Харківський національний університет
імені В. Н. Каразіна

“ХЕМОІНФОРМАТИКА ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ
для хіміків”

Лекція №6
Теорія Інформації та хімія

В. В. Іванов

Кафедра хімічного матеріалознавства

•Інформація має імовірнісну природу

$$P = P_1 \cdot P_2 \cdot \ldots \cdot P_N,$$

•Адитивність інформації

$$I(P) = I(P_1 \cdot P_2 \cdot \ldots \cdot P_N) = I(P_1) + I(P_2) + \ldots + I(P_N)$$

•Інформація – це міра неоднорідності носія інформації

$$I(P) \sim \log(P_1) + \log(P_2) + \ldots + \log(P_N)$$

текст довжиною N, алфавіт довжини M

$$N = \sum_i N_i$$

$$p_i = \frac{N_i}{N} \qquad \sum_i p_i = \sum_i \frac{N_i}{N} = 1$$

Загальна кількість можливих текстів довжиною N при умові, що число символів типу і в повідомленні N_i штук,

$$P = \frac{N!}{N_1! N_2! 5 \ N_M!}$$

число перестановок з повтореннями

Інформація в одному тексті довжиною N

$$I = -\log_2 1/P = -(\log_2 1 - \log_2 P) = \frac{\ln P}{\ln 2} = \frac{1}{\ln 2}\ln\frac{N!}{N_1!N_2!5\ \ N_M!}$$

$$I = \frac{1}{\ln 2}\left(\ln N! - \ln N_1!N_2!5\ \ N_M!\right) = \frac{1}{\ln 2}\left(\ln N! - \ln N_1! - \ln N_2!5\ \ - \ln N_M!\right)$$

За допомогою наближеної формули Стірлінга:

$$\ln N! \approx N\ln\frac{N}{e} = \boxed{N\ln N - N}$$

$$I \approx \frac{1}{\ln 2}\left(N\ln N - N_1 \ln N_1 - N_2 \ln N_2 5 \; - N_M \ln N_M\right) = -\frac{1}{\ln 2}\left(\sum_i N_i \ln N_i - N\ln N\right)$$

$$I \approx -\frac{1}{\ln 2}\left(\sum_i N_i \ln N_i - \sum_i N_i \ln N\right) = -\sum_i N_i \left(\log N_i - \log N\right) = -\sum_i N_i \log \frac{N_i}{N}$$

$$I = -N \cdot \sum_i \frac{N_i}{N} \log_2 \frac{N_i}{N} = -N \cdot \sum_i p_i \log_2 p_i$$

В перерахунку на один символ

$$I = -\sum_i \frac{N_i}{N} \log_2 \frac{N_i}{N} = -\sum_i p_i \log_2 p_i$$

Приклад № 1 монета

$$I = -2 \cdot \frac{1}{2} \log_2 \frac{1}{2} = 1 \quad \text{(біт)}$$

Приклад № 2 А якщо монета може впасти на ребро?

$$P(\text{ребро}) = \frac{2}{100} \qquad P(\text{решка}) = P(\text{орел}) = \tfrac{1}{2}\left(1 - \frac{2}{100}\right) = \frac{49}{100}$$

$$I = -\left(\frac{49}{100}\log_2\frac{49}{100} + \frac{49}{100}\log_2\frac{49}{100} + \frac{2}{100}\log_2\frac{2}{100}\right) = 1{,}121$$

Приклад №3 "Закон бутерброда з маслом"

$$I = -1 \cdot \log_2 \frac{1}{1} = 0$$

Приклад № 4 Скільки інформації утримується в одній букві алфавіту ? Для 32 буквеного алфавіту маємо 5 біт:

$$I_0 = -32 \cdot \frac{1}{32} \log_2 \frac{1}{32} = 5$$

З урахуванням Частотності

Символ	Частота	Символ	Частота
Пробел	0,175	Я	0,018
О	0,090		
Е	0,072	З	0,016
А	0,062	Ь	0,014
І	0,062	Б	0,014
С	0,045	Й	0,010
Р	0,040	Х	0,009
В	0,038	Ж	0,007
Л	0,035	Ю	0,006
К	0,028	Ш	0,006
М	0,026	Ц	0,003
П	0,023	Є	0,003
У	0,021	Ф	0,002

С урахуванням Частотності

$$I = -0{,}175 \cdot \log_2 0{,}175 - 0{,}090 \cdot \log_2 0{,}090 - 5 \; - 0{,}002 \cdot \log_2 0{,}002 \approx 4{,}350$$

Байт – 8 бит,
килобайт (KB) – 1024 байта,
мегабайт (MB) – 1024 Кбайта,
гигабайт (GB) – 1024 MB.
терабайт (TB) – 1024 MB

В наших комп'ютерах 1 байт = 8 біт

Скільки різних символів можна зберегти в 1 байтовій змінній ?

$$I = 8 = -x\left(\frac{1}{x}\log_2\frac{1}{x}\right) = -\log_2\frac{1}{x}$$

$$-8 = \log_2\frac{1}{x} \qquad x = 1/2^{-8} = 2^8 = 256$$

256 символів.
Тобто 1 байтова змінна може утримувати один з 256 символів

Зв'язок інформації і ентропії

$$\Delta S = \frac{Q}{T}$$

$$S = k \ln N$$

$k=1.38 \cdot 10^{-23}$ дж/К

$$S = I$$

$$I = \log_2 N(\text{bit}) = k \ln N = k \log_2 N \cdot \ln 2$$

$$1 \text{ бит} = k \ln 2 \approx 10^{-23} \text{ дж/град}$$

Теоретико-інформаційні індекси в структурній хімії

$$\text{Інформація} = -\sum_i \frac{N_i}{N} \log_2 \frac{N_i}{N}$$

Інформаційний склад молекулярного графу відносно околиць k-го порядку – IC_k:

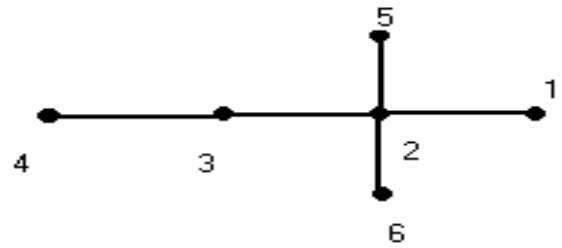
$$IC_k = -\sum_i p_i \log_2 p_i$$

$$TIC_k = NIC_k$$

$$SIC_k = IC_k / \log_2 N$$

$$BIC_k = IC_k / \log_2 N_b$$

$$CIC_k = \log_2 N - IC_k$$



$$N = P_1 + P_2 + P_3 = 5 + 7 + 3 = 15$$

$$I_D = -\left(\frac{5}{15}\log_2\frac{5}{15} + \frac{7}{15}\log_2\frac{7}{15} + \frac{3}{15}\log_2\frac{3}{15}\right) = 1{,}506$$

Індекс IC_1

Група	Атоми групи	Число атомів в групі
1	1, 5, 6, 4	4
2	2	1
3	3	1

6=4+1+1

6=4+1+1

$$IC_1 = -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6}\right) = 1{,}252$$

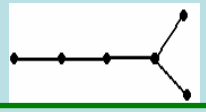
індекс IC_2 – обчислюється з урахуванням ближчих і наступних сусідів. Чотири групи еквівалентності атомів з урахуванням сусідів “второго порядку”:

Група	Атоми групи	Число атомів в групі
1	1, 5, 6	3
2	2	1
3	3	1
4	4	1

$$IC_2 = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6}\right) = 1{,}792$$

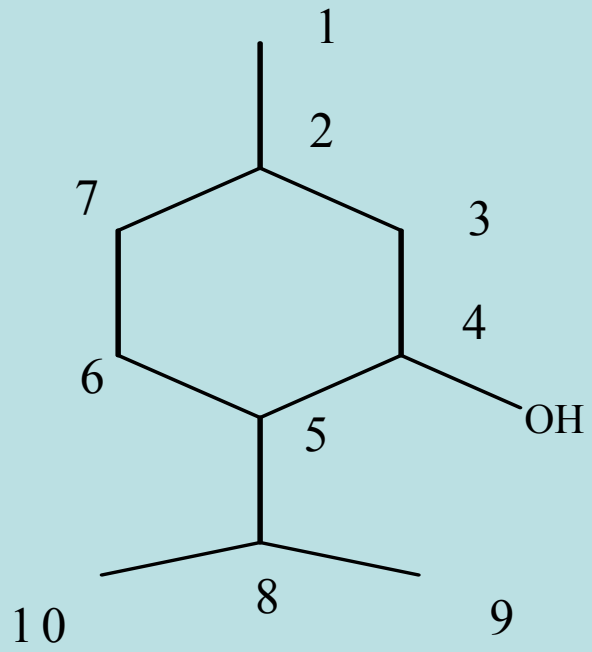
Деякі топологічні індекси для двох ізомерів гексану.





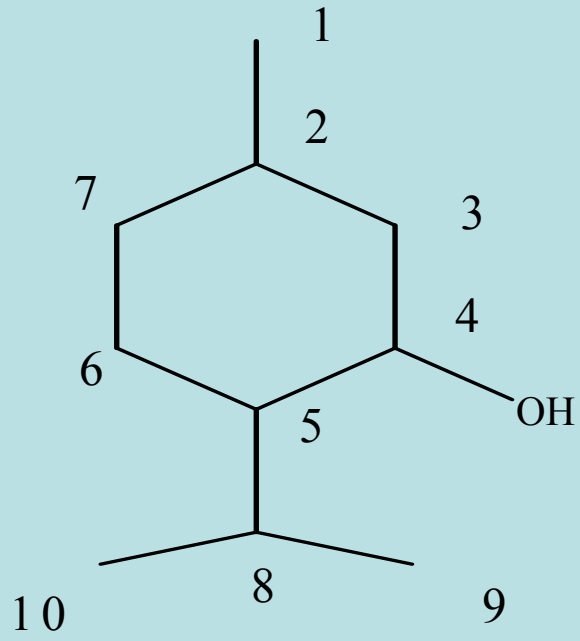
Індекс	(н-гексан)	(2,2-диметилбутан)	(2-метилпентан)
N_C	6	6	6
P_1	5	5	5
P_2	4	7	5
P_3	3	3	3
I_D	2,149	1,506	1,549
IC_0	0	0	0
IC_1	0,918	1,252	1,459
IC_2	1,585	1,792	2,252

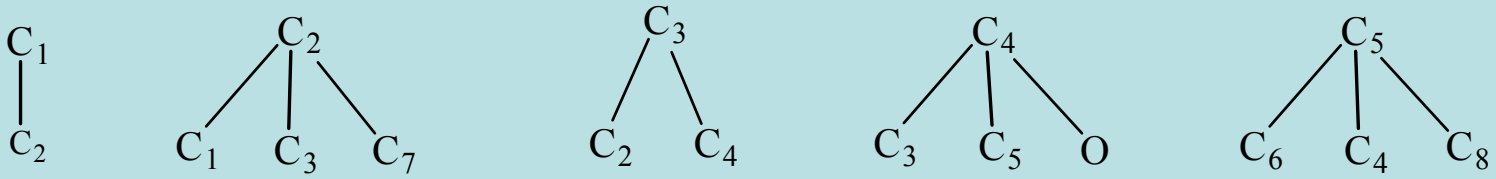
Молекула Ментолу

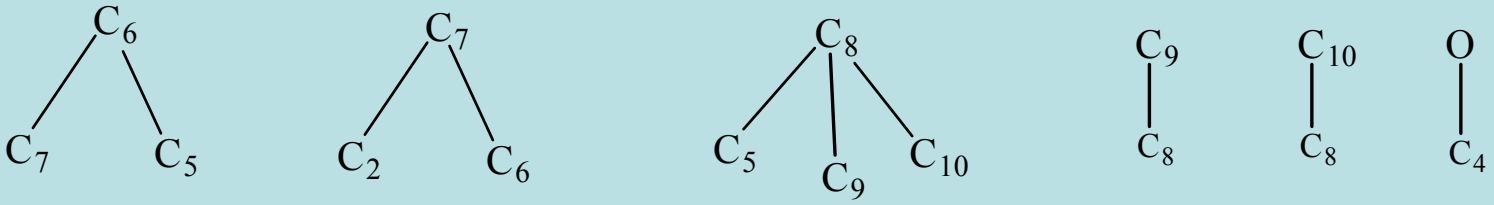


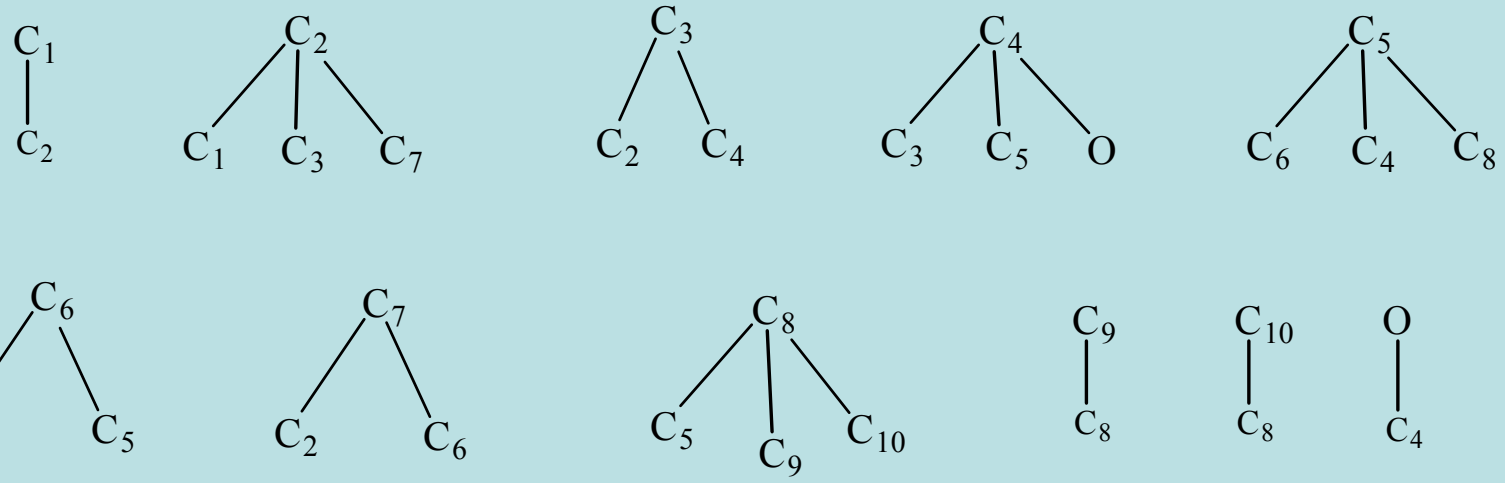
$$IC_0 = -\left(\frac{10}{11}\log_2\frac{10}{11} + \frac{1}{11}\log_2\frac{1}{11}\right) = 0{,}439$$

Молекула Ментолу









Имеется 5 групп эквивалентности.

C_1, C_9, C_{10}, – 3 атома.

C_2, C_5, C_8, – 3 атома.

C_3, C_6, C_7, – 3 атома.

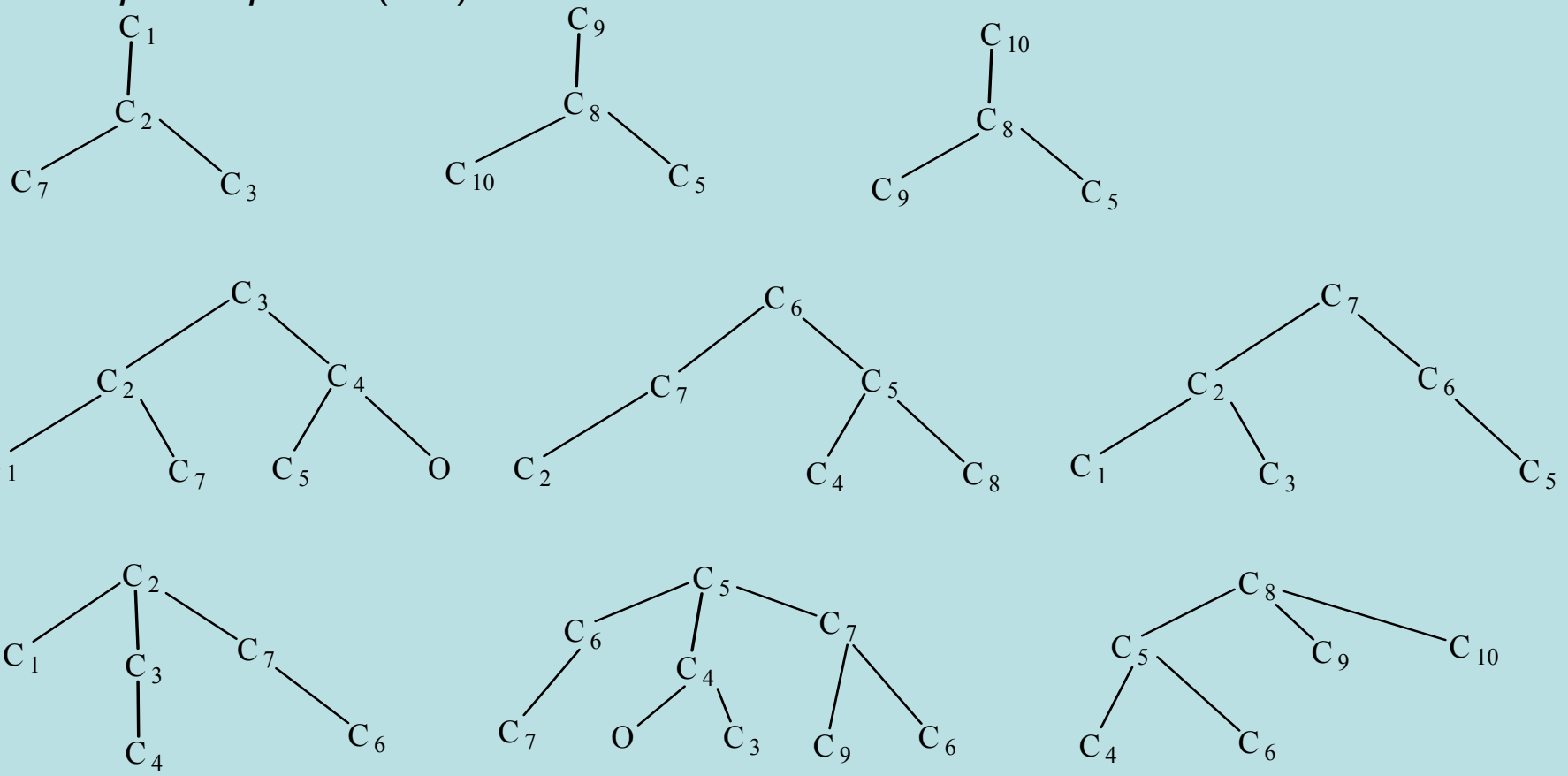
C_4 – 1 атом.

O – 1 атом.

$$IC_1 = -\left(\frac{3}{11}\log_2\frac{3}{11} + \frac{3}{11}\log_2\frac{3}{11} + \frac{3}{11}\log_2\frac{3}{11} + \frac{1}{11}\log_2\frac{1}{11} + \frac{1}{11}\log_2\frac{1}{11}\right) = 2{,}163$$

Молекула Ментола

Второй порядок (IC2).



C1, C9, C10 – 3 атома. C6, C7 – 2 атома. А также 6 групп по одному атому

$$IC_2 = -\left(\frac{3}{11}\log_2\frac{3}{11} + \frac{2}{11}\log_2\frac{2}{11} + 6\cdot\frac{1}{11}\log_2\frac{1}{11}\right) = 2{,}845$$

Приклади використання індексів

(Розчинність спиртів у воді.
Коефіцієнти регресії знайдені по 27 точкам)

$$-\lg X = 2.37\,CIC_1 - 3.52$$ R=0.92

$$-\lg X = 3.73\,CIC_1 + 8.75\,SIC_2 - 0.71\,IC_3 - 10$$ R=0.97

Where is the wisdom ?
Lost in the knowledge.
Where is the knowledge ?
Lost in the information.

T.S. Eliot (1934)

Where is the information ?
Lost in the data.
Where is the data ?
Lost in the database.
Where is the database ?
Lost in the Internet.

? (today)