, α =0.01); **F** – *отношението на Фишер на рангово скалираните параметри.*

Съществените разлики в поведението на някои от параметрите като sSR за различни библиотеки и различен тип на съединенията, потърсени в тях, обясняват липсата на преносимост на оригиналната вероятностна функция, създадена за LAST, при оценка на резултатите от търсене в библиотеката от природни съединения PHYCHEM, както и поставят под съмнение възможността за създаване на преносима функция.

3.2 Логистична регресия

Много често за създаването на подобен вид бинарни класификатори, каквато по своята същност е функцията на надеждност, се използва логистична регресия със или без смесени членове,

$$LR = \frac{e^p}{e^p + 1} = \frac{1}{1 + e^{-p}}$$

$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \quad (10)$$

Такава бе изведена от основните и комбинирани подструктурни параметрите с пакета STATISTICA 10. Бяха тествани всички алгоритми, поддържани от програмата, от които само два дадоха резултат поради наличието на корелация между променливите, затрудняваща оптимизацията. Регресионните модели, извлечени от ОИ, бяха приложени за изчисляване на надеждностите на подструктурите от ТИ и ВИ, с които са определяни характеристиките на метода – чувствителност и прецизност.

Бяха съставени три типа модели (1) от основните параметри в таблица 14 с висока разграничаваща способност между верните и неверните подструктури, (2) от комбинации между основните параметри, които показват разграничаваща способност, по-висока от всеки от основните поотделно, и някои основни параметри, неприсъстващи в комбинациите и (3) от комбинирания параметър с най-висока разграничаваща способност от таблица 15 .

За четирите най-добри модела бе отчетена чувствителност с ОИ от 10.4% до 13.3% за прагова надеждност 99%. Поради недостатъчно високата чувствителност, както и факта, че способността за отсяване на верните от останалите подструктури от логистичната регресия не е пряко зависима от понятието вероятност, функция на надеждност бе извлечена от стойността, изчислена с логистичната регресия, подобно на оригиналния метод, при който такава вероятност е табулирана за стойностите на изхода от невронните мрежи.

Чувствителността и прецизността на метода, получени с ТИ и ВИ за четирите логистични модела са представени в таблица 18.

Извличането на вероятностна функция на надеждност от стойностите на логистичната регресия не подобрява резултатите за изследваните функции. За всеки от четирите модела поне в една от извадките невярна подструктура е оценена с вероятност над 99.9%. При първата и третата от функциите, чувствителността остава ниска, около и под 10% за прагова надеждност 99%, докато при втората и четвъртата функция тя е повишена значително, но прецизностите остават по-ниски праговите.

За библиотеки LAST и LAST1000 не бе възможно съставяне на регресионни модели от основните подструктурни параметри. Тестовете показаха, че вероятностните функции, извлечени с логистичната регресия от комбинирания

параметър с най-висока разграничаваща способност за двете библиотеки, не работят коректно. За библиотека LAST с обучителната извадка бяха отчетени прецизности за трите прагови вероятности в интервала 60%-65%, а за LAST1000 не повече от 97% за най-високия праг 99%. С тестващата и валидиращата извадки за LAST1000 бяха постигнати прецизности до 85.3% и 86.3% за праг 99.0%.

Таблица 18 Прецизност и чувствителност за ТИ и ВИ на функцията на надеждност, извлечена с логистична регресия от ОИ. Таблицата включва резултатите за четири модела: I - sNA, RMSD, MaxDev, sFV, sLO, sSO и sIHS; II - sSO*sLO*sNA/sFV, sIHS, RMSD и MaxDev; III - sSO, sLOs*sNA*sIHS, MaxDev*sFV; и IV - sSO*sLO*sNA*sIHS/sFV/(MaxDev+0.1).

Извадка/модел	P% 90	P% 95	P% 99	R% 90	R% 95	R%99	H99.0	H99.9
ТИ / I	93.5	94.9	96.0	27.6	20.5	10.7	26	24
ВИ / I	96.8	97.5	99.8	27.1	20.0	11.2	1	0
ТИ / II	99.5	99.6	100.0	19.0	15.9	11.2	0	0
ВИ / II	99.6	100.0	100.0	20.2	17.0	13.4	0	0
ТИ / III	94.4	96.8	95.5	30.0	20.5	8.8	24	0
ВИ / III	94.6	96.5	99.0	31.4	23.0	10.2	4	0
ТИ / IV	100.0	100.0	100.0	17.5	14.7	9.4	0	0
ВИ / IV	99.6	99.7	100.0	20.0	16.9	12.4	0	0

***P%** прецизност за дадена прагова надеждност в проценти; **R%** чувствителност за дадена прагова надеждност в проценти; **H99.0** и **H99.9** брой на неверните подструктури в извадката с приписана надеждност по-голяма или равна на 99.0% и 99.9%, съответно.*

Получените резултати потвърждават ограниченото приложение на подхода, базиран на логистична регресия, в неговия опростен вариант, за извличане на функция на надеждност, предназначена да оценява резултатите от интерпретационно библиотечно търсене.

3.3 Критерий за сортиране на резултатите от търсене

В процеса на търсене на възможни решения, бяха тествани множество критерии, част от които основани на линейна комбинация от параметри, съставени от основните. В изразите бе включен и размерът на библиотеката, LS=1000. Такова нормиране на критерия не оказва влияние върху неговата ефективност, но е пряко свързано с възможността за пренос на функцията към

друга библиотека, съдържаща различен брой съединения. Тъй като сортиращият критерий е съставян така, че по-големите негови стойности да сортират напред в хитсписъка преимуществено верните подструктури, параметрите, които показаха положителна корелация с целевата променлива, бяха поставяни в числителя заедно с параметъра sNA.

За всички съставени критерии бе изчислена чувствителност за 90%, 95% и 99% прагова надеждност на ОИ и ТИ. Критериите и техните чувствителности за ОИ са представени в таблица 20.

Таблица 20 Чувствителност в проценти за три прагови надеждности, определена с ОИ на библиотека PHYCHEM.

№	Критерий	R%90	R%95	R%99	H99	H99.9
1	$К1 = \frac{sLO.sSO}{LS}\frac{sNA}{sFV}$	46.9	33.4	21.5	13	0
2	$К2 = \frac{sLO.sSO}{LS}\frac{sNA}{sFV}\frac{1}{MaxDev+0.1}$	50.0	39.0	27.1	18	0
3	$К3 = \frac{sLO.sSO}{LS}\frac{sNA}{sFV}\frac{sHDISumU}{(MaxDev+0.1)}$	53.3	39.8	24.8	16	0
4	$К4 = \frac{sLO.sSO}{LS.}\frac{sNA}{sFV}\frac{\sqrt{sSO+sRS}}{sRS}\frac{1}{MaxDev+0.1}$	51.6	36.0	22.3	15	0
5	$К5 = f\frac{sLO.sSO}{LS}\frac{sNA}{sFV}+(1-f)(2-MaxDev)$	46.5	38.5	28.3	19	0
6	$К6 = f\frac{sLO.sSO}{LS}\frac{sNA}{sFV}sHIS+(1-f)(2-MaxDev)$	44.7	37.3	26.5	18	0

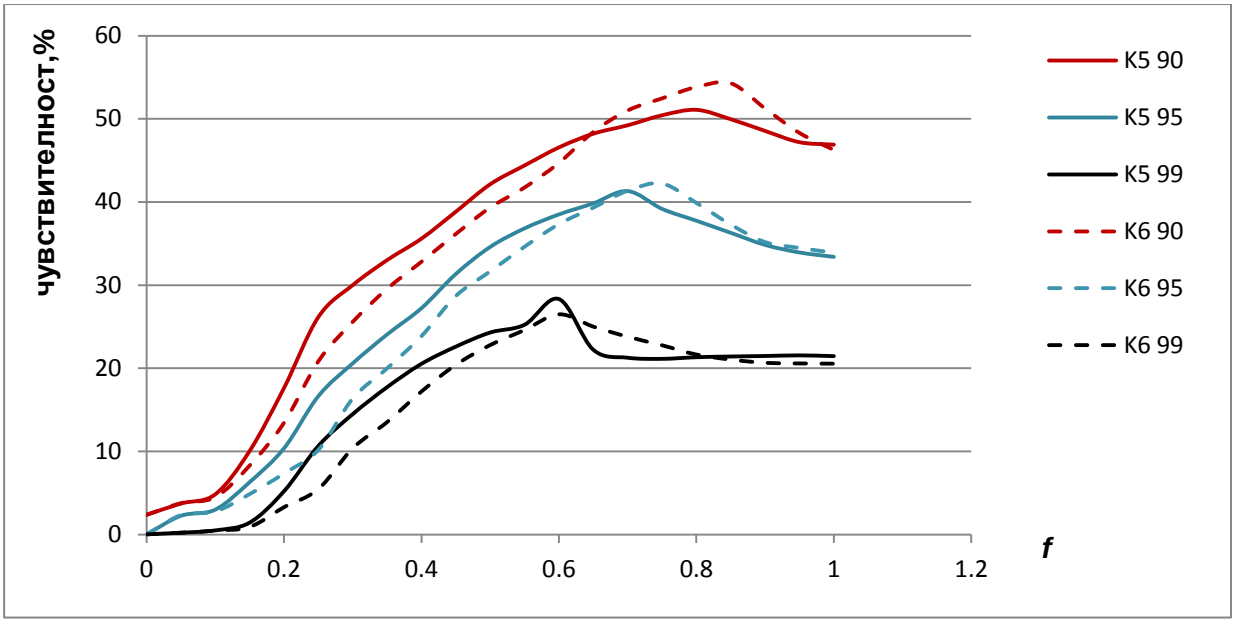
***R%** чувствителност за дадена прагова надеждност в проценти;* **H99** *и* **H99.9** *брой на неверните подструктури в извадката с приписана надеждност по-голяма или равна на 99.0% и 99.9%, съответно.*

Първия от критериите на практика е комбинираният параметър с най-висока разграничаваща способност между верните и неверните подструктури. В останалите изрази по различен начин е включено максималното отклонение на сигналите (толеранса), както и някои от другите подструктурни параметри.

За два от критериите, толерансът е добавен като втори член в линейна комбинация от два компонента. Тегловният фактор *f*, който определя участието на двата компонента в уравнението, е оптимизиран в интервала от 0.0 до 1.0 (Фигура 15).

Тъй като евентуалната полза от прилагането на такъв критерий при библиотечното търсене би била неговата способност да отсява верните подструктури с висока надеждност, за оптимална стойност на *f* ние избрахме

тази, при която бе отчетена най-висока чувствителност за 99% прагова вероятност, *f* =0.6.



Фигура 15 Влияние на тегловния фактор *f* върху чувствителността за 90%, 95% и 99% прагова вероятност за функциите, извлечени от ОИ с критерии К5 (плътна линия) и К6 (пунктир).

При сравняване на резултатите за мерките прецизност и чувствителност, получени с двете извадки ТИ и ВИ (Таблица 21), става ясно, че за някои от изследваните критерии прецизността е по-ниска при 99.0% праговата надеждност. Най-ниската стойност е 98.6% на критерий 5 за ТИ и критерий 1 за ВИ 98.4%, въпреки това при тези сортиращи критерии нито една от неверните подструктури не е оценена с надеждност 99.9% или по-висока. Критериите за ефективност на функцията, извлечена от най-добрия комбиниран параметър за библиотеки LAST и LAST1000, са представени в таблица 25.

Таблица 21 Прецизност и чувствителност на метода, намерени с подструктурите от а) ТИ и б) ВИ с функция, извлечена от ОИ на библиотека PHYCHEM.

а) с подструктурите от ТИ

№	Критерий	P% 90	P% 95	P% 99	R%90	R%95	R%99	H99.0	H99.9
1	*K1*	88.1	94.6	99.0	42.6	31.0	19.7	12	0
2	*K2*	89.3	94.9	98.9	46.2	35.6	24.9	16	0
3	*K3*	89.7	95.5	99.5	49.4	36.3	23.2	7	0
4	*K4*	89.5	95.1	98.8	48.3	34.2	20.7	15	2
5	*K5, f0.6*	89.3	95.4	98.6	42.8	34.5	25.9	22	0
6	*K6, f0.6*	89.1	94.6	98.9	41.4	34.2	24.0	16	0

б) с подструктурите от ВИ

№	Критерий	P% 90	P% 95	P% 99	R%90	R%95	R%99	H99.0	H99.9
1	*K1*	93.6	96.2	98.4	43.5	32.2	20.9	13	0
2	*K2*	95.1	96.7	98.6	45.8	36.6	26.1	14	0
3	*K3*	93.9	95.9	98.7	50.3	38.0	24.9	12	3
4	*K4*	90.8	95.4	98.9	49.5	35.0	22.3	9	0
5	*K5, f0.6*	90.3	95.4	98.6	44.4	37.0	26.9	14	0
6	*K6, f0.6*	89.4	94.0	99.0	43.9	36.3	25.6	10	3

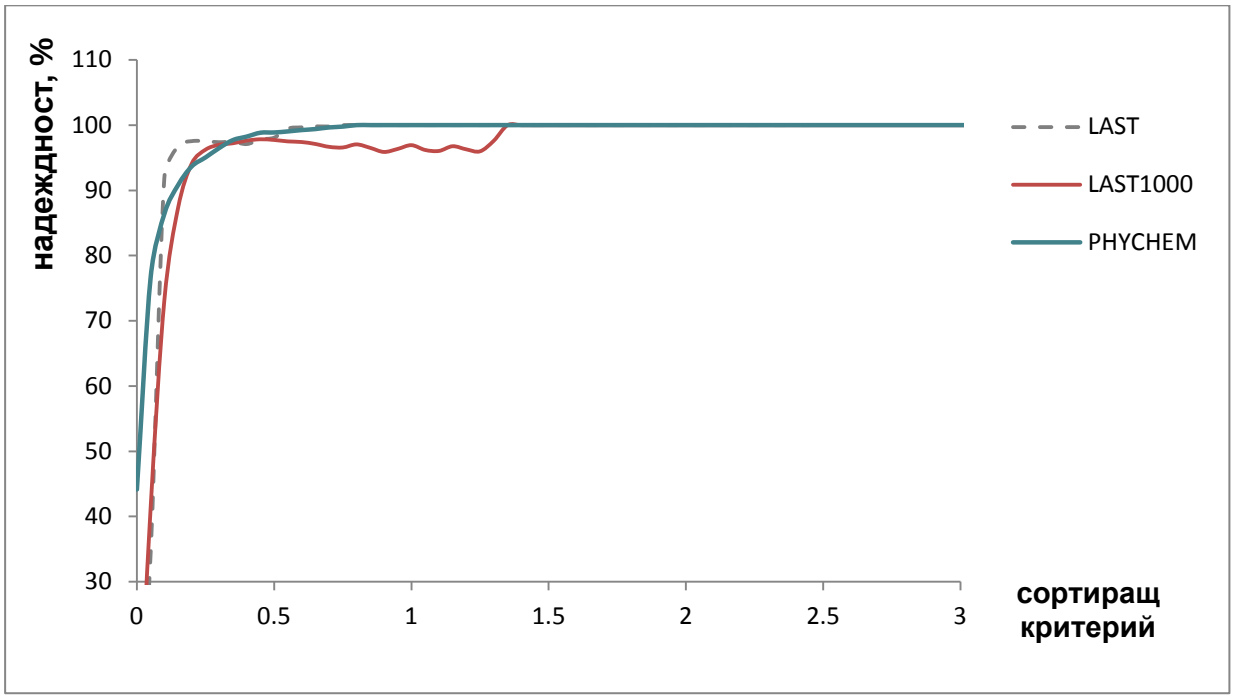
P% и **R%** *прецизност и чувствителност за дадена прагова надеждност в проценти;* **H99** и **H99.9** *– брой неверни подструктури с надеждност над 99.0 и 99.9%.*

Таблица 25 Прецизност и чувствителност на метода, изчислени за ТИ и ВИ на двете библиотеки LAST и LAST1000 с функция на надеждност, извлечена от най-добрия комбиниран параметър sSO*sLO*sNA* sHDISumU/(MaxDev+1)/sSR.

извадка/библ.	P%90	P%95	P%99	R%90	R%95	R%99
ТИ/LAST1000	93.0	95.8	100	31.9	23.7	0.8
ВИ/LAST1000	91.5	95.6	100	39.5	30.5	0.8
ТИ/LAST	94.6	98.4	99.4	32.3	22.1	1.5
ВИ/LAST	96.5	98.3	100	29.8	22.5	1.3

P% и **R%** *прецизност и чувствителност за дадена прагова надеждност в проценти;*

Графиката на функцията на надеждност, извлечена от най-добрите сортиращи критерии за трите библиотеки, представена на фигура 17, показва, че надеждността в проценти плавно нараства с нарастване на критерия само за библиотека PHYCHEM. При другите две библиотеки, кривите имат сложна форма с области на понижаване и повишаване. Разпределенията на верните и неверните подструктури и съответно тяхната надеждност в проценти по отношение на сортиращите критерии показват по-добро отсяване на верните подструктури от функцията, извлечена за библиотека PHYCHEM (Фигура 17). Максимумът за верните подструктури в тази библиотека е в интервала 99%-100%, докато при останалите две той е между 97%-98%. Това обяснява ниската чувствителност, отчетена за тези две библиотеки за праг на надеждност 99%.



Фигура 17 Зависимост на надеждността в проценти от най-добрия за всяка библиотека сортиращ критерий, с който е извлечена вероятностната функция.

Невъзможността за съставяне на ефективен сортиращ критерий за двете библиотеки от органични съединения отхвърля този тип функция като приложим за оценка на резултатите от интерпретационно библиотечно търсене, тъй като не осигурява надежден начин за актуализиране при разширяване на библиотеката с нови съединения, нито алтернатива за приложение при други библиотеки.

3.4 KNN функция на надеждност

Определянето на вероятността подструктурата да присъства в съединението, според стойностите на набор от дескриптори, на практика е задача за класификация на подструктурата към два възможни класа на верните и неверните подструктури в n-мерното пространство на нейните признаци.

Методът на най-близките ***к*** съседа (k-Nearest Neighbours) съчетава предимството на нелинейните методи с възможността за класификация на обекти, чиито класове се припокриват и/или не са компактни, а заемат повече от една отделни области от работното пространство. Най-простият от начините за определяне класа на обекта е т.н. *majority vote* – за предварително дефиниран **к** брой съседи, принадлежността на обекта се определя от класа на мнозинството най-близки до него **к** обекта в обучаващата извадка. От своя страна разпознаващата и предсказваща способност на метода обикновено са по-ниски от 100%, поради наличието на, макар и малък брой, погрешно класифицирани обекти.

Съставяне на функция

Вероятността дадена подструктура да е вярна може да се оцени с отношението на броя верни спрямо броя на всички **к** подструктури, най-близки до търсената, от подструктурите в цялата обучителна извадка. Алтернативен начин за оценка на надеждността в проценти, приписвана на всяка подструктура, е табулирането на вероятността P% за прагове от 1 до **к**, за всеки от които се намира отношението на броя верни от всички подструктури в обучителната извадка, с брой верни съседи по-голям или равен на този праг. Този метод е използван при всички направени от нас тестове, тъй като най-коректно отразява смисъла на вероятността като площ, подобно на подхода приложен за извличане на функцията на надеждност с логистична регресия и сортиращ критерий.

За извличане на кNN вероятностна функция ние използвахме Евклидово разстояние с признаците, които показаха най-висока разграничаваща способност. Това са основно подструктурните параметри с високо отношение на Фишер между класа на верните и неверните подструктури в ОИ и ТИ извадки.

Евклидовото разстояние бе изчислено с автоскалираните параметри. Броят на съседите **к** е оптимизиран в интервала от 1 до 40 за подструктурите от обучителната извадка, за които са намерени разстоянията до всички обети от същата извадка. Оптималното **к** е избрано съобразно характеристиките на постигнатата класификация, чувствителност и прецизност при 99.0% прагова вероятност за обучителната извадка. Извлечени са функции на надеждност с два набора от подструктурни параметри – шестте основни sNA, sLO, sSO, sFV, RMSD, MaxDev, както и седем, съдържащи освен изброените хистограмната променлива sIHS. За двата набора бе избрано оптимално **к**=22 и **к**=20, съответно. Въпреки по-високата чувствителност, отчетена при **к** по-малко от 20, посочените оптимални стойности бяха избрани с оглед стабилност на функцията по отношение на останалите две извадки от подструктури. При **к** по-малко от 20 за тях не бе постигната праговата прецизност от 99%.

Резултатите за класификационната способност на метода, получени за подструктурите от ТИ и ВИ при оптималното **к**, с табулирана вероятност, извлечена от обучителната извадка, показват по-ниска ефективност, в сравнение с ефективността на предложения по-рано сортиращ критерий 5 (Таблица 21). Отчетената прецизност от около 97% с кNN метода, е по-ниска от праговата 99.0%, показател, че вероятностната функцията не работи добре.

С цел подобряване на ефективността на функцията ние тествахме разновидност на метода, при която бива зададена гранична стойност за разстоянието. Обектите на разстояние по-голямо от граничното не се включват в изчисляването на вероятността, което може да се разглежда като кNN с променливо **к** за всеки класифициран обект. Граничната стойност за

разстоянието бе оптимизирана по резултатите от тестващата извадка за оптималното **к**, определено с обучителната извадка. Оценената прецизност надвиши праговата, за разлика от получената с кNN метода без отчитане на разстоянието, а също и прецизността, получена със сортиращия критерий К5 при относително постоянна чувствителност от около 23%.

Евклидовото разстояние, използвано за сравняване на подструктурата с всички подструктури от обучителната извадка на практика третира еднакво разликите за всеки признак, поради предварителното им скалиране. Подходящ начин за претегляне на разликите би било използването на отношенията на Фишер (Таблица 14) за признаците, участващи в сумата.

Друг подход, който приложихме, бе задаване на граница за стойността на всеки признак преди изчисляване на разстоянието, еквивалентен на намиране на най-близките съседи, но в област около подструктурата в пространството на скалираните признаци, описана от стените на хиперкуб с дължина на стената 2SD.

Тестване на функциите с библиотека PHYCHEM

Получените резултати за прецизност и чувствителност на метода с ТИ и ВИ за библиотека PHYCHEM (Таблица 27) показаха стабилност на функцията по отношение на подструктури, които не са използвани при нейното съставяне. Отчетената чувствителност е между 22% и 24% за 99% прагова надеждност, която е малко по-ниска в сравнение с постигнатата със сортиращ критерий, но значително по-висока от получената с логистична регресия.

По отношение на предложения по-рано сортиращ критерий, съставен за библиотеката PHYCHEM, кNN методът показва прецизност за тестващата и валидираща извадки по-висока от тази за най-добрите сортиращи критерии 2 и 5, чиито прецизности освен това са под праговата.

Таблица 27 Прецизност и чувствителност за праг 99% надеждност, постигнати с различни видовете кNN функции за а) ТИ и б) ВИ с библиотека PHYCHEM. В скоби е посочен броят на използваните подструктурни параметри.

а) **метод**	**k**	**dE***	**P%99**	**H99**	**H99.9**	**R%99**	**E_{max}**
Основен (7)	20	–	97.3	39	0	23.8	5.2
Основен (6)	22	–	97.4	35	0	22.7	4.2
dE (7)	20	2-3	98.9	15	0	23.6	5.2
dE (6)	22	2-3	99.2	11	0	22.7	4.2
Фишер (7)	38	–	99.8	2	0	17.9	1.2
Фишер (6)	38	–	99.8	2	0	18.0	1.2
Хиперкуб (7)	22	–	99.0	13	0	21.7	1.9
Критерий 2			**98.9**	**16**	**0**	**24.9**	
Критерий 5			**98.6**	**22**	**0**	**25.9**	

б) **метод**	**k**	**dE***	**P%99**	**H99**	**H99.9**	**R%99**	**E_{max}**
Основен (7)	20	–	99.0	9	0	24.0	6.6
Основен (6)	22	–	99.3	6	0	22.1	4.0
dE (7)	20	2-3	99.0	9	0	23.9	6.6
dE (6)	22	2-3	99.3	6	0	22.1	4.0
Фишер (7)	38	–	99.4	4	0	18.2	1.0
Фишер (6)	38	–	99.4	4	0	18.2	1.0
Хиперкуб (7)	22	–	99.1	7	0	21.3	2.0
Критерий 2			**98.6**	**14**	**0**	**25.9**	
Критерий 5			**98.6**	**14**	**0**	**26.9**	

к – *брой подструктури, най-близки до оценяваната*; **dE*** – *граница за Евклидовото разстояние (за методите без ограничение по разстоянието* **dE** *е равно на* $\mathbf{E_{max}}$); $\mathbf{E_{max}}$ – *максимално Евклидово разстояние в извадката*; **P%** и **R%** *прецизност и чувствителност за дадена прагова надеждност;* **H99** и **H99.9** *– брой неверни подструктури с надеждност над 99.0 и 99.9%.*

При всички тествани кNN-функции няма неверни подструктури, оценени с надеждност по-голяма или равна на 99.9%. В съответствие с данните от ОИ, най-ниска чувствителност, но най-висока прецизност над праговата са отчетени с параметрите, скалирани с отношението на Фишер.

Тестване на кNN функциите с библиотеки LAST и LAST1000

Приложимостта на кNN метода за оценка на резултатите от интерпретационно търсене бе потвърдена с другите две библиотеки с различен състав – LAST и LAST1000. Резултатите за работата на вероятностните функции за тези библиотеки са представени в таблица 28.

От таблицата се вижда, че оптималното **к** и за двете библиотеки е по-ниско от това, намерено за библиотеката от фитосъединения. Същевременно чувствителността е по-висока, достигаща 37% за библиотеката от 1000 органични съединения.

Таблица 28 Прецизност и чувствителност в проценти, постигнати с два вида кNN функции (със и без ограничение за разстоянието, dE) с шест подструктурни параметъра за а) ТИ и б) ВИ на библиотеки LAST1000 и LAST.

а) **метод**	**k**	**dE**	**P%99**	**H99**	**H99.9**	**R%99**	**E_{max}**
LAST1000	15	0	99.2	13	0	37.3	3.9
LAST1000 dE	15	0.2	99.3	7	0	21.9	3.9
LAST	16	0	99.2	127	0	31.3	4.0
LAST dE	16	0.2	99.2	78	0	18.4	4.0
LAST ANN*			**99.1**	**58**	**0**	**11.5**	

б) **метод**	**k**	**dE**	**P%99**	**H99**	**H99.9**	**R%99**	**E_{max}**
LAST1000	15	0	99.5	4	0	36.3	3.6
LAST1000 dE	15	0.2	99.6	2	0	23.3	3.6
LAST	16	0	98.2	162	0	29.9	3.9
LAST dE	16	0.2	99.2	42	0	17.8	3.9
LAST ANN*			**99.4**	**26**	**0**	**14.4**	

**за сравнение в таблицата са представени ефективностите на работа на оригиналната функция (ANN) на надеждност на библиотека* LAST, но за *търсене на съединенията от ТИ и ВИ в нея.*

Въпреки понижаването на чувствителността при метода с ограничение по разстоянието, тя е достатъчно висока, за да осигури значителен брой верни подструктури, предсказани с висока надеждност, като се има предвид общият брой верни подструктури 13 508, извлечени от библиотеката при търсене на стоте съединения от валидиращата извадка в нея.

Съпоставянето на показателите прецизност и чувствителност при кNN и ANN вероятностните функции на двете извадки за библиотека LAST показва съизмерима ефективност с известно преимущество на кNN метода. Вероятно това се дължи на разликата в типа на съединенията, използвани за обучаване на двете функции, а именно специфичното обучение на кNN функцията с извадка от природни съединения, каквито са тестващите извадки, докато оптимизацията на невронната мрежа е направена с органични съединения от общ тип.

В сравнение с изкуствените невронни мрежи, чиято оптимизация зависи от множество променливи, като брой слоеве, брой неврони във всеки слой и тяхната свързаност, активираща функция и др., допълнително предимство на кNN метода е конкретната последователност от предписания за извличане на функция на надеждност, които могат да бъдат обобщени по следния начин:

– *Избор на подструктурни параметри с висока разграничаваща способност между класа на верните и неверните подструктури.*

– *Оптимизиране на броя съседи **к** в широк интервал, например от 1 до 40, според критериите за ефективност – чувствителност и прецизност, с обучителната извадка.*

– *Оптимизиране на границата за разстоянието в интервала от нула до максималното за обучителната извадка, ако е необходимо за повишаване на прецизността.*

– *Валидиране на функцията с извадка/и от съединения, които не са използвани при обучението.*

Въпреки това, а също и малко по-добрата чувствителност, постигната с кNN метода спрямо оригиналната функция на надеждност на библиотека LAST,

един от недостатъците на предложения модел е дискретната природа на функцията. Броят на стойностите на надеждността, които могат да бъдат приписвани на подструктурите, зависи от оптималното к, така например функцията, съставена за библиотека LAST1000, ще заема само 15 фиксирани стойности със стъпка на промяна, съответстваща на 1/к или 7% приблизително.

4. Описание на резултатите от търсене, оценени с висока надеждност

За оценка на ИС на фрагментите може да се използва фактор на редукция n/N в части от единицата или проценти, където n е броят на структурите, съдържащи даден фрагмент и съответстващи на дадена молекулна формула, а N общият брои структурни изомери за тази формула [35]. Намирането на всички възможни изомери при така дефинираното отношение е затруднено за сложни съединения с голяма молекулна маса, ето защо бе потърсен алтернативен начин за оценка на информативността на подструктурите, извлечени с метода на интерпретационно библиотечно търсене.

Мярка, която отчита размера (sNA), степента на ненаситеност (sUE) и броя на хетероатомите (sHA) на генерираната подструктура, като относителни части спрямо структурата на потърсеното съединение, би могла да послужи за сравняване на ефективността на различните функции на надеждност.

$$IC = \frac{sNA}{uNA} + \frac{sUE}{uUE} + \frac{sHA}{uHA} \quad (14)$$

Така дефинираният коефициент IC може да заема стойности от 0 до 3. Ако подструктурата не съдържа хетероатоми или сложни връзки, стойност на IC 0.5 би означавала, че 50% от структурата на неизвестното съединение е разкрита.

Най-голям брой от подструктурите, оценени с висока надеждност от оригиналната функция на библиотека LAST имат коефициент на информационно съдържание между 0.2 и 0.4, подобно на резултатите получени при търсенето във PHYCHEM, но в някои от случаите, за разлика от PHYCHEM, подструктури с по-високо информационно съдържание над 2.5 са оценени като надеждни.

Информационното съдържание на подструктурите, оценени с висока надеждност от функцията, извлечена по метода на най-близките съседи за трите библиотеки, е представено на таблица 30.

Данните показват, че с kNN метода се идентифицират верни подструктури с по-високо информационно съдържание от средното за извадката и средното за верните в извадката, за разлика от функцията на надеждност, извлечена от сортиращ критерий. Освен това средното информационно съдържание на подструктурите, оценени с kNN-функцията на надеждност, е по-високо от полученото с функцията, извлечена от сортиращ критерий за библиотека PHYCHEM и оригиналната функция на библиотека LAST.

Таблица 30 Средно информационно съдържание на подструктурите, оценени с надеждност по-висока или равна на 99.0% от вероятностната функция, извлечена по метода на най-близките съседи с ограничение за разстоянието и 6 подструктурни параметъра. В колона ANN са посочени резултатите, получени с оригиналната функция на надеждност на библиотека LAST.

	PHYCHEM		**LAST1000**		**LAST**		**ANN**
извадка	**ср. в.**	**ср. 99**	**ср. в.**	**ср.99**	**ср. в.**	**ср.99**	**ср.99**
ТИ	0.49	**0.55**	0.54	**0.57**	0.50	**0.71**	**0.50**
ВИ	0.47	**0.48**	0.60	**0.79**	0.49	**0.67**	**0.49**

ср. в. - *средно на всички верни подструктури в извадката*.; **ср.99** – *средно на подструктурите с надеждност по-голяма или равна на* 99.0%.

Обяснението за получените резултати вероятно е свързано със същността на метода на най-близките съседи, при който величината от която се извлича функцията (броят на верните от всички к най-близки съседи) отразява зависимостта между подструктурните параметри и целевата променлива, но само за групи от подструктури в извадката, съдържащи подструктури с високо подобие помежду си. В тях относителния дял на верните е висок, когато се оценява вярна подструктура и обратно, за оценка на невярна е висок делът на неверните. Така, при оценката на надеждността на дадена високоинформативна подструктура, ако в обучителната извадка има близки до нея и следователно високоинформативни подструктури, които са верни, тя ще бъде оценена като надеждна. При извличането на вероятностна функция със сортиращ критерий, основният фактор влияещ върху надеждността е разграничаващата способност на критерия, която е само индиректно свързана с размера на подструктурите, посредством параметъра sNA (броя тежки атоми в подструктурата). Така при оценката на надеждността на подструктурата, единствено значение ще имат разпределенията на верните и неверните подструктури в извадката, при стойност на критерия равна или по-висока от изчислената за тази подструктура. За вярна подструктура с високо информационно съдържание, но недостатъчно висок сортиращ критерий, няма да бъде приписана висока надеждност дори при наличие на други близки до нея, високоинформативни, верни подструктури в извадката.

5. Търсене по подобие и идентичност

Един от основните проблеми при търсене за идентификация и по подобие в спектрални библиотеки е вариацията на сигналите в резултат на разликите в експерименталните условия на измерване, от които най-голямо влияние оказва видът на разтворителя. Внимателното оптимизиране на толерансите на съвпадение на химичните отмествания може да подобри ефективността на търсене при директно прилагане на спектрални мерки за подобие.

За търсене в библиотека PHYCHEM, бяха направени тестове с мярка за подобие (S_{sim}), отчитаща броя съвпаднали сигнали, при определен толеранс на съвпадение в ppm, нормиран спрямо сумата от сигналите в двата спектъра.

$$S_{sim} = \frac{2K}{M+N}$$, където (15)

K *е броят съвпаднали сигнали от двата спектъра при определен толеранс в ppm;*

M и N *са броят на сигналите в единия и другия спектър.*

Тази мярка е еквивалентна на Sorensen-Dice индекс на подобие за бинарни вектори и много близка до Танимото (Tanimoto или още Jaccard) индекса [68]. Способността на мярката да разпознава структурно еквивалентни съединения по техните спектри е тествана със съединения, чиито спектри са регистрирани в различен разтворител и стереоизомери.

Таблица31 Ефективност на мярката за спектрално подобие при различни толеранси на съвпадение на сигналите в ppm, за три нейни прагови стойности.

	праг 0.90				**праг 0.95**				**праг 0.99**			
Тол.	**б.д.е.**	**б.д.из.**	**о.б.д.**	**%**	**б.д.е.**	**б.д.из.**	**о.б.д.**	**%**	**б.д.е.**	**б.д.из.**	**о.б.д.**	**%**
0	0	0	5	**0**	0	0	3	**0**	0	0	1	**0**
0.5	1	7	20	**40**	1	3	7	**57**	1	2	4	**75**
1	4	18	54	**41**	2	10	15	**80**	2	9	12	**92**
1.5	6	27	90	**37**	5	16	27	**78**	4	12	17	**94**
2	9	30	126	**31**	7	22	51	**57**	5	14	25	**76**
2.4	10	35	175	**26**	7	27	69	**49**	5	16	30	**70**
3	10	40	215	**23**	8	29	80	**46**	5	21	37	**70**
3.5	10	44	251	**22**	9	36	101	**45**	5	26	43	**72**
4	10	47	290	**20**	10	39	126	**39**	8	27	48	**73**
4.5	10	49	335	**18**	10	43	144	**37**	8	30	53	**72**
5	10	49	370	**16**	10	47	165	**35**	8	36	60	**73**
6	10	50	452	**13**	10	48	189	**31**	8	37	66	**68**
7	**10**	**51**	**553**	**11**	10	50	222	**27**	8	38	73	**63**
8	10	51	649	**9**	**10**	**51**	**247**	**25**	8	39	78	**60**
9	10	51	741	**8**	10	51	282	**22**	**8**	**41**	**86**	**57**
10	10	51	824	**7**	10	51	313	**19**	8	41	89	**55**
15	10	51	1333	**5**	10	51	422	**14**	8	41	115	**43**
20	10	51	1898	**3**	10	51	504	**12**	8	41	124	**40**

б.д.е. *– брой на двойките спектри на едно и също съединение);* **б.д.из***. – брой двойки спектри на стереоизомери;* **о.б.д.** *– общ брой на двойките спектри с подобие над дадения праг;* **%** *– относителния дял на двойките еквиваленти съединения и стереоизомери, общо, спрямо броя на всички двойки с подобие над съответния праг.*

Присъстващите в библиотека PHYCHEM двойки спектри на едно и също съединение са 10, а съответно двойките стереоизомери – 39, от които две групи съдържат по 3 и 4 стереоизомера – общо 51 двойки, включващи всички комбинации. Резултати, представени в таблица 31 показват, че с увеличаване на толеранса на съвпадение на сигналите в спектралната мярка за подобие, все по-голям брой от еквивалентните структури биват разпознати, но същевременно общият брой на спектрите с висока степен на подобие също нараства. Практически оптималното съотношение на параметрите на търсене, толеранс и праг на подобието, гарантиращ разпознаване на дадено съединение, ако то присъства в библиотеката (търсене по идентичност), е средно висок праг за подобие 0.95 и толеранс 7-8 ppm, малко по-широк от типичния, приписван на влиянието на разтворителя – няколко ppm.

С Танимото мярката за спектрално подобие, разпознаване на всички двойки съединения от двете групи бе постигнато само при праг 0.90 и толеранс по-висок или равен на 8 ppm при относителен дял от 24% спрямо общият брой двойки. При прагове за подобието 0.95 и 0.99 само 9 и 8 еквивалентни структури са намерени, както и 47 и 42 стереоизомера, относително 46% и 57% от всички двойки с подобие над зададения праг, съответно.

Спектро-структурно подобие

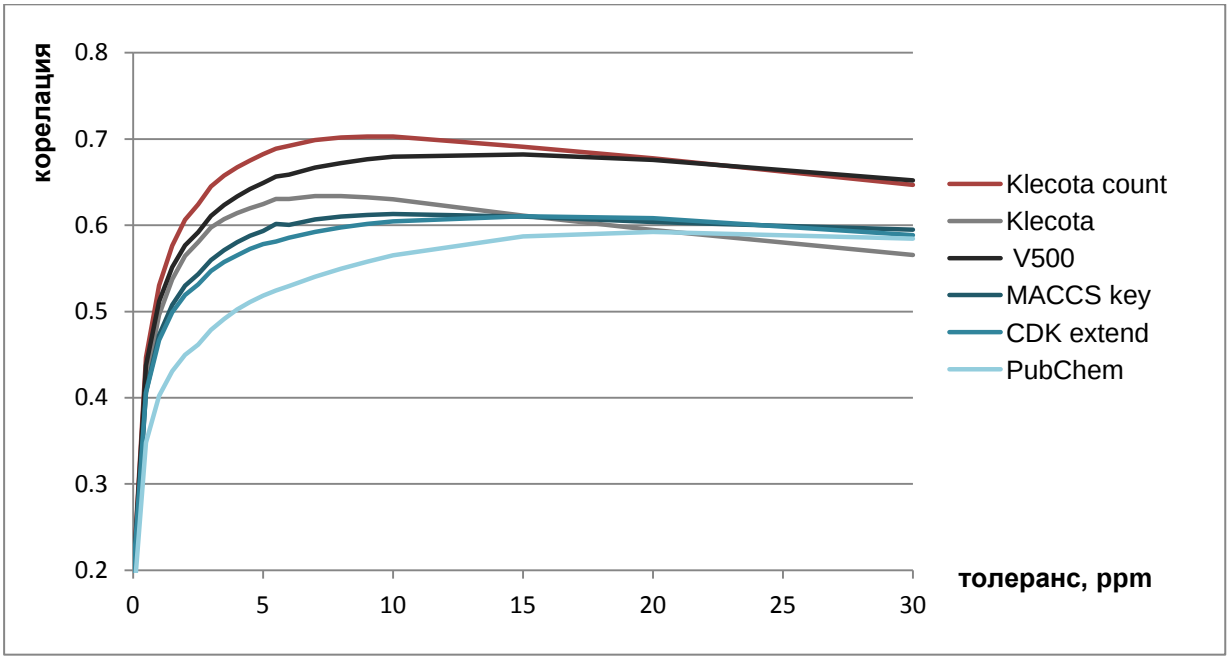
За оценка на способността на спектралното търсене да отсява структурно подобни съединения, което е пряко свързано с приложимостта на този метод за търсене по подобие, използвахме корелацията между спектралното и структурното подобие за всички двойки съединения в библиотеката.

За дефиниране на структурно подобие между двойките, структурите бяха представени в числов вид с помощта на стандартни набори от подструктурни дескриптори, известни още като структурни фингърпринти. Три фингърпринта Klekota-Roth (4860 дескриптора) [192], CDK extended (1024 дескриптора) и PubChem (880дескриптора) бяха избрани от поддържаните в програмата PaDel-Descriptor. Използван е и набор от 500 структурни дескриптора V500 с общо предназначение, които са предоставени от професор Varmuza [193].

Типът на един от фингърпринтите Klekota-Roth count, отчитащ броя на структурните фрагменти, присъстващи в структурата, освен това наложи използването на универсална мярка, боравеща с целочислени, а не само с бинарни вектори. В настоящото изследване е използвано разстояние по Джакард (Jaccard distance, пакета Vegan на R), уравнение 16.

$$D_{Jaccard} = \frac{2d}{1+d} \qquad d = \frac{\sum_i |x_i - y_i|}{\sum_i (x_i + y_i)} \tag{16}$$

Изборът на структурно представяне е направен според корелацията между спектралното и структурното подобие за всички двойки съединения в библиотеката. На фигура 25 е представена корелацията между структурното подобие, определено като единица минус разстоянието по Джакард и спектралното подобие за различни толеранси на съвпадение на сигналите. Оптималния толеранс за търсене, определен според този критерий, е между 7 и 10 ppm за структурния фингърпринт с най-висока корелация *Klekota-Roth count*.

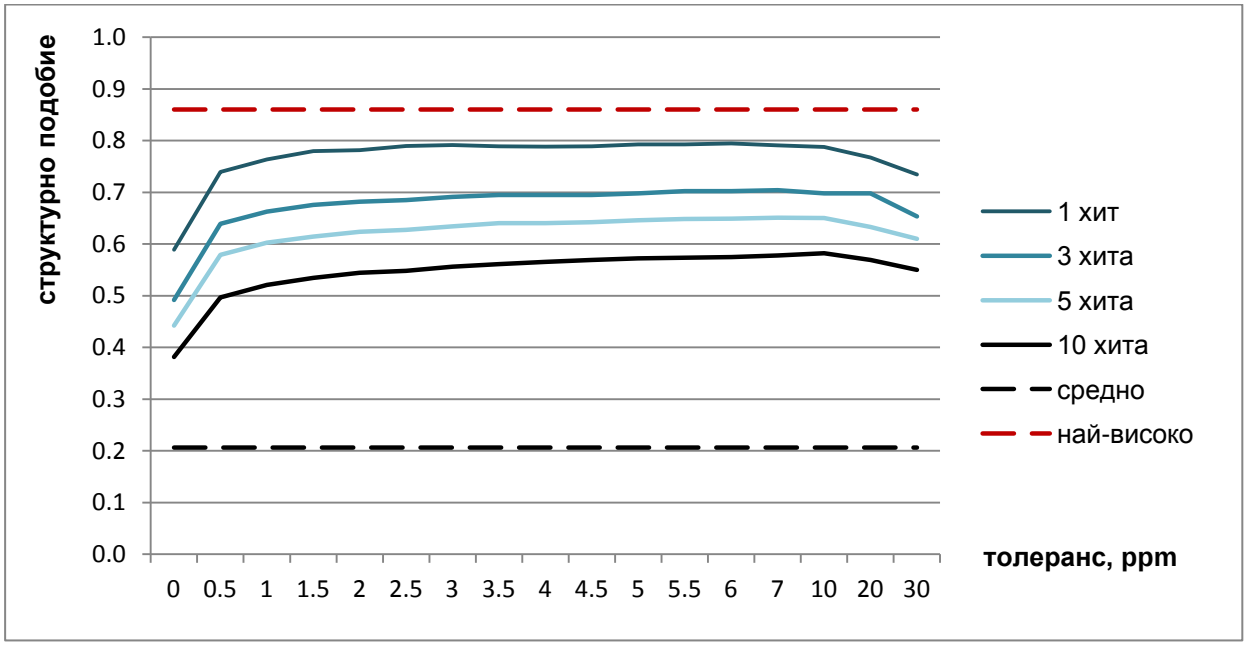


Фигура 25 Корелация между спектралното и структурното подобие (1-разстоянието по Джакард) за всички двойки съединения в библиотеката при различни толеранси на съвпадение на сигналите и пет структурни представяния.

В действителност ефективността на библиотечното търсене по подобие е пряко свързана със способността за разпознаване само на най-близките по-спектър и структура до търсеното съединения в библиотеката. Корелацията между средното спектрално и структурно подобие за първите 10 хита е по-ниско от корелацията за първия, първите три и първите пет хита в изследвания интервал на толеранса, което означава, че за тази библиотека най-информативни са хитовете на първите няколко позиции в хитсписъка. За разлика от кривата с всички двойки (Фигура 25), кривите за хитовете в началото на хитсписъка достигат максимум при толеранс 4 ppm – от 0.75 за първи хит до 0.7 за първите десет.

Реалната ефективност на метода за дадена библиотека освен това пряко зависи от нейния състав и разнообразие, т.е. от степента на подобие на най-близките структури и спектри в нея, ето защо по-коректен начин за оценка на работата на метода е по отношение на средното и най-високото подобие в библиотеката.

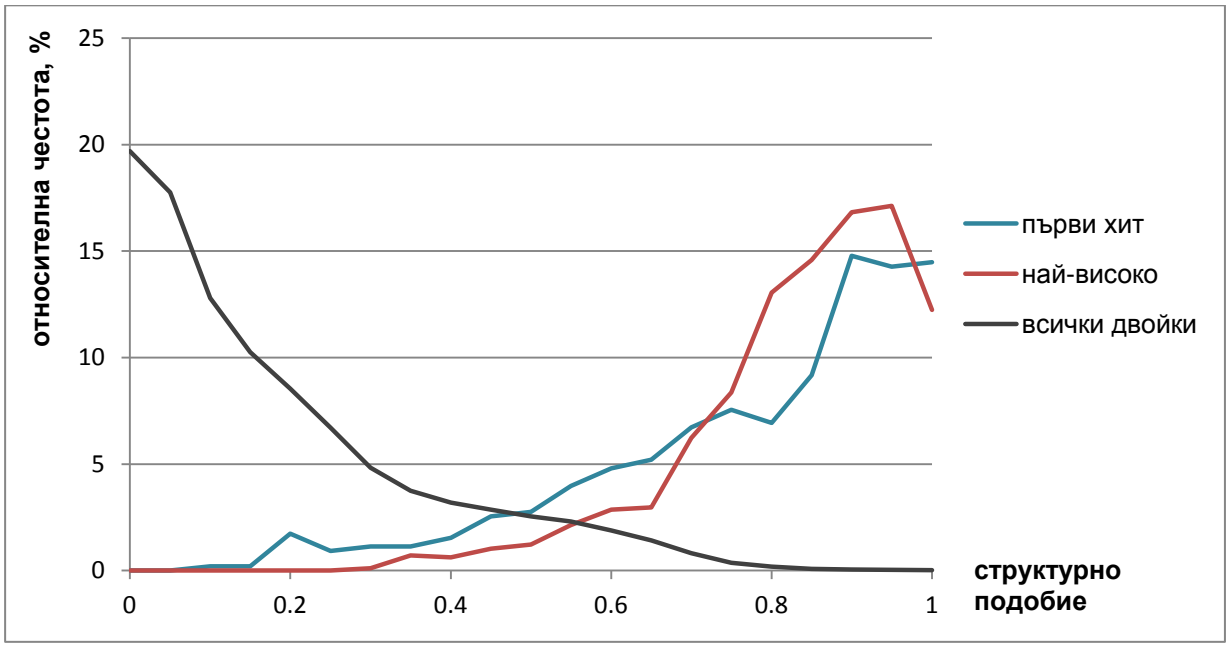
Средното структурно подобие на библиотеката може да се дефинира като средна стойност на структурното подобие за всички двойки съединения в нея, а най-високото като средно на подобието за всяко съединение в библиотеката и неговото най-подобно по структура. Структурното подобие за първите няколко хита показва достигане на постоянна стойност при толеранс около 3 ppm и плавно понижаване след 10 ppm (Фигура 27).



Фигура 27 Средно структурно подобие за всяко от съединенията в библиотеката и неговият първи хит или първите 3, 5 и 10 хита, получени при търсене в нея при различни толеранси на съвпадение на сигналите и Klekota-Roth count структурен фингърпринт. Средно структурно подобие и най-високо структурно подобие – с пунктир.

По-детайлна картина на резултатите от търсене по подобие за фиксиран толеранс, може да бъде получена от хистограмите на разпределенията на подобията за всички двойки структури в библиотеката – средно подобие за библиотеката, двойките подобия с най-подобната структура от библиотеката за всяко съединение в нея – най-високо подобие в библиотеката и подобията на двойките структура на съединението и на първия хит, получен при спектрално търсене в библиотеката за всяко от съединенията в нея (Фигура 28).

От фигурата се вижда, че нормираните разпределения на подобията за първия хит и на най-близките структури в библиотеката почти напълно се припокриват, същевременно са отместени от максимума на разпределението на структурното подобие за всички двойки съединения в библиотеката. Това означава, че спектралното търсене по подобие с избраната мярка успява да извлече от библиотеката структури, подобни на търсената, за разлика от случайния избор. Този подход е заимстван от Varmuza и съавтори [194].



Фигура 28 Хистограма на разпределенията на структурните подобия за всяко съединение в библиотеката PHYCHEM и неговия първи хит, получен при спектрално търсене по подобие в нея при толеранс 5 ppm, спрямо подобията за всички двойки и най-високите структурни подобия за библиотеката с *Klecota-Roth count* фингърпринт структурно представяне.

Една възможна мярка за количествена оценка на ефективността на търсене отчита подобието на структурите, сортирани в началото на хитсписъка и параметрите най-високо и средно подобие за библиотеката:

$$SE = \frac{S_h - S_l}{S_b - S_l}$$, където (17)

S_b – най-високото структурно подобие за библиотеката, определено като медиана или средна стойност на най-високото подобие за всяко съединение от библиотеката и неговото най-подобно в нея от всички двойки без повторения.

S_l – средното структурно подобие, определено като медиана или средна стойност на подобията за всички двойки съединения в библиотеката без повторения.

S_h – структурно подобие на първия хит, определено като медиана или средна стойност на подобието на структурите, сортирани на първо място в хитсписъка и потърсените съединения при спектрално търсене по подобие на всяко съединение в библиотеката в самата нея.

Дефинирана по този начин, мярката не отчита дисперсията на оценяваните подобия. Така висока ефективност на търсенето може да бъде приписана, дори когато разпределенията на подобията с първия хит и най-

високите подобия не се припокриват, т.е. най-подобните структури не са намерени. Предложената мярка може да бъде разширена с допълнителен коефициент *f*, който отчита площите на припокриване между двойките разпределения.

$$SE_f = \frac{S_h - S_l}{S_b - S_l} f, \qquad f = \frac{a+c}{a+b} \tag{18}$$

a – площта на припокриване между разпределенията на първия хит и най-високите подобия;

b – площта на припокриване между разпределенията на първия хит и всички двойки в библиотеката;

c – площта на припокриване между разпределенията на най-високите подобия и всички двойки в библиотеката

В таблица 32 са посочени стойностите на предложените мерки за оценка на ефективността на търсене по подобие при три толеранса. Параметрите са определени от медианите на структурното подобие за съответните двойки структури.

Таблица 32 Ефективност на търсенето по подобие в библиотека PHYCHEM, оценено с SE и SE_f за структурно подобие, дефинирано като единица минус разстоянието по Джакард между структурите, представени с Klecota-Roth count фингърпринт при SS мярка за спектрално подобие с толеранси 4, 5 и 6 ppm.

разпределение	медиана	SE	*f*	SE_f
Първи хит 4 ppm	0.85	0.94	0.89	0.84
Първи хит 5 ppm	0.87	0.97	0.91	0.88
Първи хит 6 ppm	0.85	0.95	0.90	0.85
Най-добро подобие	0.89			
Всички двойки за библ.	0.15			

От таблицата се вижда, че за дадената библиотека двете мерки дават съгласувана оценка при различни толеранси, но SE_f има по-ниски стойности. И в двата случая търсенето при толеранс от 5 ppm показва най-добри резултати.

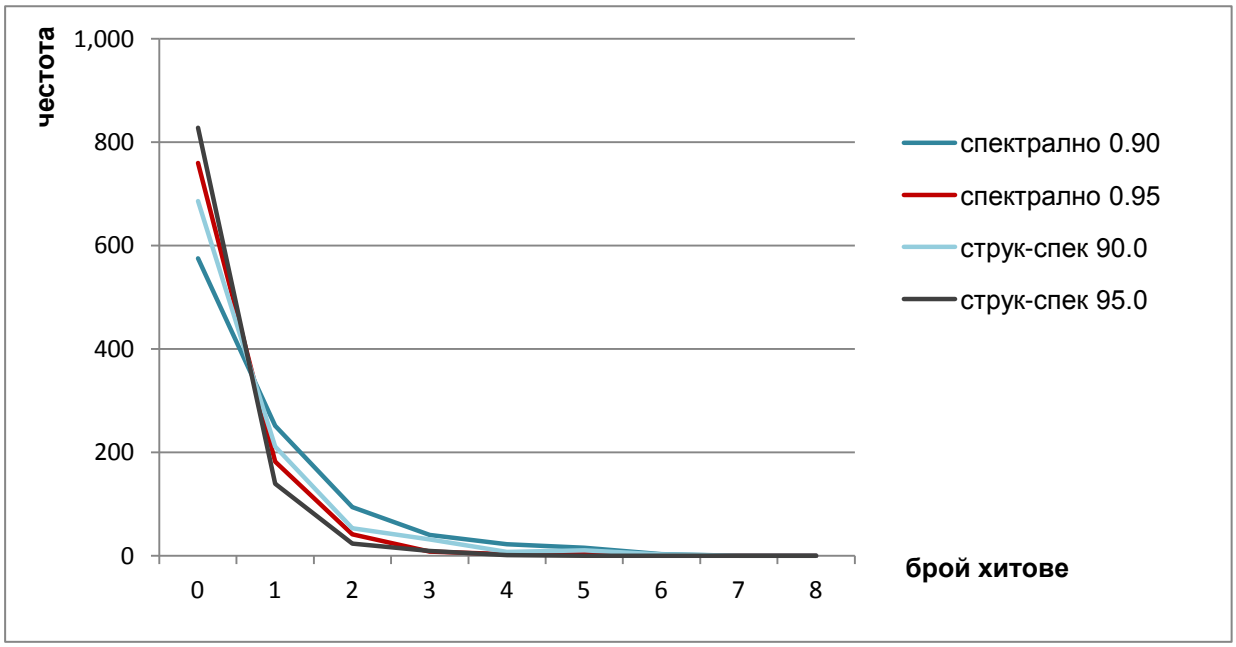
Предложената мярка за ефективност SE_f може да бъде използвана за сравняване на търсене по подобие в различни библиотеки. За тестване на търсенето по подобие на фитосъеднинения в неспециализирана библиотека от органични съединения, библиотека PHYCHEM бе потърсена в LAST по описания начин. SE_f бе изчислена с медианите на разпределенията на първия хит, най-високите 1000 структурни подобия на потърсените съединения с референтни съединения и подобията на всички двойки съединения (Таблица 33).

Таблица 33 Ефективност на търсенето по подобие на съединенията от библиотека PHYCHEM в библиотека LAST, оценено с SE и SE_f за структурно подобие, дефинирано като единица минус разстоянието по Джакард между структурите, представени с Klecota-Roth count фингърпринт при SS мярка за спектрално подобие с толеранси от 5 до 10 ppm.

разпределение	медиана	SE	*f*	SE_f
Първи хит 5 ppm	0.57	0.64	0.63	0.41
Първи хит 6 ppm	0.57	0.65	0.63	0.41
Първи хит 7 ppm	0.59	0.68	0.64	0.43
Първи хит 8 ppm	0.59	0.68	0.64	0.43
Първи хит 9 ppm	0.59	0.68	0.64	0.43
Първи хит 10 ppm	0.59	0.68	0.63	0.43
Най-високо подобие	0.81			
Всички двойки	0.13			

Подобна е наблюдаваната тенденция при структурно представяне с V500 фингърпринт. Максимум за SE_f 0.57 е достигнат при 7 ppm, което е значително по-високо от отчетеното с *Klecota-Roth count* фингърпринт, но по-ниско в сравнение с резултатите за търсене в библиотека PHYCHEM (Таблица 32).

Това показва предимството на специализираните библиотеки за целите на разкриване структурата на фитосъединения, макар проведените от нас тестове да са направени със сравнително малка по размер библиотека.



Фигура 30 Хистограми на разпределенията на броя хитове със структурно и спектрално подобие над прагове 0.90 и 0.95 за хитсписък, получени при спектрално търсене на всяко от съединенията в библиотеката с толеранс 5 ppm.

Въпреки ефективното прилагане на библиотечното търсене, близко до оптималното за библиотеката, практическата полза от библиотечно търсене по подобие зависи от нейния състав. От фигура 30 се вижда, че за голям брой от потърсените съединения, повече от половината, в хитсписъка не присъства структура с подобие по-голямо или равно на 0.90. Това означава, че търсенето по подобие ще бъде от полза само в част от случаите. Вероятно малкия размер на библиотеката ограничава нейното приложение.

На практика само за 227 от 1000 съединения във PHYCHEM е намерен поне един хит от библиотека LAST със спектрално подобие по-голямо или равно на 0.90, а за 90 от тях то е по-голямо или равно на 0.95. Въпреки значимия размер на библиотека LAST, този дял е съизмерим с търсенето във PHYCHEM (Фигура 30), а за по-голяма част от съединенията не е намерен референтен спектър с достатъчно високо подобие.

IV. ОБОБЩЕНИ РЕЗУЛТАТИ И ИЗВОДИ

1. Съставена е библиотека PHYCHEM от 1000 ^{13}C ЯМР спектъра на фитосъединения. Библиотека е претърсена за грешки ръчно и автоматично чрез предсказване на спектрите с адитивни схеми.

2. Оригиналната функцията на надеждност на библиотека LAST работи коректно за търсене на природни съединения, но с понижена чувствителност – средно 13%, спрямо 32%, отчетени при нейното валидиране с органични съединения от общ тип за праг на надеждността 99% .

3. В общия случай функция на надеждност, създадена за една библиотека не е преносима за друга.

4. Подходящи числови мерки за подбор на подструктурни параметри с висока разграничаваща способност между верните и неверните подструктури са (1) рангово бисериалния корелационен коефициент и (2) отношението на Фишер на рангово скалираните параметри. Двата критерия дават съгласувана оценка за разграничаващата способност на параметрите, предназначени за съставяне на функция на надеждност.

5. Функция на надеждност може да бъде извлечена с логистична регресия, но за сравнително малка библиотека, каквато е PHYCHEM. Това ограничение допълнително е свързано с изискването за използване назначителни изчислителни ресурси и мощен софтуер. За библиотека PHYCHEM, бяха постигнати чувствителности в интервала 9% -13% за 99% прагова надеждност.

6. Функция на надеждност може да бъде извлечена от сортиращ критерий, съставен от подструктурните параметри с най-висока разграничаваща способност, но ефективна такава не бе получена за библиотеки LAST и LAST1000. За библиотека PHYCHEM най-високата постигната чувствителност е 26-27% и прецизност 98.6 за праг 99%.

7. Методът на най-близките съседи осигурява универсален и относително лесен начин за извличане на функция на надеждност, посредством процедура от ясно дефинирани последователни стъпки. Най-високи прецизности са получени с вариант на метода, при който се задава ограничение за разстоянието до най-близките съседи. Съставени са ефективни функции на надеждност и за трите библиотеки PHYCHEM, LAST и LAST1000 с прецизности над праговите, и чувствителности съответно средно за ТИ и ВИ 22% с PHYCHEM, 37% с LAST1000 и 18% с LAST.

8. Предложен е критерий за оценка на информационното съдържание на подструктурите, резултат от интерпретационното библиотечно търсене. Функцията на надеждност, извлечена по метода на най-близките съседи, оценява с висока надеждност подструктури с по-високо информационно съдържание от средното за верните в извадката, както и по-високо в сравнение с функцията, извлечена от сортиращ критерий за библиотека PHYCHEM и оригиналната функция на библиотека LAST.

9. Високото съответствие между спектралното подобие и структурното подобие, изчислено от структурните фингърпринти *Klekota-Roth count* и *V500* с мярката Джакард индекс, показва, че тези структурни пресдставяния могат да бъдат използвани при оптимизиране на методите търсене в библиотека от природни съединения.

10. Мярката за спектрално подобие, базирана на Sorensen-Dice индекс, е подходяща за търсене по идентичност в библиотека от фитосъединения. Задаване на праг за подобие 0.95 при толеранс 7-8 ppm, осигурява разпознаването на идентични съединения, чиито спектри са снети в различен разтворител, както и на стереоизомери.

11. Предложената мярка за ефективност на търсене по подобие, позволява сравняването на различни методи за търсене и различни библиотеки, според способността за намиране на най-подобните по-структура съединения. Най-висока ефективност за търсене по подобие на фитосъединения бе отчетена при толеранс 5 ppm за библиотека PHYCHEM и 7 ppm за LAST със Sorensen-Dice мярка за спектрално подобие.

Методът на интерпретационно търсене в библиотека от природни съединения PHYCHEM може да бъде прилаган успешно в процеса на разкриване структурата на съединения със сложен строеж. За подобряване на ефективността на търсене е необходимо разширяване състава на библиотеката с по-голямо разнообразие от съединения и оптимизиране на вероятностната функция с по-голям брой тестови подструктури.

V. ПРИНОСИ

1. Съставена е нова библиотека PHYCHEM от 1000 ^{13}C ЯМР спектъра на фитосъединения, която се поддържа от програмата INFERCNMR.

2. Съставена е вероятностна функция за оценка на надеждността на подструктурите, извлечени при интерпретационно търсене на природни съединения в библиотека PHYCHEM.

3. Предложена е процедура от конкретни и лесни за изпълнение стъпки с цел съставяне на вероятностна функция по метода на най-близките съседи, предназначена за оценка на надеждността на подструктурите, извлечени в резултат на интерпретационно търсене в библиотеки с различен състав.

4. Предложена е мярка за оценка на информационното съдържание на подструктурите, извлечени в резултат на интерпретационно библиотечно търсене, която може да бъде използвана при сортиране на подструктурите с висока надеждност в хитсписъка или за сравняване на ефективността на работа на различни функции на надеждност.

5. Намерени са оптималните толеранси на съвпадение на сигналите за търсене по подобие и идентичност на фитосъединения – 5 и 7 ppm съответно за библиотека PHYCHEM и 7 ppm за търсене по подобие в библиотека LAST със спектрална мярка за подобие, базирана на Sorensen-Dice индекс.

6. Предложена е мярка за оценка на ефективността на търсене по подобие, която отчита подобието на получените структури в хитсписъка, спрямо средното и максимално структурно подобие на библиотеката.

VII. ИЗПОЛЗВАНА ЛИТЕРАТУРА

1 E. Fukushi, Advanced NMR approaches for a detailed structure analysis of natural products. *Biosci. Biotechnol. Biochem.*, 2006, 70 (8), 1803-1812.

2 Ch. Peng, Sh. Yuan, Ch. Zheng, Y. Hui, Efficient Application of 2D NMR Correlation Information in Computer-Assisted Structure Elucidation of Complex Natural Products. *J. Chem. Inf. Comput. Sci.,* 1994, 34 (4), 805-813.

4 M. Elyashberg, K. Blinov, S. Molodtsov, A. Williams, Elucidating "Undecipherable" Chemical Structures Using CASE Approaches, *Magn. Reson. Chem.*, 2012, 50 (1), 22–27.

35 H. Schriber, E. Pretsch, General Characteristics of Good-List and Bad-List Entries for Structure Generators from Spectra. *J. Chem. Inf. Comput. Sci*. 1997, 37 (5), 879-883.

68 L. Bodis, A. Ross, E. Pretsch, A novel spectra similarity measure, *Chemometr. Intell. Lab.*, 2007, 85, 1-8

86 P. Penchev, K.-P. Schulz M. Munk, INFERCNMR: A ^{13}C NMR Interpretive Library Search System. *J. Chem. Inf. Model*. 2012, 52, 1513−1528.

89 J. López-Pérez, R. Theron, E. del Olmo, B. Santos-Buitrago, J. Adserias, C. Estévez, C. Cuadrado, D. López, G. Santos-García, NAPROC-13: A Carbon NMR Web Database for the Structural Elucidation of Natural Products and Food Phytochemicals, series Advances in Intelligent Systems and Computing, 8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB) 2014, Volume 294.

179 П. Пенчев, Компютърна интерпретация на молекулни спектри с цел разкриване на структурата на органични съединения. Дисертация за придобиване на степен доктор на науките , Пловдив, 2016.

182 M. Elyashberg, A. Williams, K. Blinov, Methods of NMR Spectrum Prediction and Structure Verification. In: Contemporary Computer-Assisted Approaches to Molecular Structure Elucidation, Royal Society of Chemistry, Cambridge, 2012.

184 A. Yongye, J. Waddell, J. Medina-Franco, Molecular Scaffold Analysis of Natural Products Databases in the Public Domain, *Chem. Biol. Drug Des.,* 2012, 80(5), 717-724.

192 J. Klekota, F. Roth, Chemical substructures that enrich for biological activity, *Bioinformatics*, 2008, 24(21), 2518–2525.

193 H. Scsibrany, M. Karlovits, W. Demuth, F. Müller, K. Varmuza, Clustering and similarity of chemical structures represented by binary substructure descriptors, *Chemom. Intell. Lab.*, 2003, 67 (2), 95–108.

194 W. Demuth, M. Karlovits, K. Varmuza, Spectral similarity versus structural similarity: infrared spectroscopy, *Anal. Chim. Acta*, 2003, 490, 313–324.

ПУБЛИКАЦИИ ПО ТЕМАТА НА ДИСЕРТАЦИЯТА

S. Nachkova, S. Milenkova, P. Penchev, P. Bozov; Interpretive library search of plant compound spectra in a ^{13}C NMR database. *Chemistry of Natural Compounds*, 2015, 51(5), 993-996. IF0.473

S. Nachkova, S. Milenkova, P. Bozov, P. Penchev, Interpretive search in a ^{13}C-NMR spectral library of plant compounds. *Scientific Works: University of Ruse "Angel Kanchev"* , 2013, 52, 10(1), 47-51.

УЧАСТЕЯ В НАУЧНИ ФОРУМИ

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

S. Nachkova, S. Milenkova, P. Bozov, P. Penchev*; Interpretive search in a ^{13}C-NMR spectral library of plant compounds*. Scientific session of University of Ruse "Angel Kanchev", 25-26 October 2013, Razgrad, Bulgaria

S. Nachkova, S. Milenkova, P. Bozov, P. Penchev, *A ^{13}C NMR Interpretive Library search System*, V^{-th} International Conference of the Young Scientists – Plovdiv' 2013, Natural and Technical Sciences, Plovdiv, 13-16 June 2013.

БЛАГОДАРНОСТИ

Изказвам признателност на моя научен ръководител проф. Пламен Пенчев, както и на моите колеги за тяхната отзивчивост и подкрепа, за всичко на което ме научиха, за сърдечното отношение и търпение при съвместната ни работа, за които искрено благодаря.