Проект 10. Пропускная способность канала с выпадениями и перемешиванием

Валерия Кирова, Ксения Кувшинова

Университет «Сириус», г. Сочи

21 августа 2020 г.

DNA digital data storage Введение

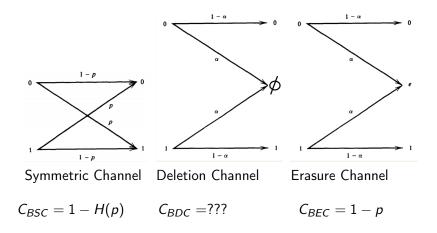
- Процесс кодирования и декодирования бинарных данных из последовательностей нуклеотидов.
- Имеет огромный потенциал использования, но очень высокие затраты по использованию.
- В июне 2019 года были закодированы все 16 GB английской Википедии с помощью искусственной ДНК.
- Иммобилизация одноцепочечной ДНК на твердом носителе, амплификация последовательностей полимеразной цепной реакцией (ПЦР) и маркировка отдельных оснований ДНК комплементарными основаниями, меченными флуоресцентными маркерами.

DNA digital data storage

Более подробно

- ДНК представляет собой длинную молекулу, состоящую из четырех нуклеотидов, по сути четырёхбуквенный алфавит.
- Неэффективно синтезировать длинные нити ДНК, а также невозможно пространственно упорядочить их.
- На сегодняшний день информация хранится на коротких последовательностях ДНК (не более нескольких сотен нуклеотидов).
- Простейшей математической моделью системы хранения ДНК является канал перемешивания.

Типы каналов



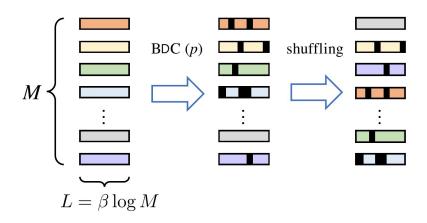
$$X=[X_1,...,X_M]$$
 - вход канала - M строк длины L . $Y=[Y_1,...,Y_M]$ - выход канала - M строк неизвестной длины. $S\in\{1,2,...,M\}^M$ - случайное перемешивание.

X и Y между собой связаны:

$$Y_{S(k)}^{[l]} = egin{cases} X_k^{[l]}, & ext{c вероятностью } 1-p \ \emptyset & ext{c вероятностью } p \end{cases}$$

- Всего кодовых слов: 2^{MLR}
- Количество информационных битов в строке длины L: LR_{BDC}
- Количество информационных битов для хранения индексов: log M
- Количество эффективных битов: LR_{BDC} log M
- Скорость схемы:

$$\frac{M(LR_{BDC} - \log M)}{ML} = R_{BDC} - \frac{1}{\beta}$$



$$C = (1 - \rho_0)(1 - \frac{1}{\beta})$$

- Конкатенация каналов.
- p₀ вероятность того, что на выходе из канала мы не получим исходную последовательность, т.е. что произойдёт хотя бы одна ошибка.
- $\frac{1}{\beta}$ вероятность того, что произойдёт перемешивание.

$$C=C_{BDC}-rac{1}{eta}$$
 - ёмкость канала с выпадениями и перемешиванием

• Перемешивание понижает ёмкость на $\frac{1}{\beta}$ бит.

Верхние и нижние границы

$$C_I = p \log p + (1-p)(1 + \log(1-p))$$
 - нижняя граница

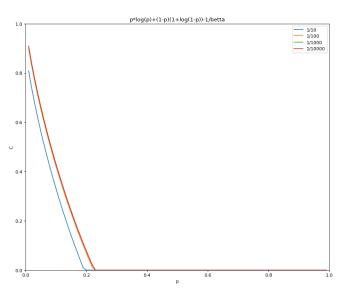
 Выводится из общей нижней границы для канала с выпадениями и стираниями:

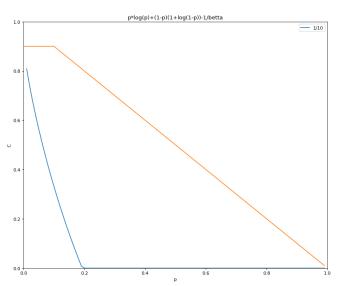
$$C(p,r) \leq (1-p)\log q + p\log p - H_q(r,p)$$
, где q - мощность алфавита, $H_q(r,p) = -(1-p-r)\log(1-p-r) - p\log(rac{p}{q-1})$

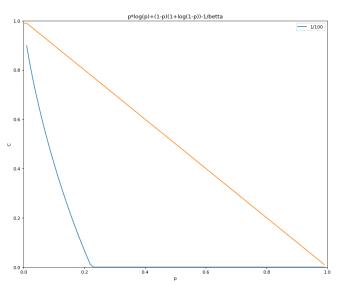
Верхние и нижние границы

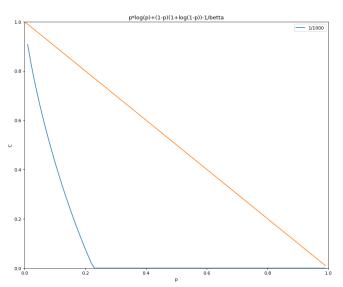
$$\mathcal{C}_u = \min(1-p, 1-rac{1}{eta})$$
 - верхняя граница

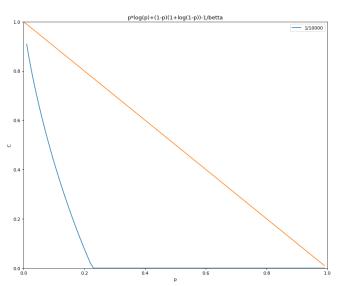
• Очевидно!











- А.С. Долгополов, Границы пропускной способности канала в ошибками синхронизации. Проблемы передачи информации, том 26, Москва, 1990
- Seiyun Shin, Reinhard Heckel, Ilan Shoromony, Capacity of the Erasure Shuffling Channel, ICASSP 2020
- Ilan Shoromony, Reinhard Heckel, Capacity Results for the Noisy Shuffling Channel, ISIT 2019
- Ilan Shoromony, Reinhard Heckel, DNA-Based Storage: Models and Fundamental Limits, ISIT 2017