Лекция 10

Спектрално и структурно подобие: използване в спектроскопията

10.1. Спектрални признаци. Спектралният признак е число, което се изчислява по определен алгоритъм от спектъра. Например в мас-спектрометрията на органични съединения „суровият" мас-спектър е много „некачествен" образ, когато трябва да се използва при сравнение със спектрите на други съединения. Най-главната причина за това е, че спектрите на хомолози (съединения с едни и същи функционални групи, които се различават в структурата си с една метиленова група, CH_2, с маса 14) дават различаващи се по маса фрагменти, и която маса е кратна на 14. Ето защо от мас-спектрите се изчисляват спектрални признаци, които се използват за класификация или сравнение на спектрите. Един такъв признак е т.н. „модуло 14", който е сума от йонните токове (интензитетите) при масови числа, различаващи се с 14 единици: $mod14(m,5) = I_m + I_{m+14} + I_{m+28} + I_{m+42} + I_{m+56}$. В ИЧ и Раман спектроскопията спектралните признаци са по-директно свързани със спектъра, например интензитет на пик в даден спектрален регион. Но като цяло спектралният признак не се отъждествява с числова характеристика, взета директно от спектъра.

10.2. Спектралната крива като вектор и образ. Спектрите се измерват при определена висока разделителна способност, изисквана от специфичността на научния проблем за който се използват. В компютърните колекции те се пазят във файлове (най-вече в текстов формат) също във висока разделителна способност. Но когато те се съберат в бази от данни, наречени спектрални библиотеки, те обикновено са с по-ниска разделителна способност. Например за ИЧ и Раман спектрите оптималният интервал между две данни на абсцисата (вълново число) е от 2 cm^{-1} до 4 cm^{-1}.

Целите спектрални криви се пазят като поредица от числа, показващи ординатата на спектралната крива. В повечето конмпютърни системи данните са еквидистантни по абсцисата. Например в ИЧ библиотеките на Sadtler се пазят стойностите на абсорбцията, взети през 4 cm^{-1} от 500 cm^{-1} до 3700 cm^{-1} [1]. По същият начин се пазят данните и в програмата IRIS [2]. Определено този набор от числа представлява един многомерен вектор – в дадения пример това е 801 мерен вектор: ширината на целия спектрален интервал е 3200 cm^{-1} (= 3700 cm^{-1} - 500 cm^{-1}) и той е разделен на 800 (= 3200 cm^{-1} / 4 cm^{-1}) интервала, които са ограничени от 801 стойности по абсцисата. Това че данните са еквидистантни по абсцисата няма никакво значение – важното е при компютърното сравнение на две спектрални криви стойност на абсорбцията в единия спектър при определено вълново число да се сравнява със стойност на абсорбцията в другия спектър при същото вълново число – вижте лекция 5 в частта за извадките от данни.

10.3 Сравняване на спектрални криви. Директният метод за сравняване на спектрални криви е изчисляване на разстоянието между тях в многомерното пространство. Разбира се има и други критерии, които ще бъдат засегнати в настоящата лекция.

Разстоянието между образите е най-естественият критерий и е интересно да разгледаме защо това е така. Ако използваме Евклидово разстояние, то под „близки обекти" се разбира „образи с малко разстояние между тях", и това математически изисква стойностите на съответните признаци да са близки в двата образа. Ако разстоянието между образите е голямо, то те се различават силно поне в един от признаците си, според формула (3.1) от лекция 3.

В спектроскопията се използват растояние в Манхатън и Евклидово растояние, но с цел изчисляване на мярка за подобие в някакъв интервал (обикновено 0 – 1000) те са леко изменени.

***Малко философия**: Силата на всички физични взамодействия в природата спада с разстоянието. Това води до редица последствия, но най-първото забелязаното от хората е, че в области, които са близки по разстояние, има подобни или еднакви обекти – подобни/еднакви животни, растения, релеф, климат и т.н., а отдалечените области са много различни по тези признаци.*

Освен разстоянието между спектралните криви или образи (наборите от спектрални признаци) друг критерий за подобие между тях е тяхната корелация и скаларното им произведение.

Най-използваните сравнения между ИЧ (а също и Раман) спектрални криви са следните: (1) *средно квадратично отклонение*, (2) *средно абсолютно отклонение*, (3) *скаларно произведение*, и (4) *коефициент на корелация*. Нека с A^U_k означим абсорбцията при k-тото вълново число в непознатия спектър, с A^R_k - тази в референтния,), а N е броят на разглежданите абсорбционни стойности. С съответните хит-лист качествени индекси (hit quality index, HQI), по които се сортират резултатите от търсенето в библиотеката, са пропорционални на:

- средното квадратичното отклонение между спектрите:

$$S_1 = \sqrt{\sum_k (A^U_k - A^R_k)^2 / N} \qquad (10.1)$$

- средното абсолютно отклонение между спектрите:

$$S_2 = (1/N)\sum_k \left|A^U_k - A^R_k\right| \qquad (10.2)$$

- скаларното произведение между спектрите:

$$S_3 = \frac{\sum_k A^U_k A^R_k}{\left|A^U\right| \cdot \left|A^R\right|} \qquad (10.3)$$

- коефициента на корелация между спектрите:

$$S_4 = \frac{\sum_k (A_k^U - \overline{A^U})(A_k^R - \overline{A^R})}{\sqrt{\sum_k (A_k^U - \overline{A^U})^2 * \sum_k (A_k^R - \overline{A^R})^2}}, \qquad (10.4)$$

Величините $|A_U|$ и $|A_R|$ в (10.3) са големините на векторите (спектрите), а средни стойности на спектралните признаци (координатите) на тези вектори са величините в (10.4), отбелязани с черти над тях.

$$\left|A^U\right| = \sqrt{\sum_k (A_k^U)^2} \quad \text{и} \quad \left|A^R\right| = \sqrt{\sum_k (A_k^R)^2}$$

$$\overline{A^U} = \frac{\sum_k (A_k^U)}{K} \quad \text{и} \quad \overline{A^R} = \frac{\sum_k (A_k^R)}{K}$$

Сумите в (10.1) – (10.4), както и последните четири израза са по **k**, което се променя от 1 до **K**; **K** = 801 в програмата **IRIS**. Мерките S_1 и S_2 показват различието между спектрите. Тяхната максимална теоретична стойност е 1.0 при условие, че спектрите са нормирани в интервала 0.0 - 1.0 а.е. Те заемат минималната си стойност 0.0 при напълно идентични спектри. В програмата те са преобразувани в **HQI** по формулата:

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

S_3 и S_4 показват подобието между спектрите, като тяхната максимална стойност е 1.0 за напълно идентични спектри. Минималната стойност на S_3 е 0.0 (за напълно ортогонални спектри) поради това, че стойностите на абсорбцията са по-големи или равни на нула. Минималната стойност на S_4 е -1.0 за отрицателно корелирани спектри. Ето защо съответните **HQI** се изчисляват по следния начин:

$$HQI_3 = 999\ S_3$$

$$HQI_4 = 999\ (1 + S_4)/\ 2$$

Математическият анализ на хит-качествените индекси показва редица техни недостатъци и предимства. Така например първият и вторият HQI са зависими от положението и вида на базовата линия, докато това не се отнася за третия и четвъртия. HQI_3 е донякъде (но слабо) зависим от нивото на базовата линия поради това, че тя представлява вектор с посока, различна от тази на спектъра. Поради втората степен в HQI_1 той е по-нечувствителен към малки разлики между спектралните криви в сравнение с HQI_2. Коефициентът на корелация (и съответно HQI_4) е добра оценка за (не)зависимостта на две серии от данни само ако те са разпределени нормално, което очевидно не е изпълнено за стойностите на абсорбцията в спектралните криви. Поради липса на нормировка HQI_1 и HQI_2 са пригодни за използване само в целия обхват на спектъра, докато третият и четвъртият алгоритъм могат да се използват в произволен спектрален интервал.

Същите мерки могат да се приложат и за УВ-Вид електронни спектри. Протонният ЯМР спектър, въпреки ясно изразената си крива с различните мултиплети с немалка ширина на пиковете е неподходящ за директно сравнение, защото мултиплетната му структура зависи от силата на магнитното поле на апарата, на който е измерен спектъра. ^{13}C ЯМР и мас-спектрите също представляват криви в компютърното си представяне, но техните пикове са с малка ширина и последната не свързана със строежа на молекулата. За тези два вида спектри е важно местоположението на пиковете (по абсцисата), а за мас-спектрите - и техния интензитет. Сравненията на тези два вида спектри са извън рамките на този курс – в магистърската програма „Спектрохимичен анализ" се изучава работата с библиотеки от ^{13}C-ЯМР и мас-спектри.

10.4. Търсене в спектрални библиотеки. Сравнението на спектри се прилага при търсене на спектри в спектрална библиотека. Спектърът на

непознатото съединение се сравнява последователно със спектрите на референтните (библиотечните) съединения и в резултат на това сравнение се получава списък от хитове, които се сортират по тяхните хит-качествени индекси. Ако непознатото съединение има спектър в библиотеката, то може да бъде надеждно идентифицирано. При липса на негов спектър, първите хитове в хит-списъка са спектри на съединения със структура, подобна на тази на непозното съединение: от структурите на хитовете могат да се направят редица изводи за структурата на непознатото съединение.

Освен последователно търсене, е възможно и индексирано търсене по различни числови характеристики, сред които са пиковете на спектъра. За ^{13}C ЯМР и мас-спектри се използва целият спектър. Индексираното търсене използва т.н. обратно индексиране (reverse index) – методите за това търсене са много добре и отдавна разработени, използват се във всякакви по род бази от данни, но са извън обсега на този курс – вижте следната статия: http://en.wikipedia.org/wiki/Reverse_index.

10.5. Химична структура на библиотечните спектри. Тя се представя в химичната информатика по различен начин, в зависимост от целите за които се използва. Исторически, първите индивидуални съединения са описвани с техния състав, т.е. молекулна формула, а след развитие на структурната теория – като набор от атоми и връзки, с които първите са свързани по между си. След откриване на електрона Люис и Косел, независимо един от друг, създават модел, в който химичната връзка е изградена от електронна двойка, обща за двете ядра. Модерното схващане на химичната структура е освен вида на атомите и връзките, в нея да се включват и пространствените координати на атомите и електронния строеж на молекулата. Трябва да се прави разлика между конфигурация и конформация: ***Конфигурация на молекулата*** е пространственото

разположение на атомите в молекулата. В по-тесен смисъл се разбира разположението на атомите един към друг, еднозначно определено само от дължините на връзките и валентните ъгли. ***Конформация на молекулата*** - всяко едно пространствено разположение на атомите в пространството, което се получава при завъртане около единична връзка на една част от молекулата спрямо останалата. Възможни са безкраен брой такива разположения на атомите в молекулата.

Засега в химичните бази от данни, както и в спектралните, е застъпено най-вече представянето на химичната структура с 2D-таблици на свързаност. Освен описание на типа на атомите и химичните връзки, в това представяне се пазят и двумерните координати на атомите, което позволява показване на химичната структура във вид, познат на химиците и спектроскопистите. Самата таблица на свързаност (connectivity table) позволява търсене по подструктури, което според съвременния стандарт е задължително в спектралните библиотеки [3,4]. Субструктурното търсене се изучава в дисциплината „Химическа информатика", чийто курс е преминат от Вас и затова не е предмет на настоящата лекция.

ЛИТЕРАТУРА

1. The Sadtler IR Search Software Manual, Sadtler Research Labs, Division of Bio-Rad Laboratories, Inc, 1988.
2. K. Varmuza, P. Penchev, H. Scsibrany. *Maximum Common Substructures of Organic compounds Exhibiting Similar Infrared Spectra*. **J. Chem. Inf. Comp. Sci.**, **38**, 420-427 (1998).
3. W.A. Warr; Computer-Assisted Structure Elucidation. Part 1: Library Search and Spectral Data Colections. **Anal. Chem.**, 1993, **65**, A1087-A1095.
4. H. Somberg; *Infrared Databases - Their Use, Structure and Implementation on a Microcomputer System*, pp. 64-91 in: J. Zupan (Ed.); *Computer-supported Spectroscopic Data Bases.* Ellis Horwood, Chichester, UK, 1986.