**Проблема аннотации биологических последовательностей**.

1. Периодичность кодирующих районов
2. Кодирующие потенциалы и методы поиска генов
3. Применение марковских моделей.
4. Использование теории информации
5. Применение нейронных сетей

# Where are the coding regions?

TCAGCGAAGATGAGATAGTTTTTAAAGGTGGGATTTCCCCACCTTTAAAAAGCGAGAAGTCCCGGTTTTAA
AGAGGAGTAAAATCCTCTTTTTCTAGCCCACTCAGGTGGTTTTTTTGGTTTTCGCTCCTTGCCGCATCTTC
TGTGCCTTTGATGGCGGCTGGTTGGGGTGAAAGGCTGCATATTCCAGAATTTCAGACAGTAGATTGTTTTT
GAAATCTTCCGTTTTATCGTTGACGAACTTAACCATCCTGTTGAAATCATCTTCCTTTGATACACCTTCAG
GAAATGCCTTAGGAACTGATGTTTGGCTATCCAAGGCATCTTGCAATATCTGCACGATCTCCGAATTCATT
GATCGCCCATTGGCCTTTGCTCTGGCGGCAACTGCGTCACGCATACCGTCAGGCATCCTAACTGTAAATCT
CTCAATGAAAGCTGGATCTTCTTTTTCAGTCATCATCTTAAACCATAAAAATTTATACAAAACACACTAGC
ATCATATTGACATTACCCACAATGACATCATAATGGTGTCAGGCATCAAAATGATGTCATCATGACAAGGG
GAAAGTAAATGCAAGATGTTCTCTATACAGGTCGTAAGAACGACAGCTTTCAGCTTCGTCTGCCTGAGCGA
ATGAAAGAAGAGATCCGTCGCATGGCAGAGATGGACGGCATTTCGATTAATTCTGCAATCGTGCAGCGCCT
TGCTAAAAGCTTGCGTGAGGAAAGAGTTAATGGGCAGTAAAAACAGCGAAGCCCGGAAGTGTGGGGACACT
AACCGGGCTTCTAATGTCAGTTACCTAGCGGGAAACCAACAATGACCAGTATAGCAATCTTTGAAGCAGTA
AACACTATCTCTCTTCCATTCCACGGACAGAAGATCATAACTGCGATGGTGGCGGGTGTGGCGTATGTGGC
AATGAAGCCCATCGTGGAAAACATCGGTTTAGACTGGAAGAGCCAGTATGCCAAGCTCGTTAGTCAGCGTG
AAAAGTTCGGGTGTGGTGATATCACCATACCTACCAAAGGTGGTGTTCAGCAGATGCTTTGCATCCCTTTG
AAGAAACTGAATGGATGGCTCTTCAGCATTAACCCAGCAAAAGTACGTGATGCAGTTCGTGAAGGTTTAAT
TCGCTATCAAGAAGAGTGTTTTACAGCTTTGCACGATTACTGGAGCAAAGGTGTTGCAACGAATCCCCGGA
CACCGAAGAAACAGGAAGACAAAAAGTCACGCTATCACGTTCGCGTTATTGTCTATGACAACCTGTTTGGT
GGATGCGTTGAATTTCAGGGGCGTGCGGATACGTTTCGGGGGATTGCATCGGGTGTAGCAACCGATATGGG
ATTTAAGCCAACAGGATTTATCGAGCAGCCTTACGCTGTTGAAAAAATGAGGAAGGTCTACTGATTGGCGT
ATTGGAAGGCGCAAAAAGAAAAGCCAGCAGATGGGCTGCTGGCATTCATTGGGTATATGAACTTTCGGAGA
ACATATGAAGTCAATTATCAAGCATTTTGAGTTTAAGTCAAGTGAAGGGCATGTAGTGAGCCTTGAGGCTG
CAAGCTTTAAAGGCAAGCCAGTTTTTTTAGCAATTGATTTGGCTAAGGCTCTCGGGTACTCAAATCCGTCA

Таблица 1. Характерные длины в геномных последовательностях ДНК (в па́рах нуклеотидов для двухцепочечных молекул и в числах нуклеотидов для одноцепочечных молекул)

| Механизм | Длина, $l$ |
|---|---|
| **Строение спирали ДНК** | |
| Чередование пуринов и пиримидинов в Z-форме | 2 |
| Шаг спирали в B-форме | 10,2 – 10,5 |
| Шаг спирали в A-форме | 10,8 |
| Шаг спирали в Z-форме | 12,0 |
| **Упаковка хроматина** | |
| Нуклеосома | $200 \pm 40$ |
| Шаг спирали в 30 нм фибрилле | $(200 \pm 40) \times 6$ |
| Петли | $2 \times 10^4 - 10^5$ |
| Субъединица из петель | $\sim 10$ петель |
| **Изгибные характеристики** | |
| Длина Куна для двухцепочечной ДНК | 300 – 400 |
| Длина Куна для одноцепочечной ДНК | 12 – 14 |
| **Структура белков** | |
| Кодон (отвечает одной аминокислоте) | 3 |
| Шаг $\alpha$-спирали | $(3{,}6 \pm 0{,}2) \times 3$ |
| Чередование водородных связей в элементах $\beta$-структуры | $2{,}0 \times 3$ |
| Длины элементов вторичной структуры | $(8 - 20) \times 3$ |
| Длины белков | $(70 - 2000) \times 3$ |

# Применение преобразование Фурье для поиска кодирующих последовательностей ДНК

- Задана последовательность:$x_j$, $j=1,\ldots,N$
- $x_j$ может принимать $a,t,c,g$
- $U_\alpha(x_j)$ – бинарная индикаторная последовательность
- $U_\alpha(x_j)=1$ если $x_j=\alpha$
- $U_\alpha(x_j)\neq 1$ если $x_j\neq\alpha$
- $U_\alpha(x_j)$ – вводится для каждого символа из используемого алфавита

# Применение преобразование Фурье для поиска кодирующих последовательностей ДНК

| Sequence | G | G | A | T | A | T | C | A | C | T | T | T | A | G | A | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Apply $U_A$ | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| Apply $U_T$ | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| Apply $U_G$ | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| Apply $U_C$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

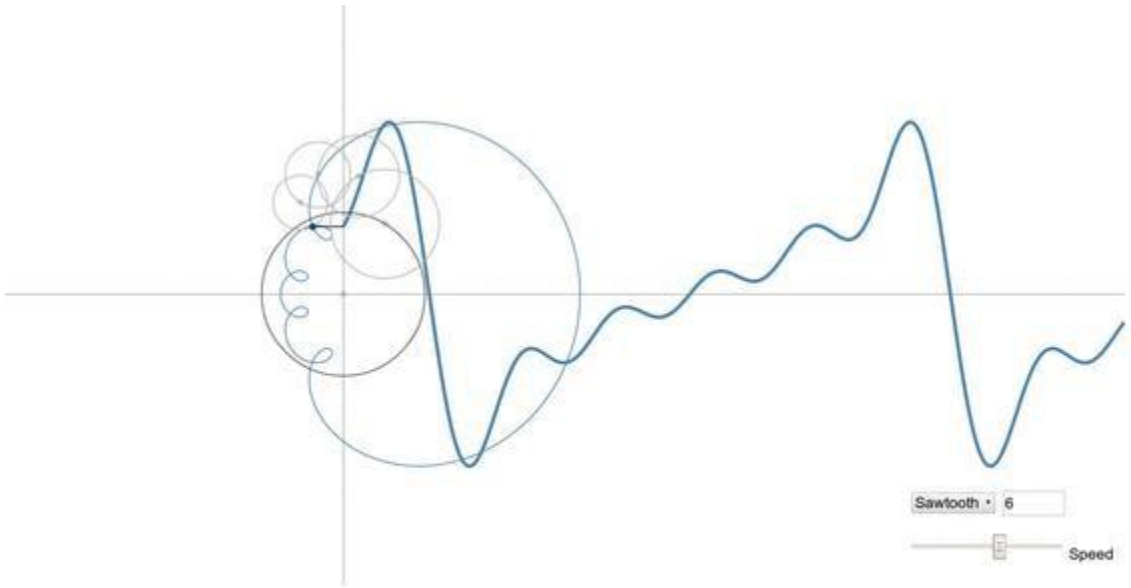
# Применение преобразование Фурье для поиска кодирующих последовательностей ДНК

$$S(f) \equiv \sum_{\alpha} S_{\alpha}(f) = \sum_{\alpha} \frac{1}{N_2} \left| \sum_{j=1}^{N} U_{\alpha}(x_j) \exp 2\pi i f j \right|^2$$

$f$ – дискретная частота;
$f=k/N,\ k=1,2,\ldots,N/2$

$$\overline{S} \equiv \frac{2}{N} \sum_{k=1}^{N/2} S(k/N) = \frac{1}{N} \left( 1 + \frac{1}{N} - \sum_{\alpha} P_{\alpha}^2 \right)$$

$P_\alpha$ – вероятность встретить символ α в изучаемой последовательности


Sawtooth
6
Speed

# Применение преобразование Фурье для поиска кодирующих последовательностей ДНК

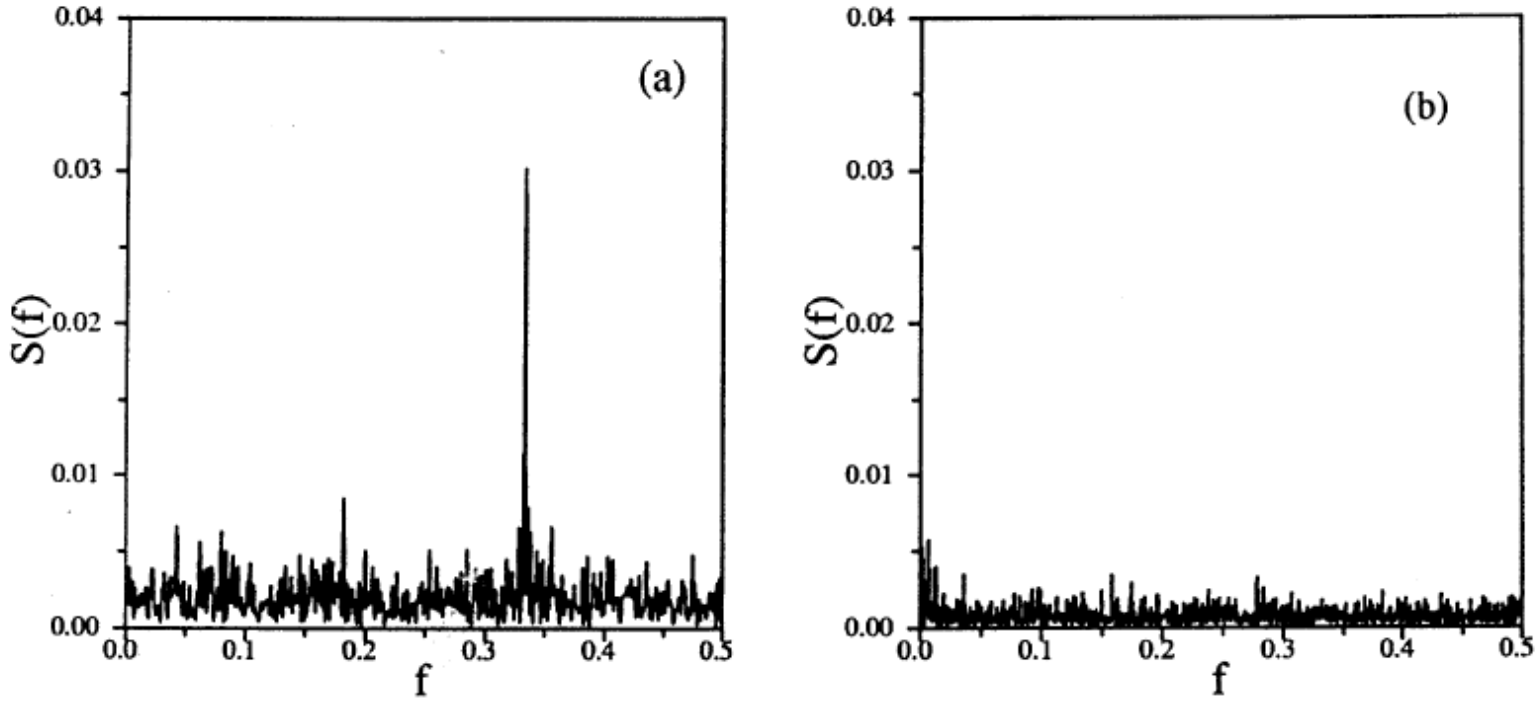
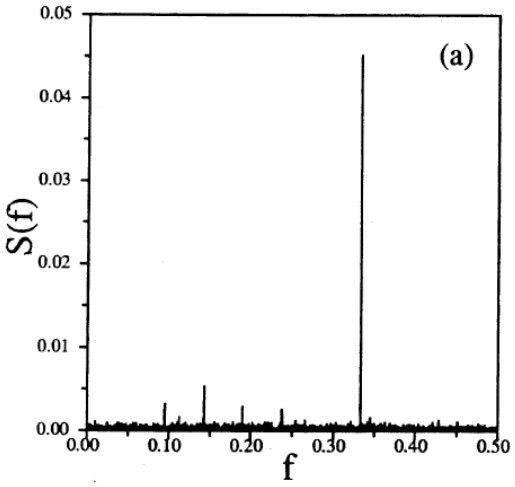
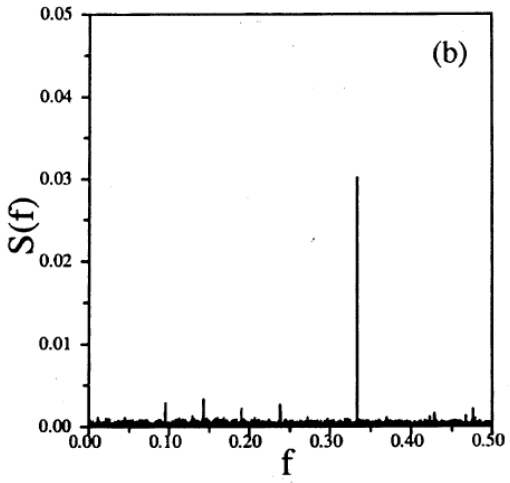
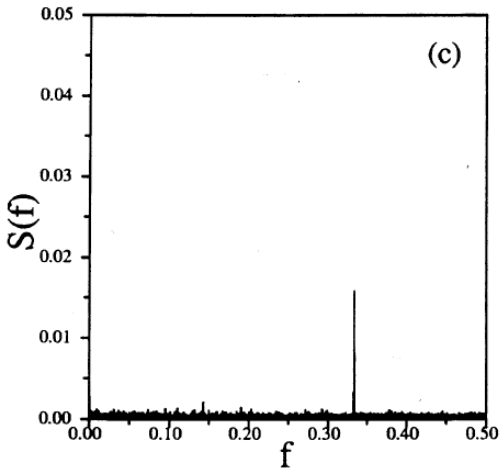



**Fig. 1.** Typical Fourier spectra for (a) a coding stretch of DNA and (b) a non-coding stretch from *S.cerevisiae* chromosome III.

# Применение преобразование Фурье для поиска кодирующих последовательностей ДНК



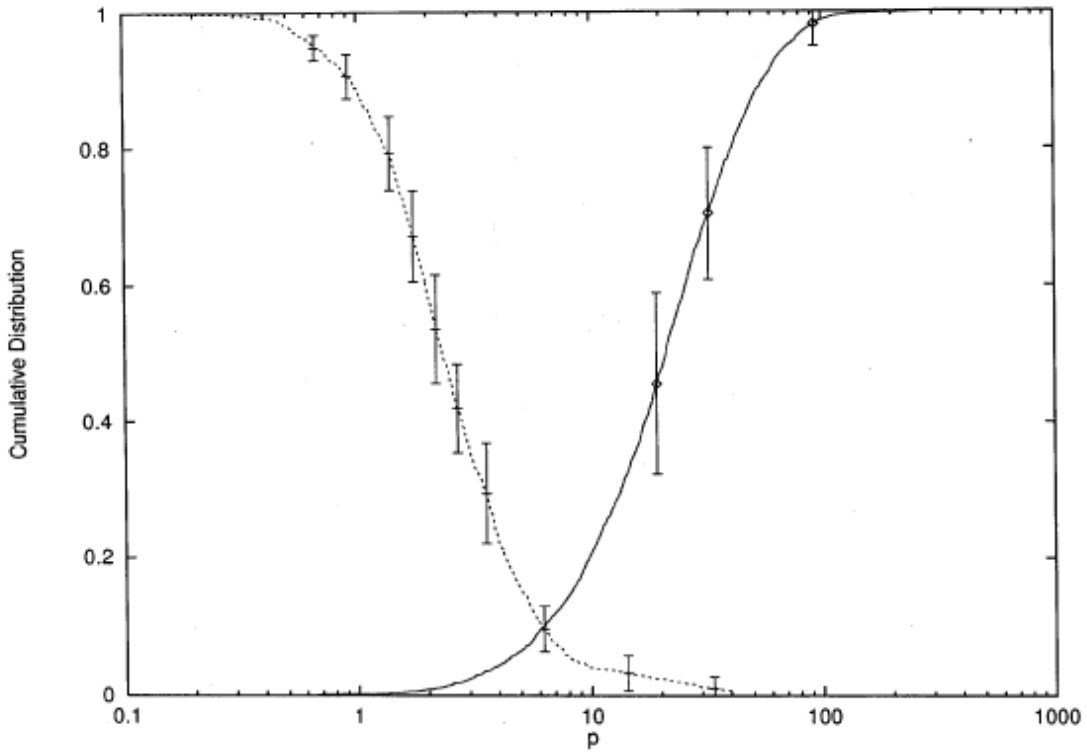




a – последовательность гена миозина человека

b – ДНК=>aa seq.(перекодировка с равномерным использованием кодонов) => ДНК

b – ДНК=>aa seq.(перекодировка в произвольный генетический код) => ДНК

# Применение преобразование Фурье для поиска кодирующих последовательностей ДНК




$$P = S(1/3)\sqrt{S}$$

Правая кривая (непрерывная) – доля кодирующих последовательностей с P меньше, чем абсцисса

Левая кривая (прерывистая) – доля некодирующих последовательностей с P больше, чем абсцисса

# Применение преобразование Фурье для поиска кодирующих последовательностей ДНК

**Table II.** Summary of results for *S.cerevisiae* chromosomes III and VIII, and *H.influenzae*

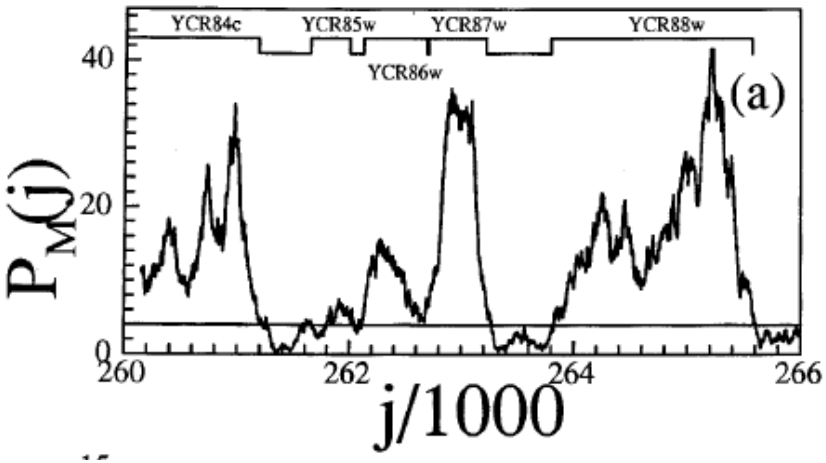
| | Chromosome III | Chromosome VIII | *H.influenzae* |
|---|---|---|---|
| ORFs | 216 | 267 | 1727 |
| ORFs detected[1] | 187 | 226 | 1499 |
| False positives detected | 0 | 0 | 0 |
| Specificity | 1.0 | 1.0 | 1.0 |
| Sensitivity | 0.87 | 0.85 | 0.87 |
| Genes reported | 54 | 140 | 933 |
| Genes detected | 44 | 123 | 867 |
| Sensitivity | 0.81 | 0.88 | 0.93 |

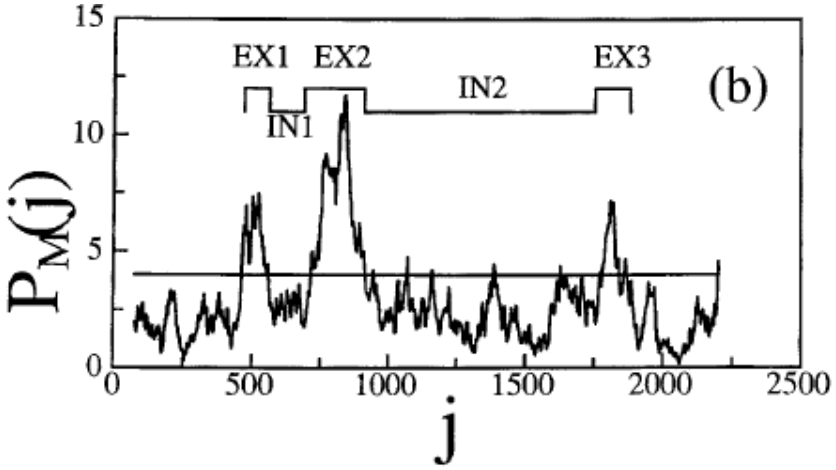
# Применение преобразование Фурье для поиска кодирующих последовательностей ДНК

**Table IIIA.** Summary of results for human and *C.elegans* genomic sequences

| | *C.elegans* | Human |
|---|---|---|
| Genes reported | 146 | 24 |
| Genes detected | 146 | 24 |
| Exons reported | 982 | 141 |
| Exons detected | 837 | 119 |
| Exons > 100 bp reported | 844 | 93 |
| Exons > 100 bp detected | 764 | 86 |

# Применение преобразование Фурье для поиска кодирующих последовательностей ДНК




**а** - Идентификация кодирующих последовательностей в геноме *S.cerevisiae* в хромосоме III в окне длиной 351 основание.

**б** - ß-глобин коз. Фрагмент длиной в 2278 нуклеотидов.

**j** – меняется с шагом в 3 основания

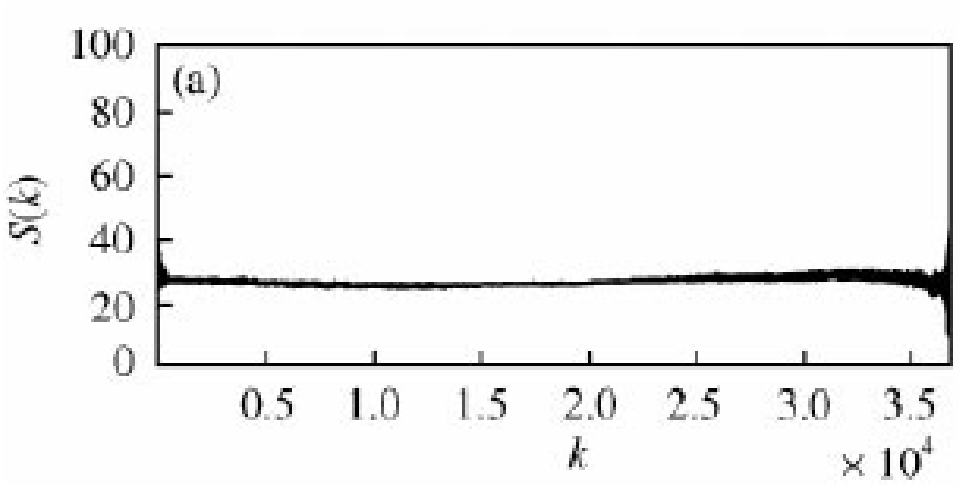
# Применение преобразование Фурье для поиска кодирующих последовательностей ДНК

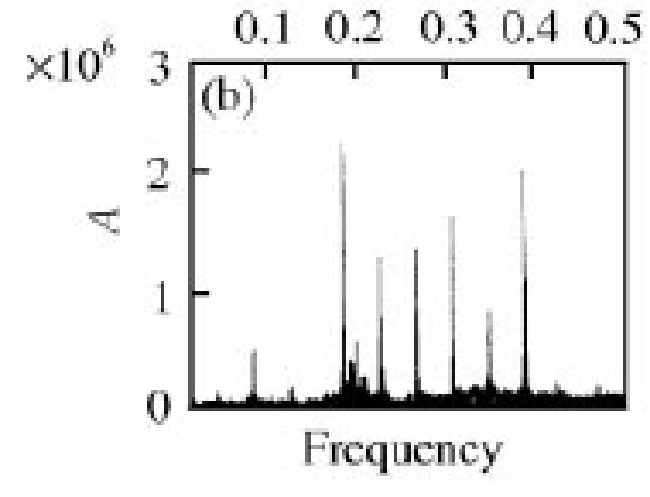
- Последовательность $x(i)$, $i=F,L$

- Если $x(i)=x(i+k)$ => $g_{i,i+1+k}=1$
- Если $x(i)\neq x(i+k)$ => $g_{i,i+1+k}=0$

$$S(k)=\sum_{i=F}^{L-1-k}\frac{g_{i,i+1+k}}{(L-F-k)}$$

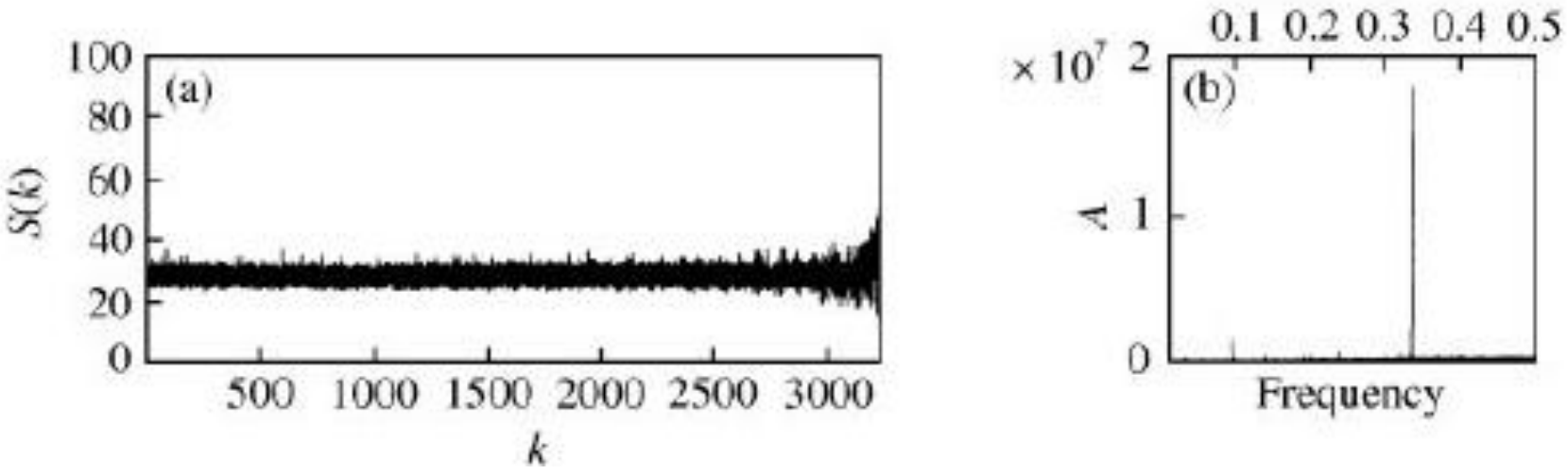
# Применение преобразование Фурье для поиска кодирующих последовательностей ДНК




Район Центромеры 22 хромосомы человека (accession number AP000543)

# Применение преобразование Фурье для поиска кодирующих последовательностей ДНК




Человеческий MyHC-perynatal 3’gene (Y00821), coding sequence 1-3237

# Расчет взаимной информации

$$H(1) = -\sum_{i=1,n} p(i) log_2 p(i)$$

$$p(i) = n(i) / L$$

$$H(2) = -\sum_{i=1,m} f(i) log_2 f(i)$$

**tcggtagt**

**atcgtacg**

a/a – первый символ, a/t – второй символ,… всего 16 символов для $n$=4 и $m$=4

$$H(1,2) = -\sum_{i=1,nm} t(i) log_2 t(i)$$

$I(1,2)=H(1)+H(2)-H(1,2)$ средняя общая информация на букву

# Применение взаимной информации для поиска кодирующих областей

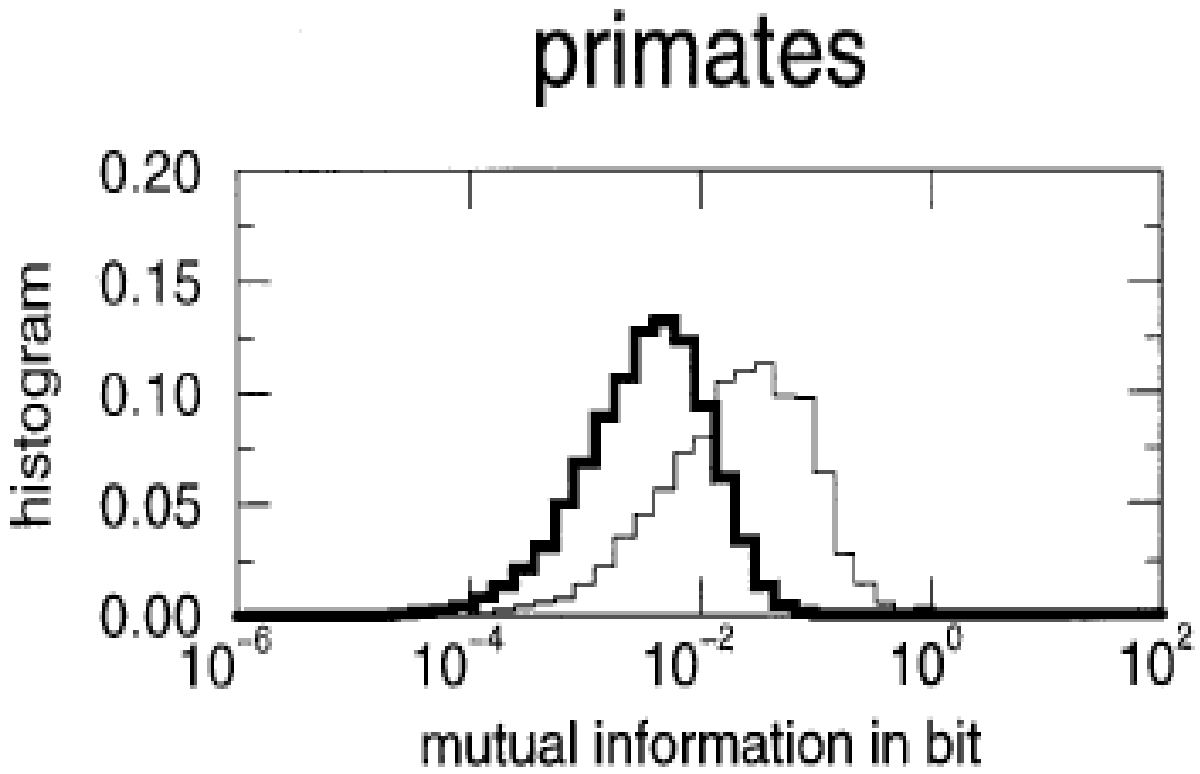
- I(x,y)=H(x)+H(y)-H(x,y)

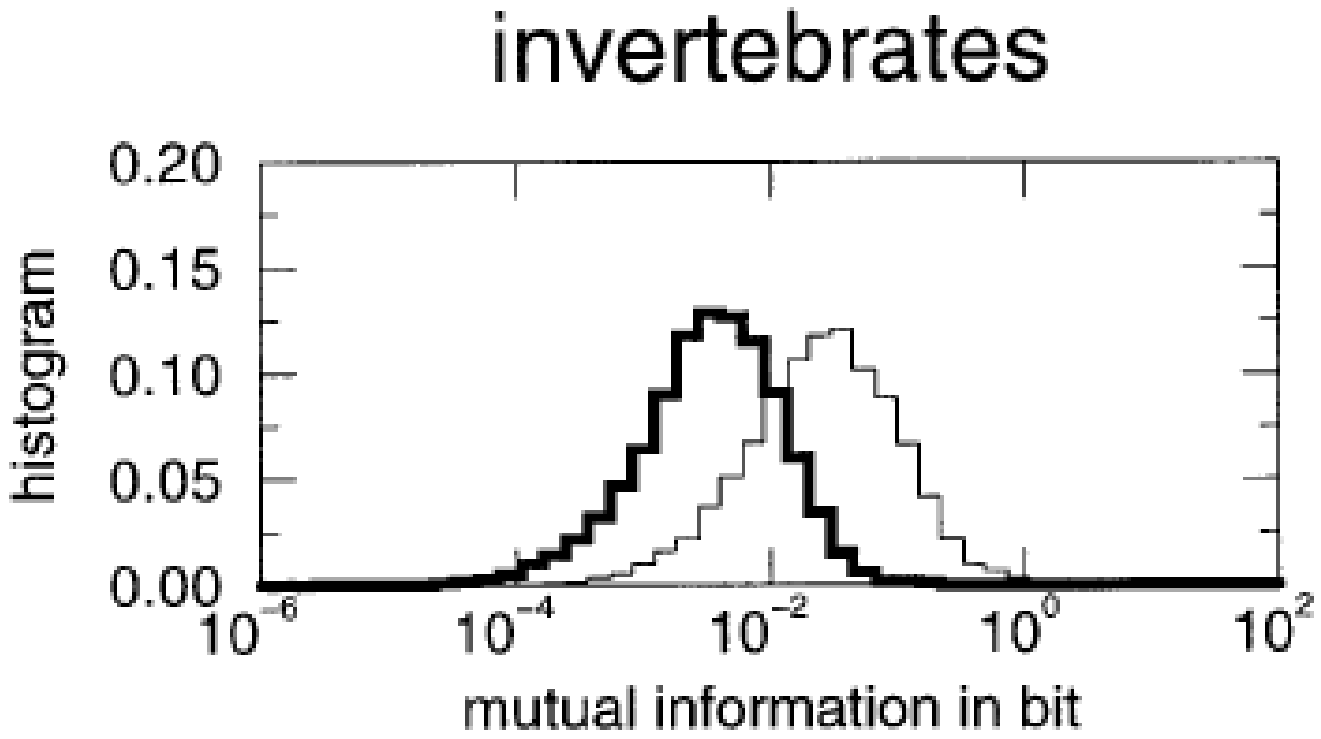
- $H(x)=-\Sigma p(i)\log_2(p(i))$

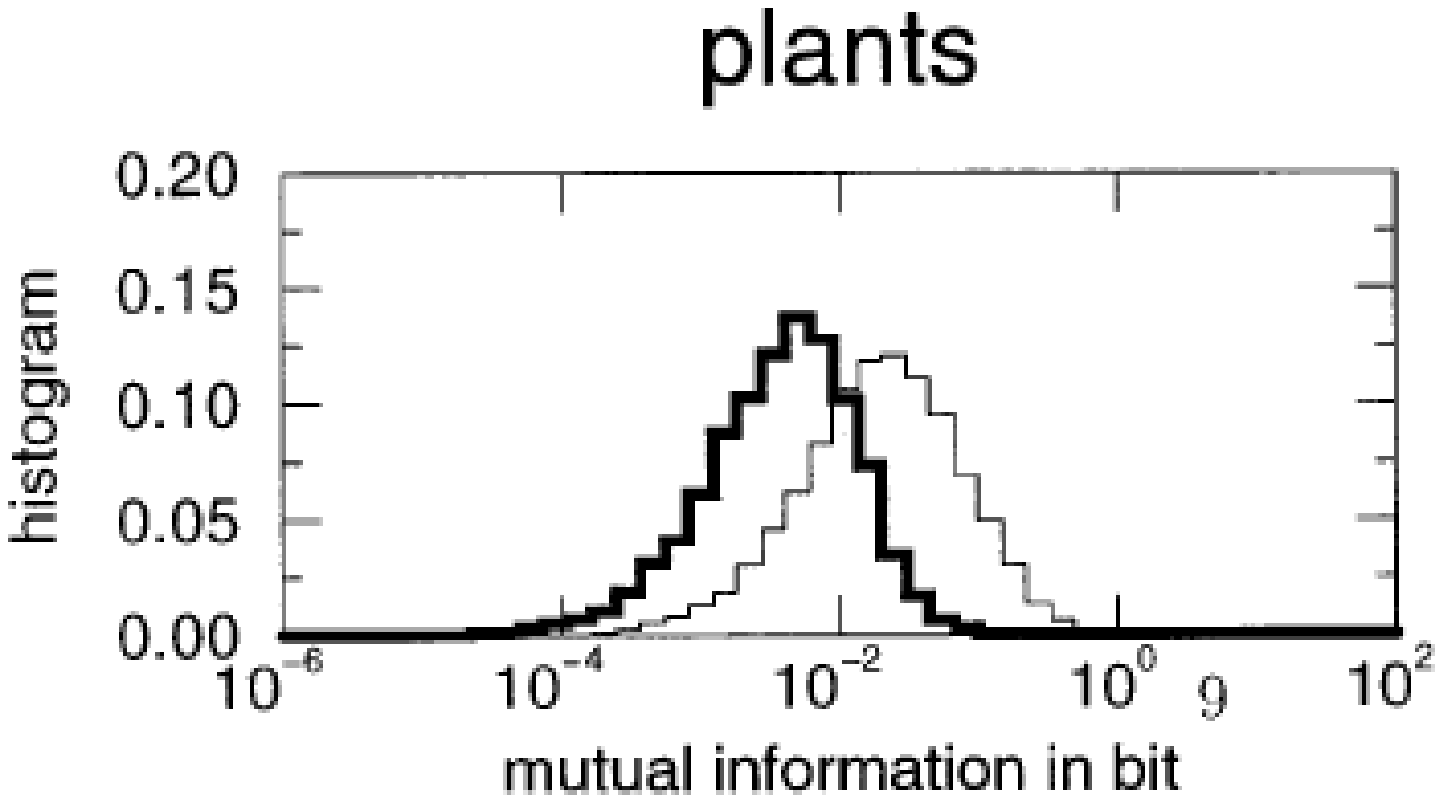
- atc**g**cg**a**tc**g**ta**g**tc**g**
- **a**tc**g**cg**a**tc**g**ta**g**tcg

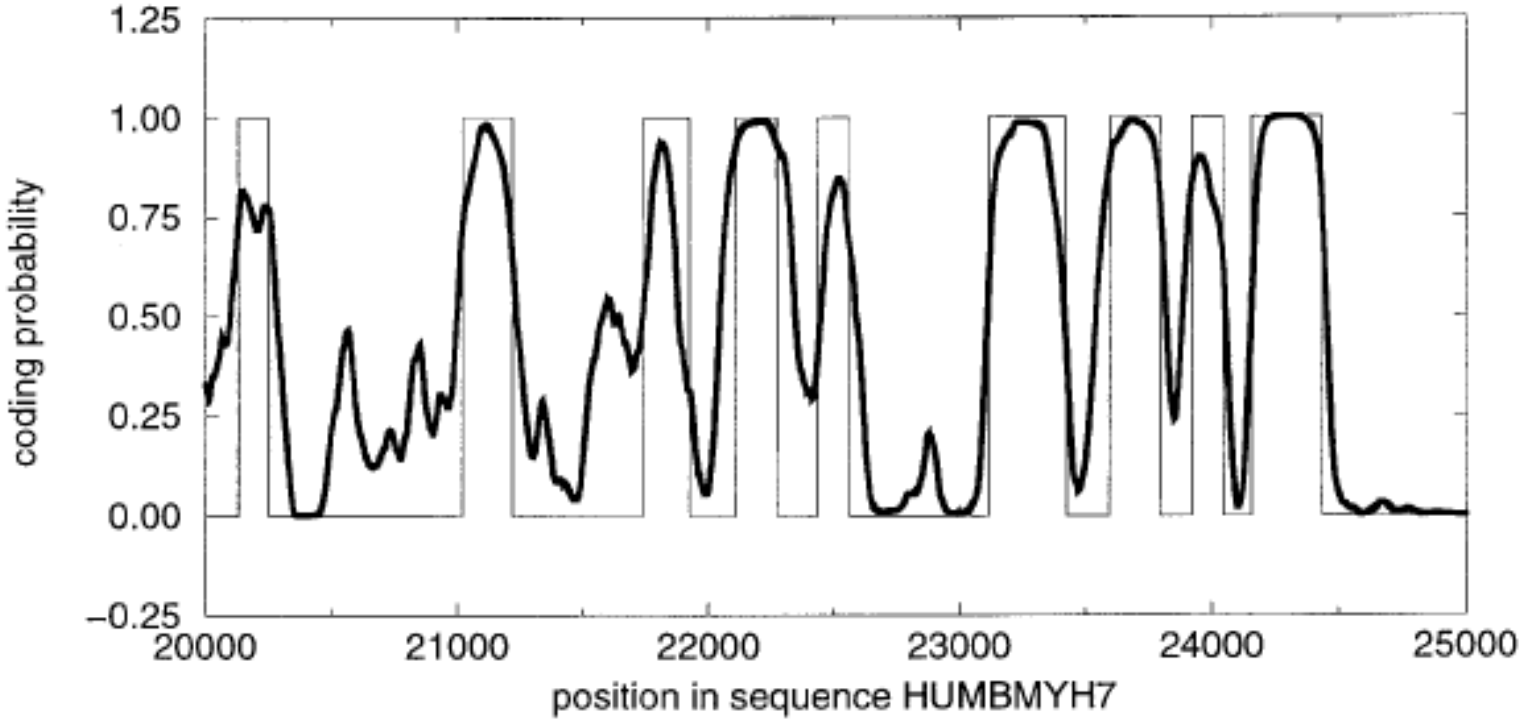
  n=3,6,9…
- n=2,5,8
- n=1,4,7

| | a | t | c | g |
|---|---|---|---|---|
| a | | | | |
| t | | | | |
| c | | | | |
| g | | | | |


primates
0.20
0.15
0.10
0.05
0.00
histogram
$10^{-6}$
$10^{-4}$
$10^{-2}$
$10^{0}$
$10^{2}$
mutual information in bit

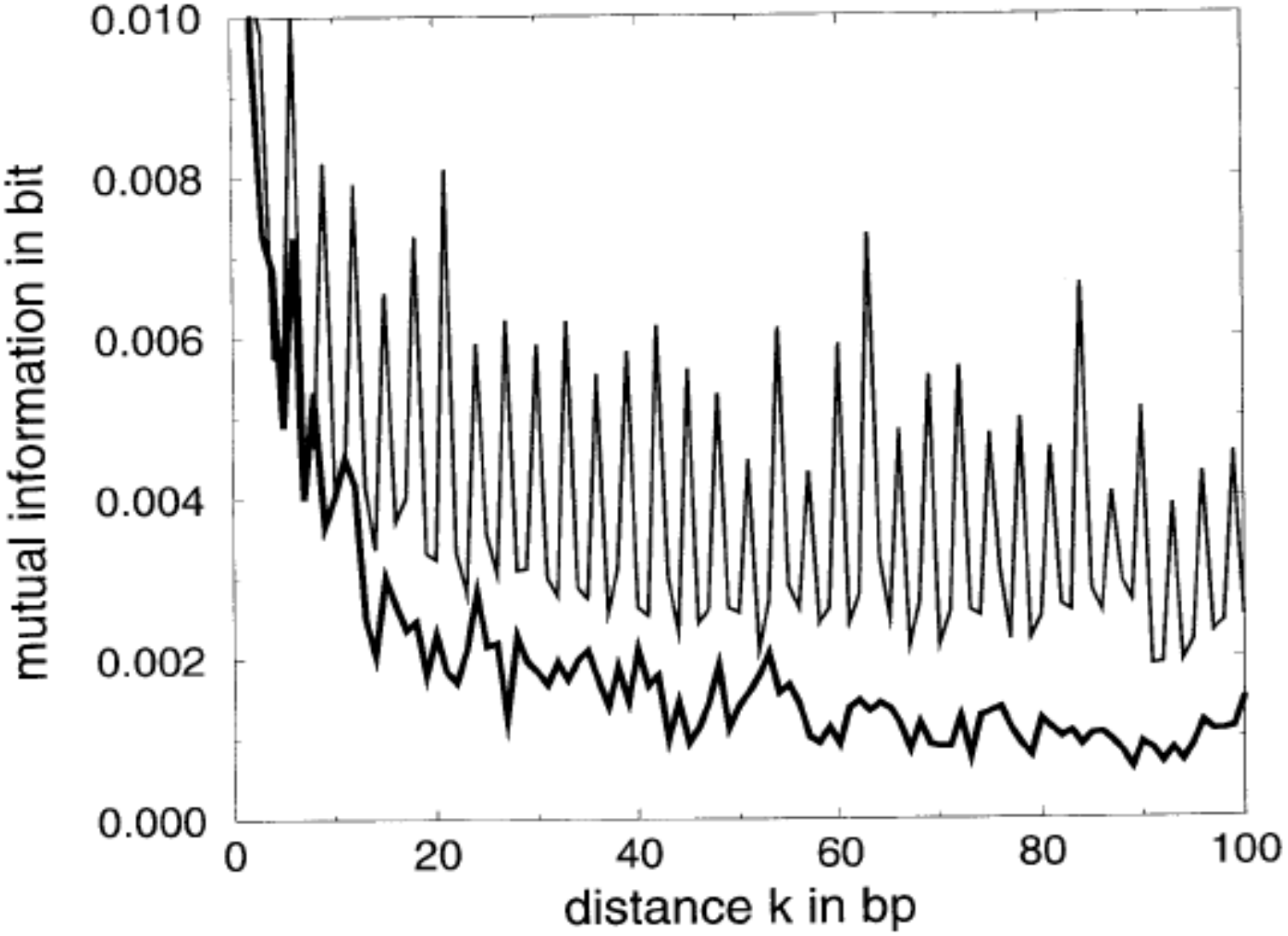

invertebrates
histogram
0.20
0.15
0.10
0.05
0.00
$10^{-6}$
$10^{-4}$
$10^{-2}$
$10^{0}$
$10^{2}$
mutual information in bit


plants
0.20
0.15
0.10
0.05
0.00
histogram
$10^{-6}$
$10^{-4}$
$10^{-2}$
$10^{0}$
$10^{2}$
mutual information in bit




Выделялось окно в 108 оснований. Строилось распределение для $I(l)$ для кодирующих последовательностей. Расчитывалось $\overline{I}$

Затем окном пробегалась изучаемая последовательность и для каждой позиции считалась $I(l)$ и рассчитывалась площадь $I < I(l)$


0.010
0.008
0.006
0.004
0.002
0.000
mutual information in bit
0
20
40
60
80
100
distance k in bp

# Распознование кодирующих областей

$$D = \sum_{i=1}^{64} \frac{(f_i - m_i)^2}{m_i}$$

Шульман, 1981

$f_i$ - наблюдаемая частота кодона $i$, $i$=1,…,64

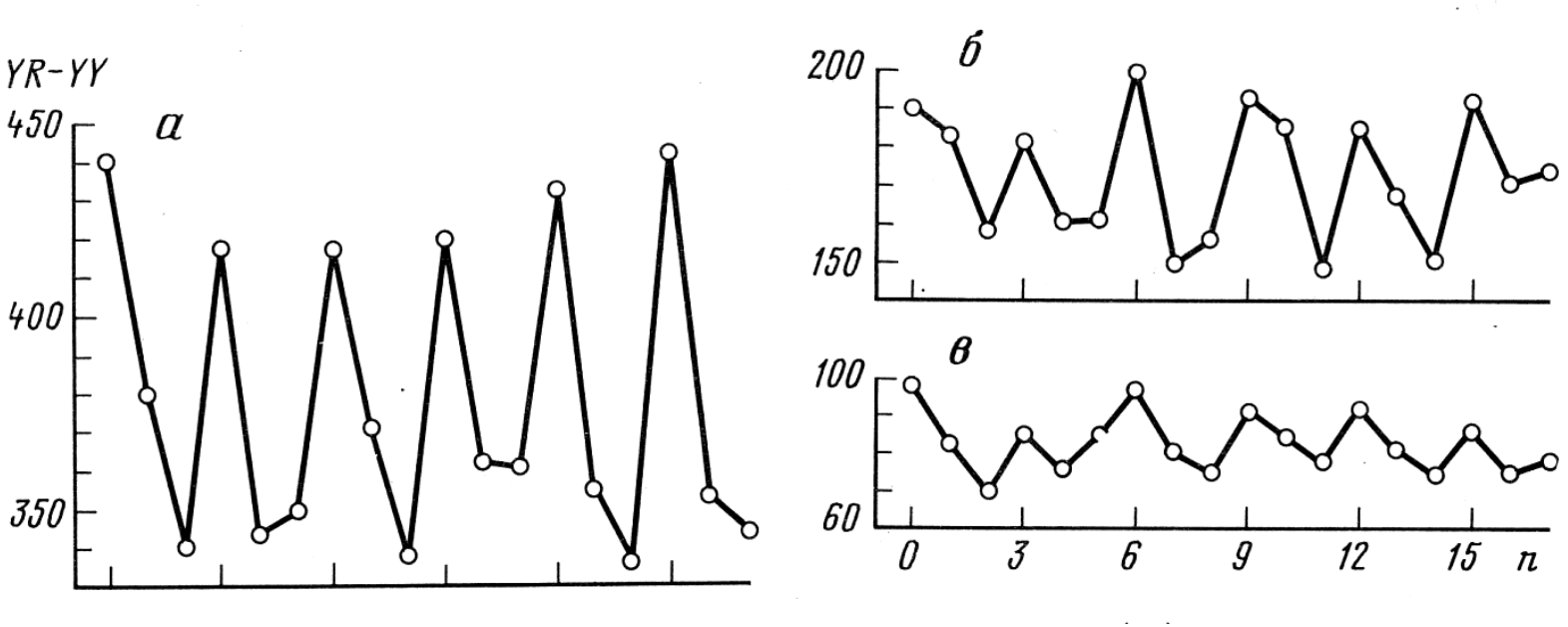
$m_i$ – ожидаемая частота кодона $i$, $i$=1,…,64

N – число кодонов, $P(a) = \frac{n(a)}{3N}$

$$m_i = p(a)\,p(b)\,p(c)N$$

# Распознование кодирующих областей

- RNY – закономерности, Шефферд, 1981
- RNYRNYRNYRNY…




Частота встречаемости $YR(N)_KYY$

а - ФХ174, б-E.coli, с-в геноме морского ёжа

# Распознование кодирующих областей

- Последовательность ДНК переводится в код R,Y
- Выделяются три рамки считывания
- Для каждой рамки подсчитывается число мутаций для перевода последовательности в код RNY
- Рамка с наименьшим числом мутаций – истинная
- $YR(N)_kYR$ - лучше

# Распознование кодирующих областей

- Метод Фиккетта

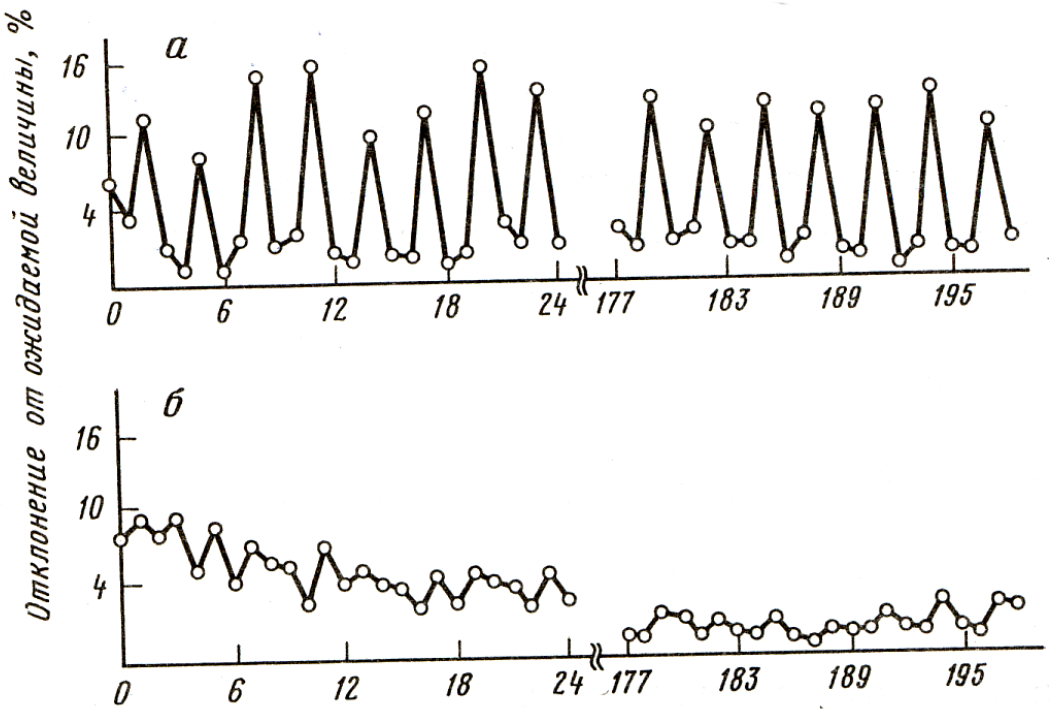



График автокорреляций встречаемости тимина $T(n)_kT$

а – кодирующие области

б – некодирующие области

# Распознование кодирующих областей

$$T_p = \frac{\max(T_1, T_2, T_3)}{\min(T_1, T_2, T_3) + 1}$$

$T_1$ – число тиминов в позициях 3n-2, n=1,2,3…
$T_2$ – число тиминов в позициях 3n-1, n=1,2,3…
$T_3$ – число тиминов в позициях 3n, n=1,2,3…

$A_f$, $T_f$, $C_f$, $G_f$ –частоты оснований *a*,*t*,*c*,*g* в рассматриваемой нуклеотидной последовательности

$$F = 0.33T_p + 0.18C_p + 0.26A_p + 0.31G_p + 0.14T_f + 0.12C_f + 0.11A_f + 0.15G_f$$

# Распознование кодирующих областей

- 0.32<F<1.37 на реальных последовательностях ДНК

Правила предсказания кодирующих свойств по функции Фиккетта

| Значение | Вероятность кодирования | Предсказания |
|---|---|---|
| От 0,32 до 0,43 | 0,00 | Не кодирует |
| " 0,43 " 0,53 | 0,04 | " |
| " 0,53 " 0,64 | 0,07 | " |
| " 0,64 " 0,74 | 0,29 | " |
| " 0,74 " 0,84 | 0,40 | Не ясно |
| " 0,84 " 0,95 | 0,77 | " |
| " 0,95 " 1,05 | 0,92 | Кодирует |
| " 1,05 " 1,16 | 0,98 | " |
| " 1,16 " 1,26 | 1,00 | " |
| " 1,26 " 1,37 | 1,00 | " |

# Распознование кодирующих областей

- Метод Стадена и Маклахана
- $Z=a_1b_1c_1a_2b_2c_2\ldots a_{n+1}b_{n+1}c_{n+1}$
- f(abc) – частота кодона abc в генах из данного организма
- В геноме выделены три рамки считывания. $Q_i$ – доля генов в *i*-ой рамке считывания

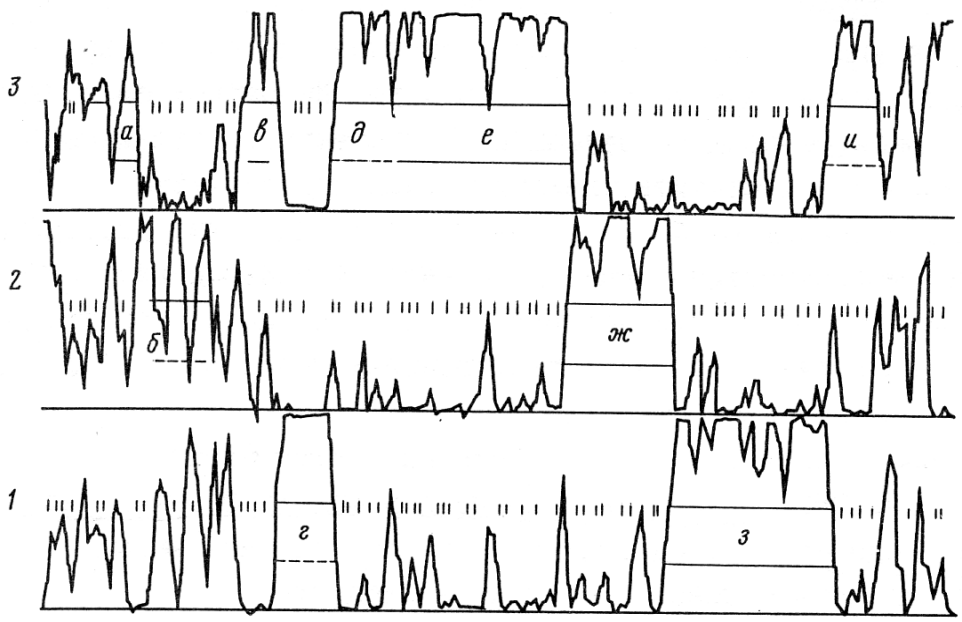
$$p_1 = Q_1 f(a_1b_1c_1)\ldots f(a_nb_nc_n)$$
$$p_2 = Q_2 f(b_1c_1a_2)\ldots f(b_nc_na_{n+1})$$
$$p_3 = Q_3 f(c_1a_2b_2)\ldots f(c_na_{n+1}b_{n+1})$$

# Распознование кодирующих областей

$$P_1 = \frac{p_1}{p_1 + p_2 + p_3}$$

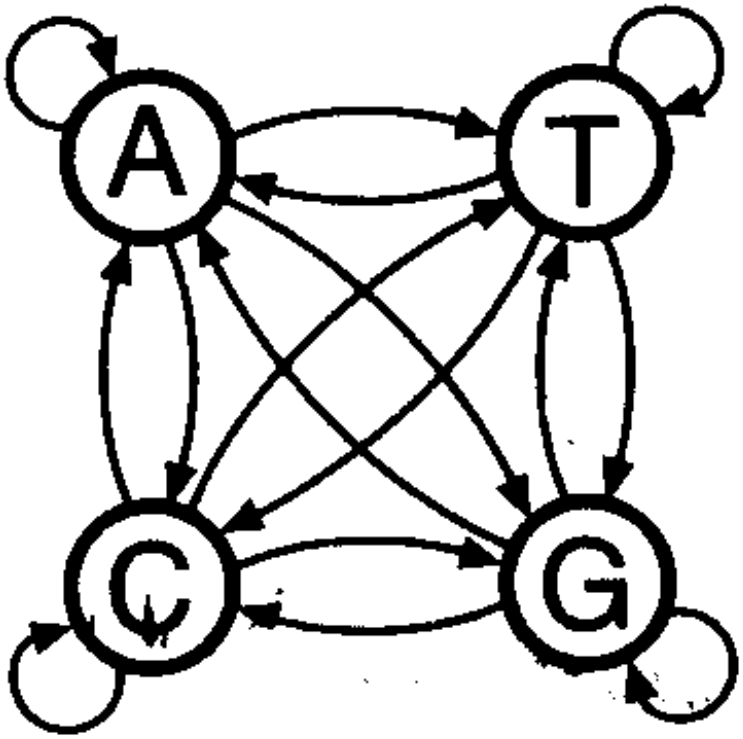
$$P_2 = \frac{p_2}{p_1 + p_2 + p_3}$$

$$P_3 = \frac{p_3}{p_1 + p_2 + p_3}$$




Применение "метода селекции кодонов" к atp(unc)-оперону E. coli
Отмечены гены: а - atp1; б - atpB; в - atpE; г - atpF; д - atpH; е - atpA; ж -atpG; з - atpD; и - atpC; 1,2,3 - номера рамок

# Марковские модели



# Цепи Маркова

- $\{E_1, E_2, \ldots, E_k\}$ - множество состояний некоторой системы.
- Изменение состояний происходит во время $t_1, t_2, \ldots, t_n, \ldots$
- Однородные цепи Маркова – переход из $E_i$ в $E_j$ зависит только от состояния $E_i$

# Матрица переходов

$$\mathrm{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \cdots & \cdots & & \cdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{bmatrix}$$

$$0 \le p_{ij} \le 1$$

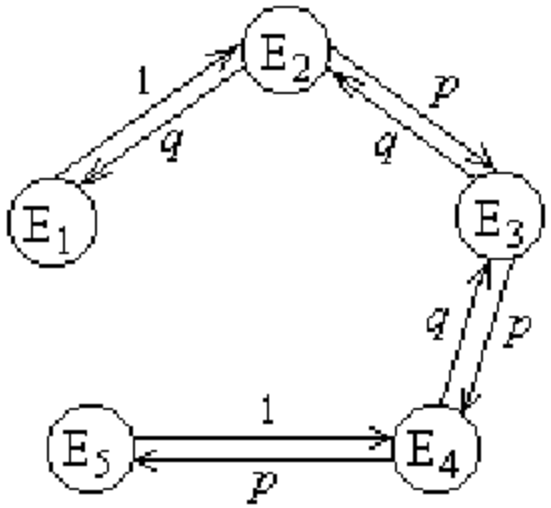
$$\sum_i p_{ij} = 1$$ для любого *i*

- Вектор $a=(a_1,a_2,\ldots,a_k)$ – вектор начальных вероятностей
- $a_i=P(E_i)$ – вероятность появления состояния $E_i$ в начальном испытании
- Матрица перехода за $n$ шагов равна $P^n$
- Если вероятность попасть из $E_i$ в $E_j$

  за n шагов не равна 0, то $E_j$ достижимо из $E_i$
- $E_i$ существенное состояние, если для каждого $E_j$, достижимого из $E_i$, $E_i$ достижимо из $E_j$. В противном случае состояние $E_i$ называется несущественным.

# Пример марковской цепи

- Состояния 1, 2, 3, 4, 5; в точках 1 и 5 находятся отражающие стенки.




Переход влево с вероятностью *q* и в право с вероятность *p*

# Матрица вероятностей переходов

$$
\mathrm{P} = \begin{array}{c} \\ E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \end{array}
\begin{array}{c}
\begin{array}{ccccc} E_1 & E_2 & E_3 & E_4 & E_5 \end{array} \\
\begin{bmatrix}
0 & 1 & 0 & 0 & 0 \\
q & 0 & p & 0 & 0 \\
0 & q & 0 & p & 0 \\
0 & 0 & q & 0 & p \\
0 & 0 & 0 & 1 & 0
\end{bmatrix}
\end{array}
$$

# Поиск GC богатых участков

atcgatcgcgcgtcgaacgcgattcgcgcacgtcgtacga

# Поиск GC богатых участков

$a^{+}_{st}$

| + | A | C | G | T |
|---|---|---|---|---|
| A | 0.180 | 0.274 | 0.426 | 0.120 |
| C | 0.171 | 0.368 | 0.274 | 0.188 |
| G | 0.161 | 0.339 | 0.375 | 0.125 |
| T | 0.079 | 0.355 | 0.384 | 0.182 |

$a^{-}_{st}$

| − | A | C | G | T |
|---|---|---|---|---|
| A | 0.300 | 0.205 | 0.285 | 0.210 |
| C | 0.322 | 0.298 | 0.078 | 0.302 |
| G | 0.248 | 0.246 | 0.298 | 0.208 |
| T | 0.177 | 0.239 | 0.292 | 0.292 |

$$\sum_t a^{+}_{st} = 1$$

$$\sum_t a^{-}_{st} = 1$$

$$S(x) = \log \frac{P(x|\text{model}+)}{P(x|\text{model}-)} = \sum_{i=1}^{L} \log \frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}}$$

$$= \sum_{i=1}^{L} \beta_{x_{i-1}x_i}$$

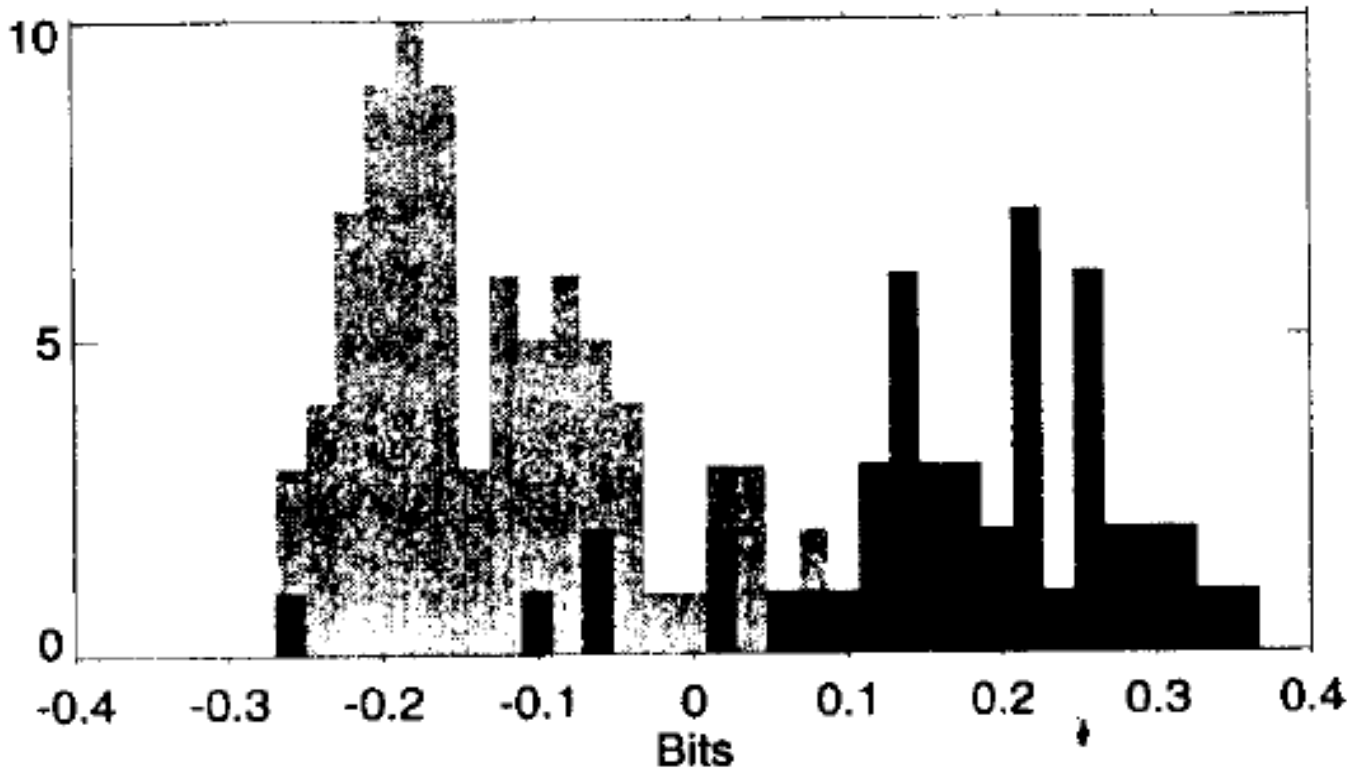
| $\beta$ | A | C | G | T |
|---|---|---|---|---|
| A | −0.740 | 0.419 | 0.580 | −0.803 |
| C | −0.913 | 0.302 | 1.812 | −0.685 |
| G | −0.624 | 0.461 | 0.331 | −0.730 |
| T | −1.169 | 0.573 | 0.393 | −0.679 |

# Подсчет веса последовательности

- atcgatgc – *s(i)*
- at,tc,cg,ga,tg,gc

- Перемешивается последовательность *s*(*i*) *N* раз.
- Определяется множество *w* для случайных последовательностей
- На множестве определяется $\bar{w}$ и *D*
- Рассчитывается $Z = \frac{w - \bar{w}}{\sqrt{D}}$

# Плотность распределения для *S(x)/L* (*L* – длина последовательности)



# Поиск кодирующих областей при помощи марковских моделей.

$$Z = a_1, a_2, \ldots . a_n, \quad a_i = T, C, A, G$$

***P(K|Z)*** **– вероятность того, что последовательность принадлежит к кодирующей области**

***P(N|Z)*** **вероятность того, что последовательность принадлежит к некодирующей области**

$$P(K|Z) = P(k_1|Z) + P(k_2|Z) + P(k_3|Z)$$

- Формируем две обучающие выборки :
  - Одна- некодирующие последовательности, другая кодирующие последовательности
- Рассчитываем три вектора начальных вероятностей $P^i(a)$, $a=T,C,G,A$
- Рассчитываем три матрицы переходных вероятностей $P^i(b|a)$ для трех рамок считывания. Эта вероятность встретить основание *b* в *i+1* позиции кодона при условии, что основание *a* присутствует в *i* ой позиции
- P(b|a)=P(ba)/P(a)

$$P(Z|H) = P(a_1) \cdot P(a_2|a_1) \cdot \ldots \cdot P(a_n|a_{n-1}).$$

$$P(Z|k1) = P^1(a_1) \cdot P^1(a_2|a_1) \cdot P^2(a_3|a_2) \cdot \ldots \cdot P^2(a_n|a_{n-1})$$
$$P(Z|k2) = P^2(a_1) \cdot P^2(a_2|a_1) \cdot P^3(a_3|a_2) \cdot \ldots \cdot P^3(a_n|a_{n-1})$$
$$P(Z|k3) = P^3(a_1) \cdot P^3(a_2|a_1) \cdot P^1(a_3|a_2) \cdot \ldots \cdot P^1(a_n|a_{n-1})$$

***P(Z|H)*** – вероятность случайного обнаружения фрагмента Z в некодирующей области

# Формула Байеса

где

- $P(A)$ — априорная вероятность гипотезы $A$;
- $P(A \mid B)$ — вероятность гипотезы $A$ при наступлении события $B$ (апостериорная вероятность);
- $P(B \mid A)$ — вероятность наступления события $B$ при истинности гипотезы $A$;
- $P(B)$ — вероятность наступления события $B$.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(B) = \sum_{i=1}^{N} P(A_i)P(B|A_i)$$

Формула Байеса позволяет «переставить причину и следствие»: по известному факту события вычислить вероятность того, что оно было вызвано данной причиной.

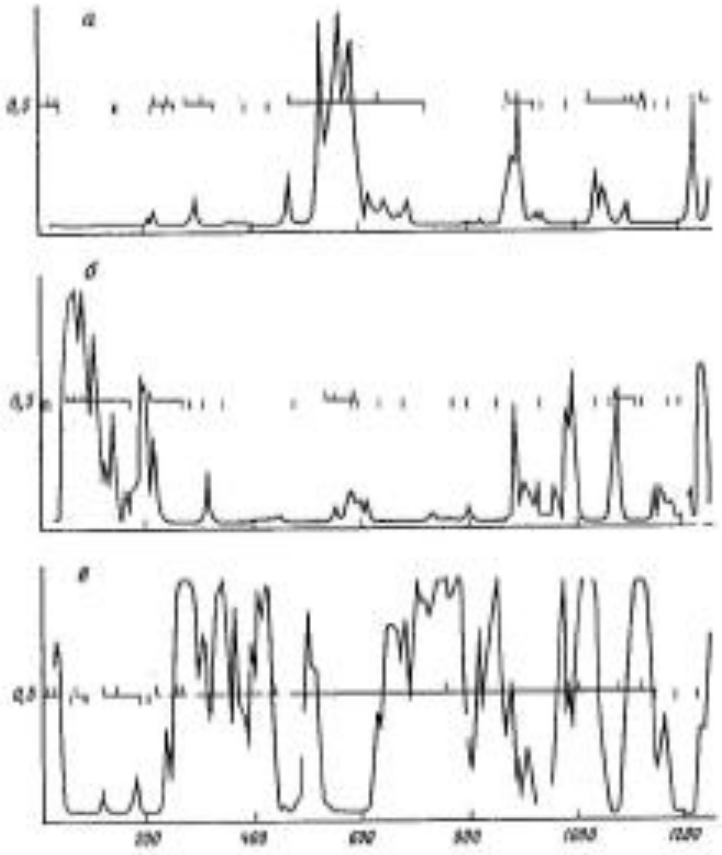
Применяем формулу Байеса:

$$P(k_1|Z) = \frac{P(Z|k_1)\cdot P(k_1)}{\sum P(Z|k_i)\cdot P(k_1)+P(Z|H)P(H)}$$

**P(H) и P($k_i$) являются априорными вероятностями.**

**P(H)=0.5**

**P($k_i$)=1/6**




Р и с. 3.8. Графики функций-индикаторов кодирующих областей для последовательности ECARAC в трех рамках считывания (а-в)

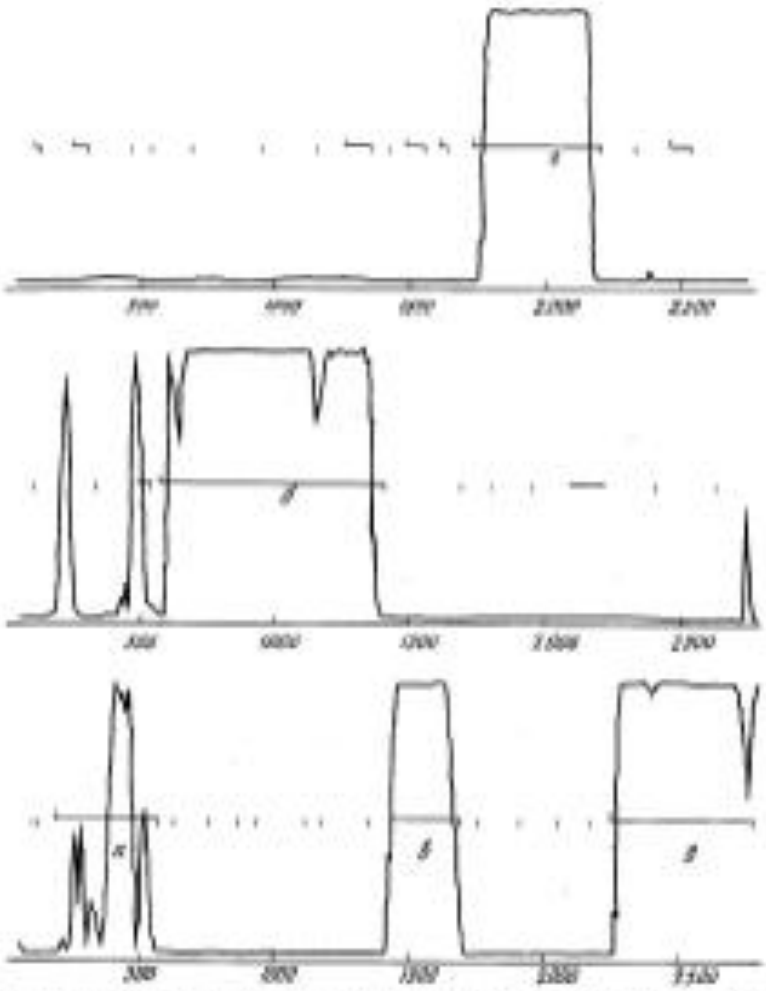


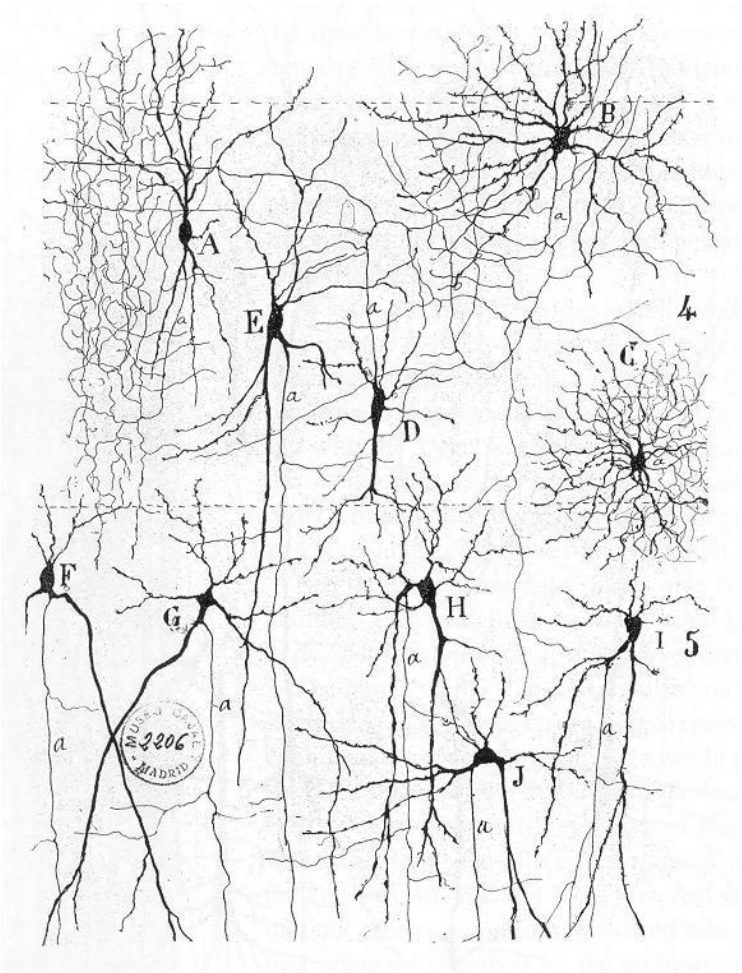
Рис. 3.12. Применение метода марковских цепей для последовательности ВСАТРХ в случае $\Psi=(2,32)$
Отмечены типы а - atp1; б - atp2; в - atp0; г - atpP; д - atpE

# Нейронные сети

- Предпосылка:
  - Известно, что биологические системы (люди, животные) прекрасно справляются со сложными задачами распознавания образов;
- Основная идея:
  - Применить знания о работе мозга (людей, животных) для решения задач распознавания образов;

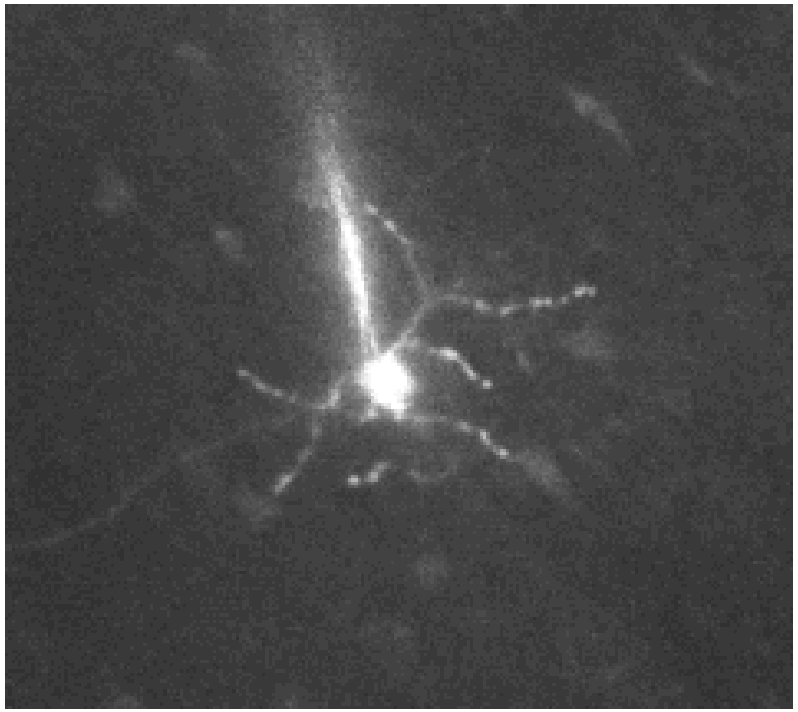
# Биологические нейронные сети

- 1872-1895 гг.
  - Понятие нейрона и нейронной сети;
  - Первые предположения о принципе работы;



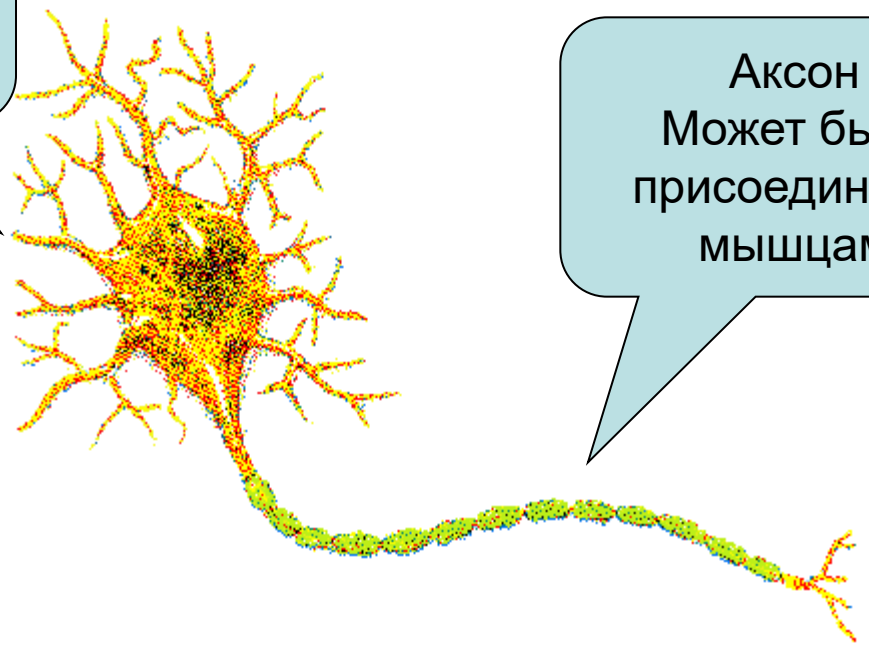
# Биологический нейрон

# Биологический нейрон



# Биологический нейрон
# Передача импульса

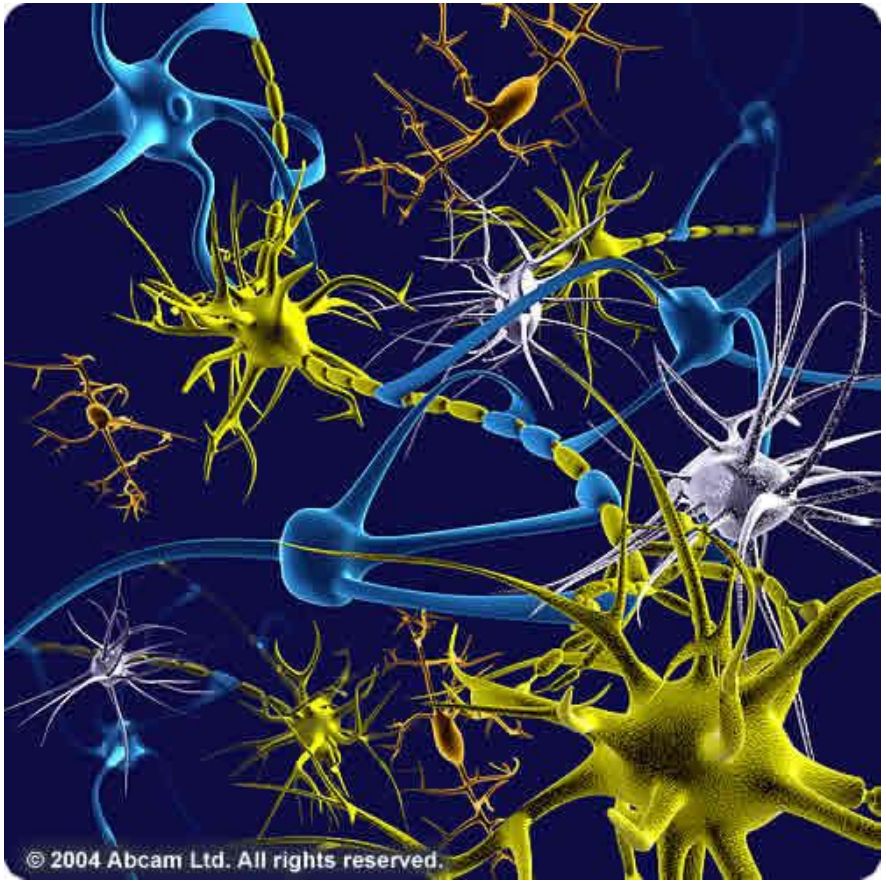
Дендриты
Например, могут быть присоединены к рецепторам

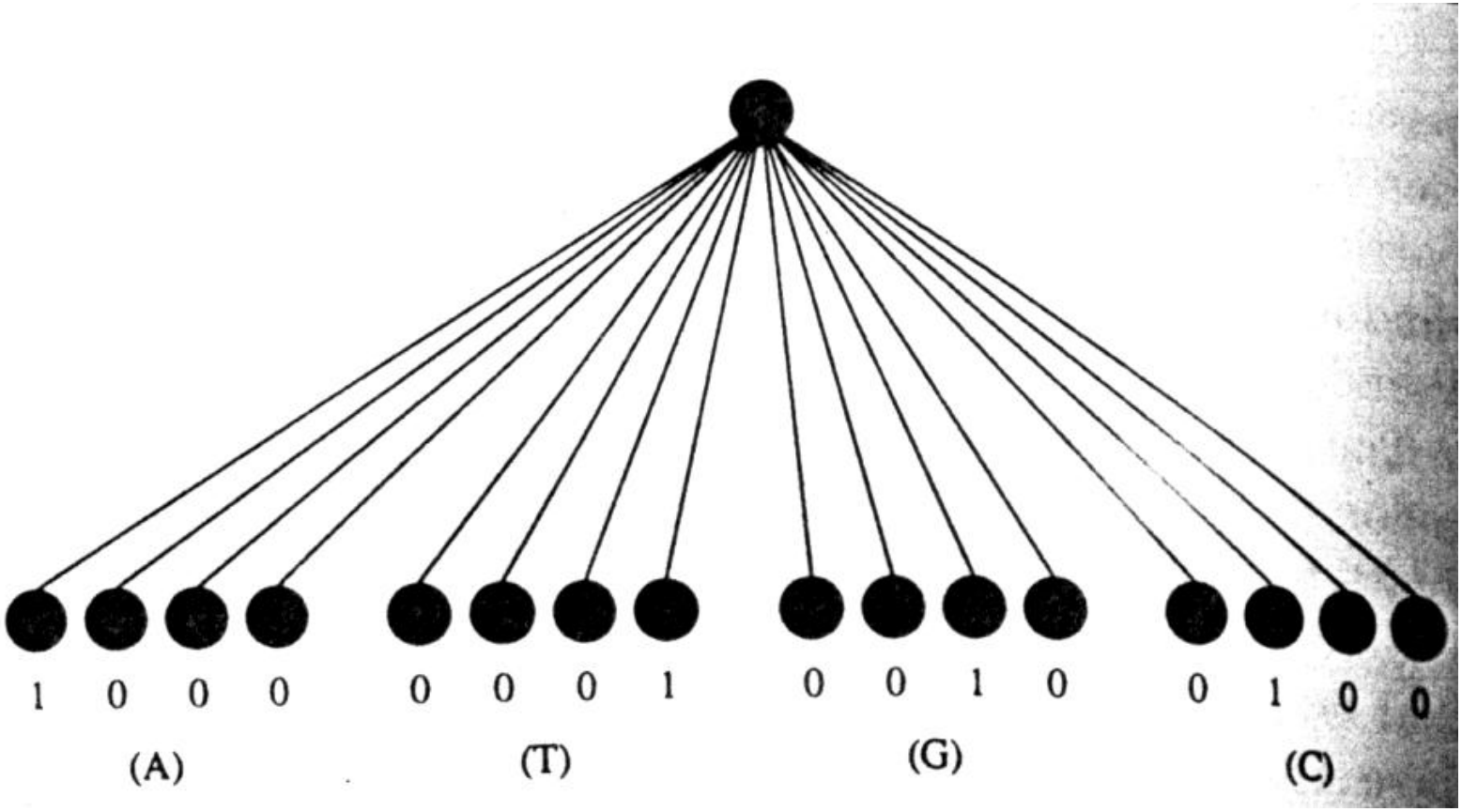


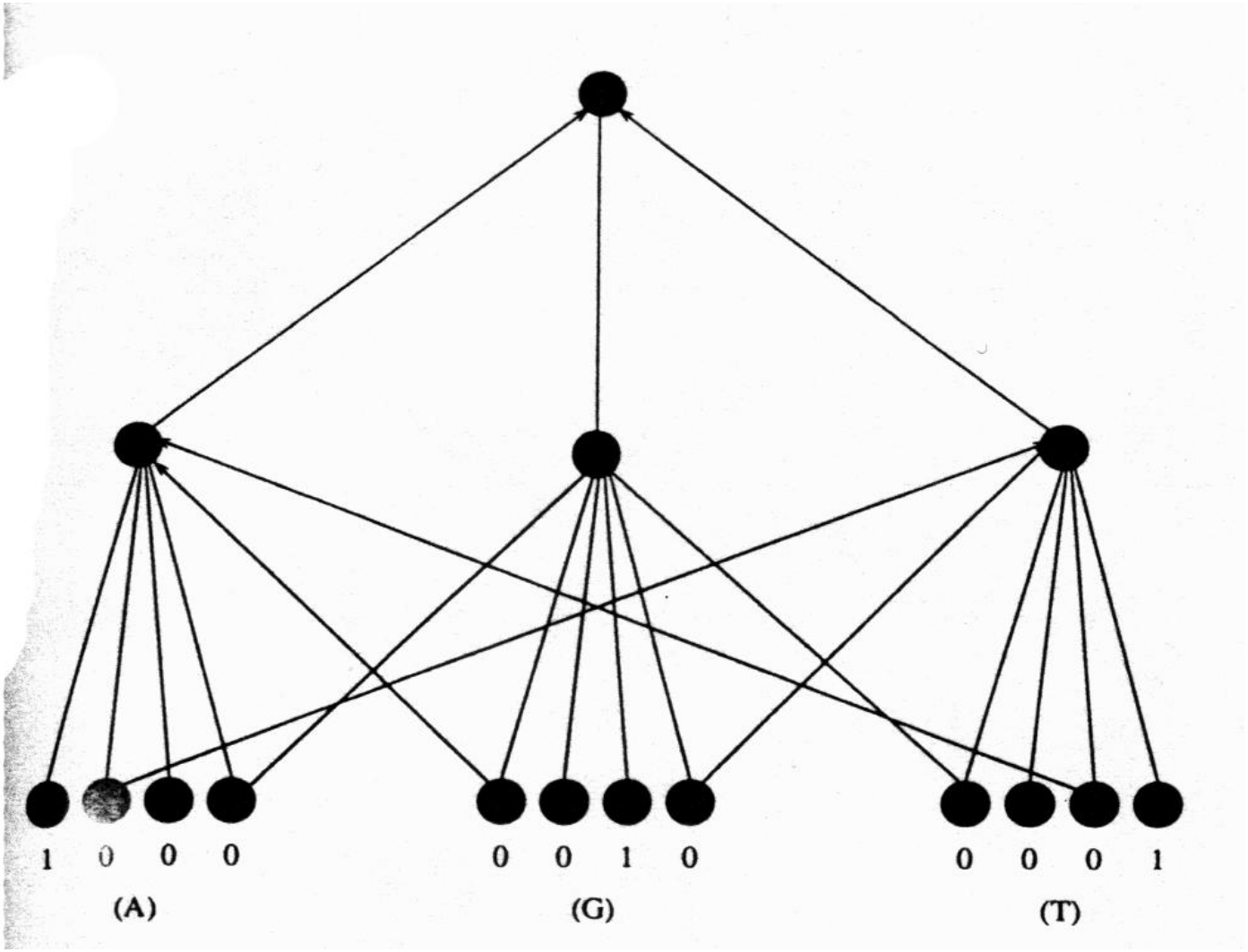
Аксон
Может быть присоединен к мышцам

# Нейронная сеть

- Совокупность соединенных между собой нейронов;
- Сеть осуществляет преобразование входного сигнала с рецепторов в выходной, являющейся реакцией организма на внешнюю среду




1 0 0 0
0 0 0 1
0 0 1 0
0 1 0 0
(A)
(T)
(G)
(C)


1 0 0 0
(A)
0 0 1 0
(G)
0 0 0 1
(T)

$$\mathrm{Out} = g\left(\sum_j T_j I_j + \theta\right), \qquad (2)$$

$I_j$ – входные величины (0 или 1)

OUT- состояние выходного нейрона

$T_j$ – численные веса, определяемые алгоритмом обучения

Θ – добавочная константа, определяемая в ходе обучения

g(x) – монотонная функция изменяющаяся от 0 (отрицательные x) до 1 (положительные x)

$$g(x) = 0{\cdot}5(1 + \tanh(x)), \qquad (3)$$

$$E = \sum_{p} (t^{(p)} - \mathrm{Out}^{(p)})^2 C^{(p)}. \qquad (4)$$

E – выходной сигнал

$C^{(p)}$ – веса последовательностей для обучения

+1 – для кодирующих районов

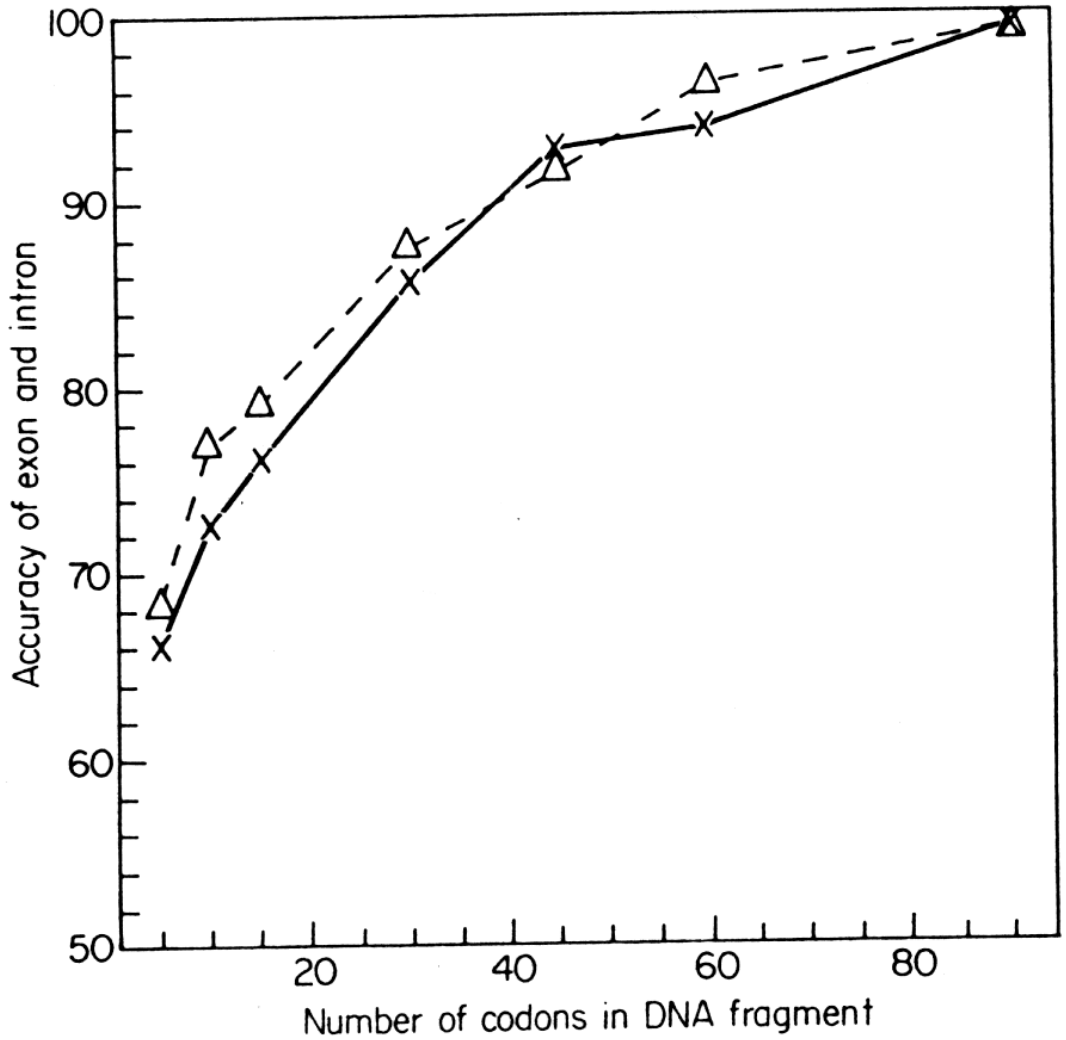
+0.1 для некодирующих участков

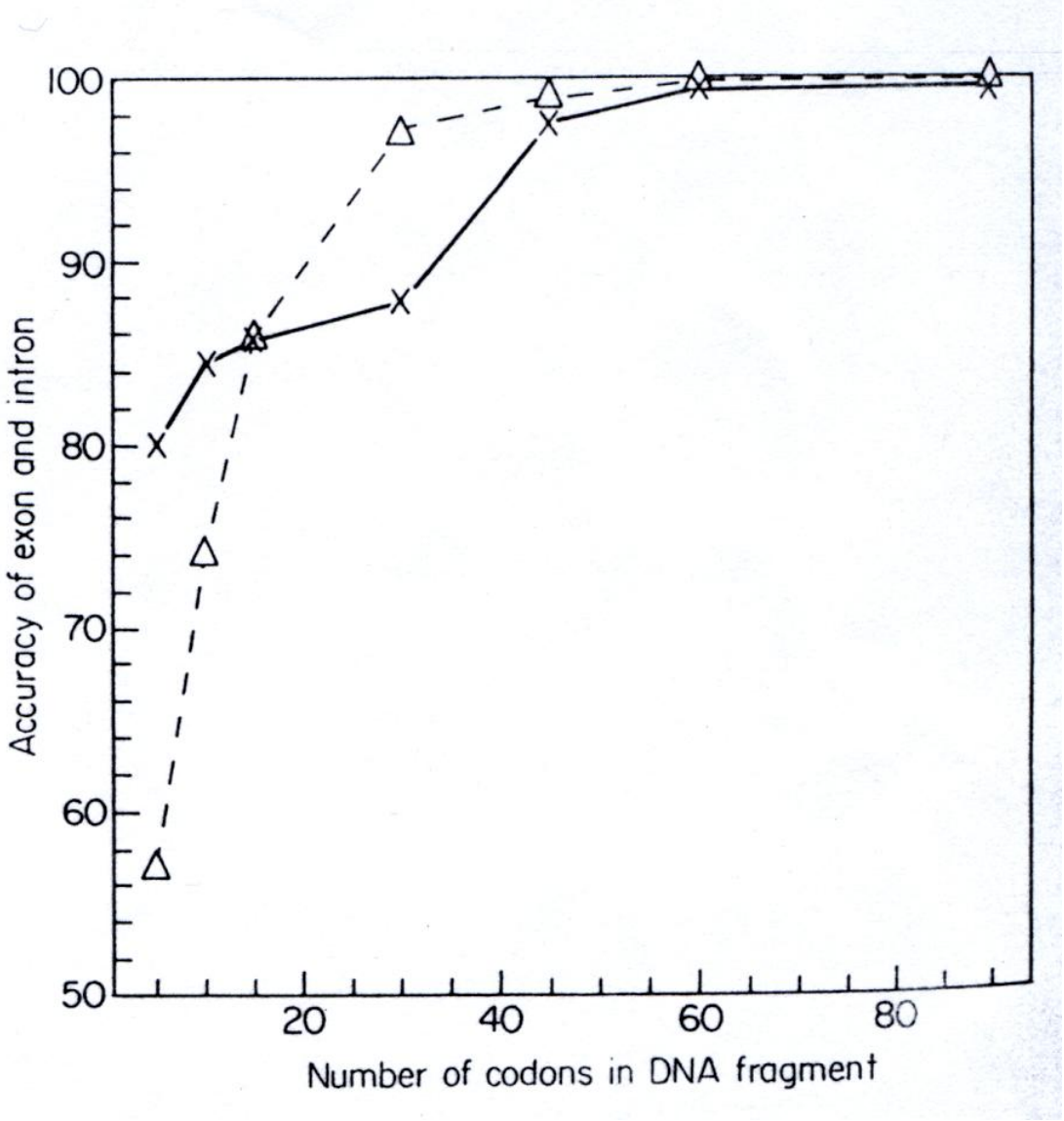
Задача - минимизировать E.

Используется градиентный метод

$$\Delta T_j = - \frac{\partial E}{\partial T_j} \times \varepsilon,$$

$$\Delta \theta = - \frac{\partial E}{\partial \theta} \times \varepsilon,$$

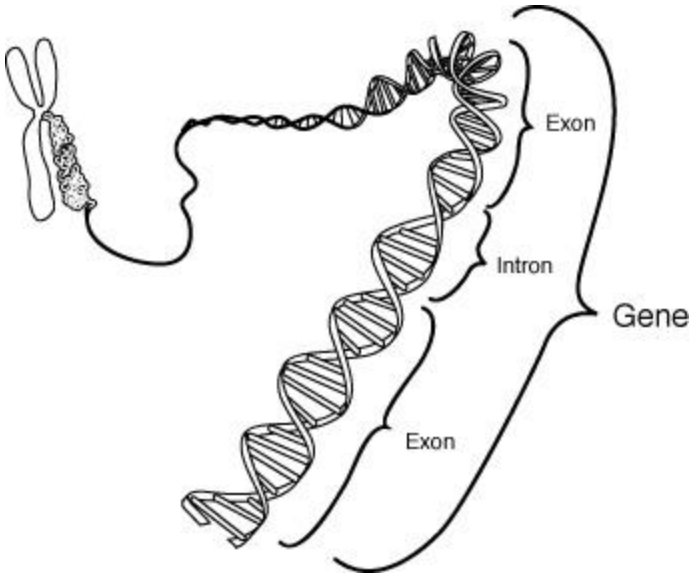

100
90
80
70
60
50
Accuracy of exon and intron
20
40
60
80
Number of codons in DNA fragment

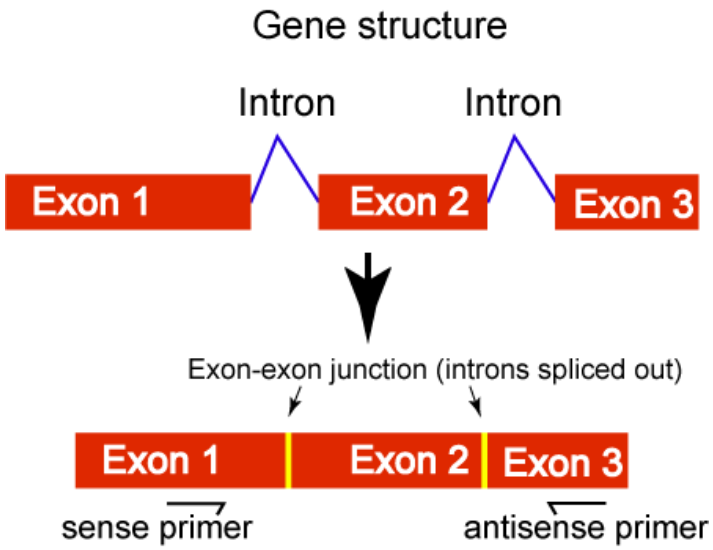



Использование дикодоновых частот

**GENSCAN** webserver at MIT is a great tool for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.
**GENEID** a program to predict genes, exons, splice sites and other signals along a DNA sequence.
**JIGSAW** a program that predicts gene models using the output from other annotation software. It uses a statistical algorithm to identify patterns of evidence corresponding to gene models.
**Artemis** is a free DNA sequence viewer and annotation tool that allows visualisation of sequence features and the results of analyses within the context of the sequence, and its six-frame translation.
**AUGUSTUS** is an open source program that predicts genes in eukaryotic genomic sequences.It has a protein profile extension (PPX) which allows to use protein family specific conservation in order to identify members and their exon-intron structure of a protein family given by a block profile.By incorporating mRNA alignments, EST alignments, conservation and other sources of informationcan predict alternative splicing and alternative transcripts, the 5'UTR and 3'UTR including introns.
**EuGene** is an open integrative gene finder for eukaryotic and prokaryotic genomes- it is characterized by its ability to simply integrate arbitrary sources of information in its prediction process, including RNA-Seq, protein similarities, homologies and various statistical sources of information.
**PseudoPipe** is a stand alone computational pipeline for pseudogene annotation.
**FusionSeq** is a computational framework to identify fusion transcripts from paired-end RNA-sequencing.

# Exon prediction in Eukaryotic DNA using Genescan: Net result is a protein sequence




GeneScan looks for start and stop codons, promoters, splice sites, polyA tails, provides statistics for coding potential