



*Харківський національний університет
імені В. Н. Каразіна*

**“ІНФОРМАТИКА І ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ
для хіміків”**

Лекція №6

Теорія Інформації та хімія

В. В. Іванов

Кафедра хімічного матеріалознавства

- Информация имеет вероятностную природу

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

- Аддитивность информации.

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

- Информация – это мера неоднородности носителя информации

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

текст длиной N , алфавит длиной M

$$N = \sum_i N_i$$

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

Общее число возможных сообщений длиной N при условии, что символов типа i в сообщении N_i штук,

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

число перестановок с повторениями

Информация в одном сообщении длиной N

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

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

С помощью приближенной формулы Стирлинга:

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

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

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

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

В расчете на символ

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

Пример № 1 монета

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

Пример № 2 А если монета упала на ребро?

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

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

Пример №3 "Закон бутерброда с маслом"

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

Пример № 3 Сколько информации содержится в одной букве алфавита? Для 32 буквенного алфавита получаем ровно 5 бит:

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

С учетом Частотности

Символ	Частота	Символ	Частота
Пробел	0,175	Я	0,018
О	0,090	Ы	0,016
Е, Ё	0,072	З	0,016
А	0,062	Ь, Ь	0,014
И	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	Э	0,003
У	0,021	Ф	0,002

С учетом Частотности

$$I = -0,175 \cdot \log_2 0,175 - 0,090 \cdot \log_2 0,090 - \dots - 0,002 \cdot \log_2 0,002 \approx 4,350$$

$$I(\text{"Война и мир"}) \approx 1,5 \cdot 10^7 (\text{бит})$$

Байт - 8 бит,
килобайт (КВ) - 1024 байта,
мегабайт (МВ) - 1024 Кбайта,
гигабайт (ГВ) - 1024 МВ.
терабайт (ТВ) - 1024 МВ

$$I(\text{"Война и мир"}) \approx 1,5 \cdot 10^7 (\text{бит}) = 1,875 \cdot 10^6 (\text{байт}) \\ = 1831,05(\text{КВ}) = 1,78(\text{МВ}).$$

В наших компьютерах 1 байт = 8 бит

Сколько различных символов можно сохранить в 1 байтовой переменной ?

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

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

256 символов.

Т.е. 1 байтовая переменная может содержать один из 256
СИМВОЛОВ

Связь информации и энтропии

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

$$S = k \ln N$$

$$k = 1.38 \cdot 10^{-23} \text{ Дж/К}$$

$$S = I$$

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

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

Теоретико-информационные индексы в структурной химии

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

Информационное содержание молекулярного графа относительно окрестностей k -го порядка – IC_k :

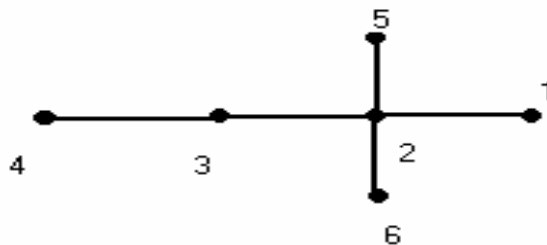
$$IC_k = -\sum_i p_i \log_2 p_i$$

$$TIC_k = NIC_k$$

$$SIC_k = IC_k / \log_2 N$$

$$BIC_k = IC_k / \log_2 N_b$$

$$CIC_k = \log_2 N - IC_k$$



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

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

Индекс IC_1

Группа	Атомы группы	Число атомов в группе
1	1, 5, 6, 4	4
2	2	1
3	3	1

$$6 = 4 + 1 + 1$$

$$6=4+1+1$$

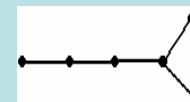
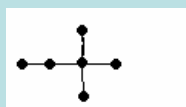
$$IC_1 = - \left(\frac{4}{6} \log_2 \frac{4}{6} + \frac{1}{6} \log_2 \frac{1}{6} + \frac{1}{6} \log_2 \frac{1}{6} \right) = 1,252$$

индекс IC_2 – вычисляется с учетом ближайших и следующих за ними соседей. Четыре группы эквивалентности атомов с учетом соседей “второго порядка” имеет вид

Группа	Атомы группы	Число атомов в группе
1	1, 5, 6	3
2	2	1
3	3	1
4	4	1

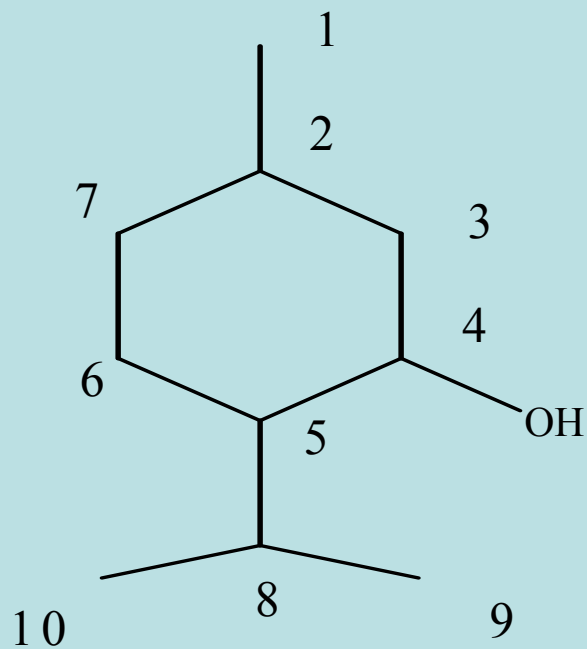
$$IC_2 = - \left(\frac{3}{6} \log_2 \frac{3}{6} + \frac{1}{6} \log_2 \frac{1}{6} + \frac{1}{6} \log_2 \frac{1}{6} + \frac{1}{6} \log_2 \frac{1}{6} \right) = 1,792$$

Некоторые топологические индексы для двух изомеров гексана.



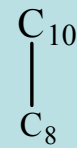
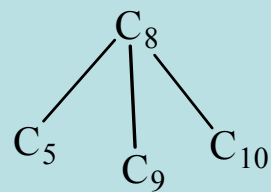
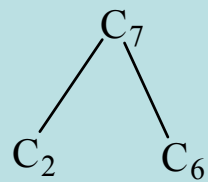
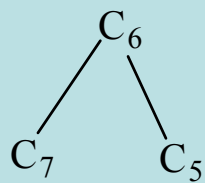
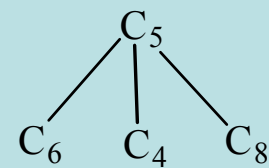
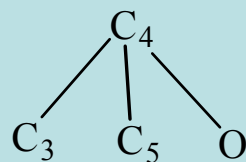
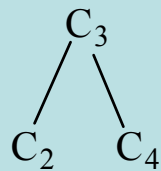
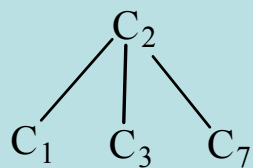
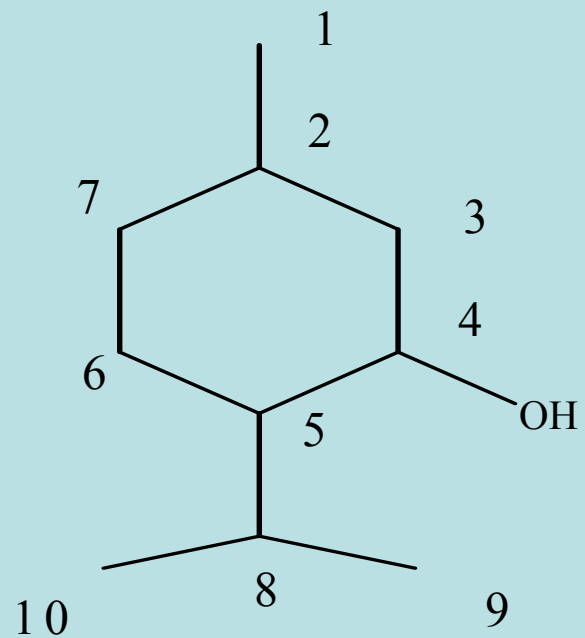
Индекс	(н-гексан)	(2,2-диметилбутан)	(2-метилпентан)
N_C	6	6	6
P_1	5	5	5
P_2	4	7	5
P_3	3	3	3
I_D	2,149	1,506	1,549
IC_0	0	0	0
IC_1	0,918	1,252	1,459
IC_2	1,585	1,792	2,252

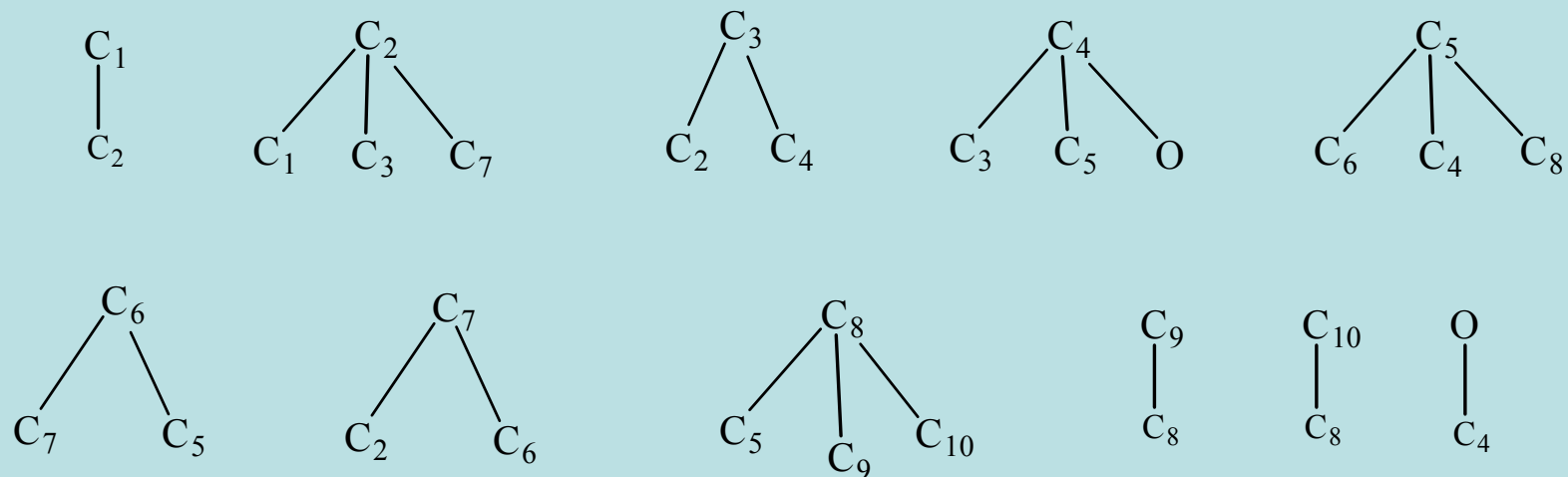
Молекула Ментола



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

Молекула Ментола





Имеется 5 групп эквивалентности.

C₁, C₉, C₁₀, – 3 атома.

C₂, C₅, C₈, – 3 атома.

C₃, C₆, C₇, – 3 атома.

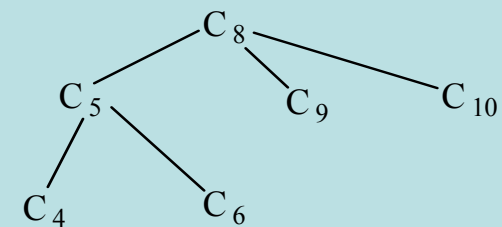
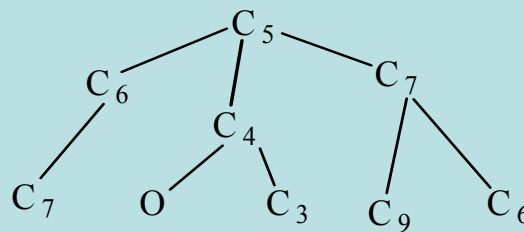
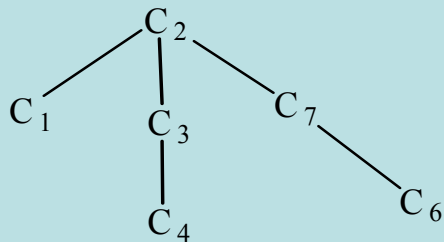
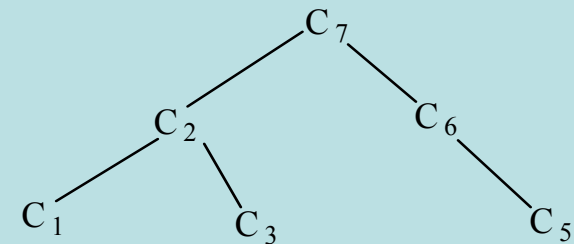
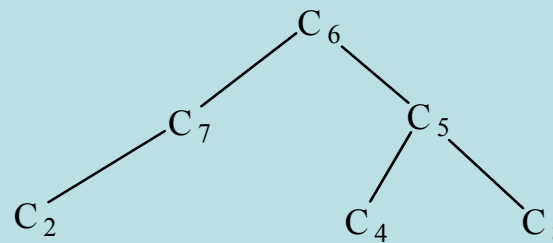
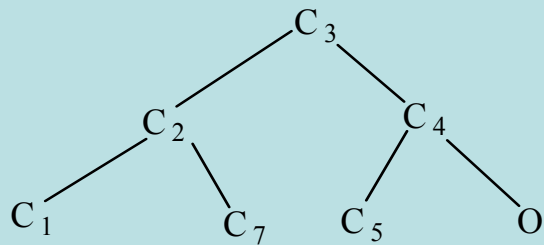
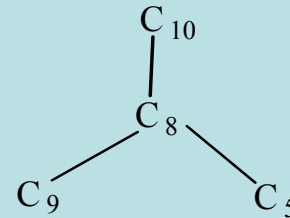
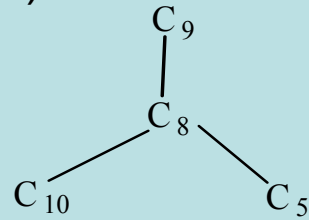
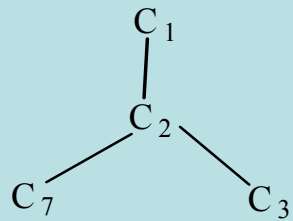
C₄ – 1 атом.

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

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

Примеры использования индексов

(Растворимость спиртов в воде.

Коэффициенты регрессии найдены по 27 точкам)

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

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

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 internet.

NN (today)