

## Метод 1

Нуклеотидам А, G, С і Т ставляться у відповідність вектори: А (1, 0.8), G (1, 0.6), С (1, 0.4), Т (1, 0.2). Послідовність отримуємо, сумуючи вектори, що відповідають нуклеотидам з послідовності.

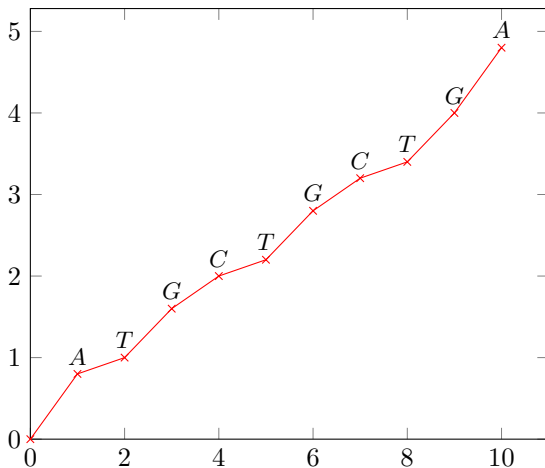


Рис.: Графічне представлення послідовності ATGCTGCTGA

## Метод 2

Спочатку використаємо попередній метод щоб отримати числову послідовність  $(x_i, y_i)$ . Далі використаємо наступну формулу для обчислення результуючої послідовності:

$$\frac{x_i - \vec{y_i}}{\frac{1}{2}n(n+1) - y_n},$$

де  $\vec{y_i}$  це  $y$ -компонента вектора, що відповідає  $i$ -тому нуклеотиду при використанні методу 1,  $n$  це розмір ДНК послідовності.

## Метод 3

Нуклеотидам А, G, C, Т ставимо у відповідність вектори  $(-1, 0)$ ,  $(1, 0)$ ,  $(0, 1)$ ,  $(0, -1)$ . Починаємо з точки  $(0, 0)$  і рухаємось по відповіднім векторам. Точки через які ми проходимо утворюють послідовність, причому точка стільки разів зустрічається у послідовності, скільки разів ми в неї потрапили.

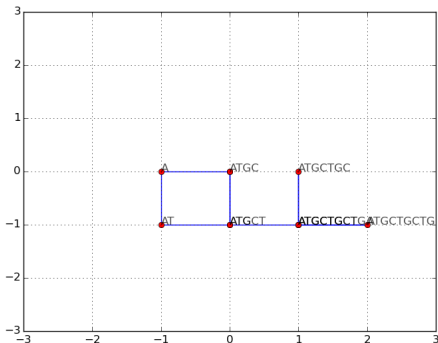


Рис.: Графічне представлення послідовності ATGCTGCTGA

## Метод 4

Розташовуємо нуклеотиди у вершинах квадрата зі стороною 1:  $A=(0,0)$ ,  $G=(1,1)$ ,  $C=(0,1)$ ,  $T=(1,0)$ . Координати послідовності рахуються ітеративно, рухаючись на половину відстані між попередньою позицією і точкою квадрата, якій відповідає наступний нуклеотид у напрямку цієї точки. Ітеративну процедуру можна задати наступним чином:

$$p_i = p_{i-1} - 0.5(p_{i-1} - g_i)$$

$$i = 1, \dots, n; p_0 = (0.5, 0.5),$$

де  $g_i$  - координати, що відповідають  $i$ -тому нуклеотиду,  $n$  - довжина послідовності ДНК.

## Метод 4

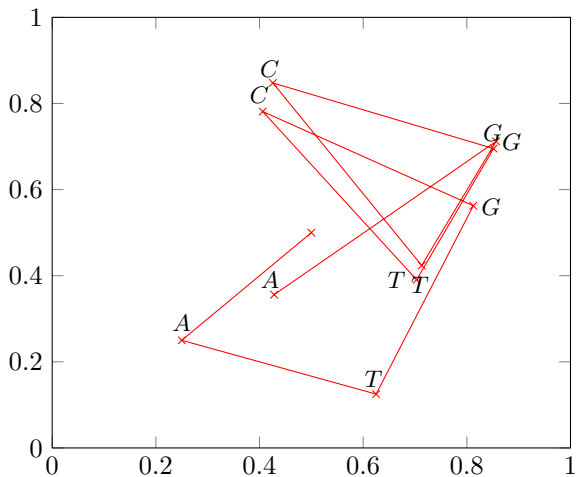


Рис.: Графічне представлення послідовності ATGCTGCTGA

## Метод 5,6

Використовуємо попередній метод, щоб отримати послідовність  $p_i$ , отримуємо результуючу, як суму всіх попередніх:

$$z_i = \sum_{j=1}^i p_i$$

Отримуємо за допомогою методу 4 послідовність  $p_i$  і, щоб отримати результуючу послідовність, кожній точці ставимо у відповідність число:

$$z_i = x_i + y_i,$$

де  $p_i = (x_i, y_i)$ .

## Метод 7

Нуклеотидам А,С ставимо у відповідність  $-1$ , а нуклеотидам Т,Г ставимо у відповідність  $1$ . Починаючи з точки  $0$  рухаємось ітеративно:

$$p_i = p_{i-1} - \frac{(g_i - p_{i-1})}{2} \text{sign}(g_i)$$

де  $g_i$  число яке відвідає  $i$ -тому нуклеотиду. Тобто ми, подібно до методу 4, рухаємося на пів відстань до числа яке відповідає  $i$ -тому нуклеотиду.

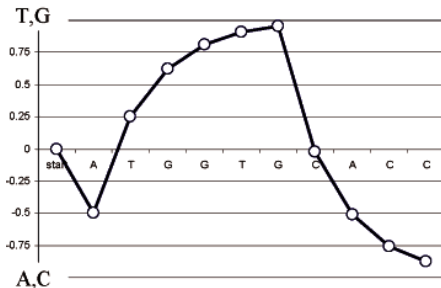


Рис.: Графічне представлення послідовності ATGGTGCACC

## *p*-СТАТИСТИКА

$G$  і  $G'$ -генеральні сукупності.  $x = (x_1, \dots, x_n) \in G$  і  $x' = (x'_1, \dots, x'_m) \in G'$ ,  $x_{(1)} < \dots < x_{(n)}$ ,  $x'_{(1)} < \dots < x'_{(n)}$  - порядкові статистики. Припустимо, що  $F_G(u) = F_{G'}(u)$ .

$A_{ij}^{(k)} = \{x'_k \in (x_{(i)}, x_{(j)})\}$ . Якщо  $F_G(u) = F_{G'}(u)$ :

$$P\left(A_{ij}^{(k)}\right) = P\left(x'_k \in (x_{(i)}, x_{(j)})\right) = p_{ij}^{(n)} = \frac{j-i}{n+1} = \frac{q}{n+1}, q = j-i$$

$$p_{ij}^{(1)} = \frac{h_{ij}^{(n)} m + g^2/2 - g\sqrt{h_{ij}^{(n)}(1-h)m + g^2/4}}{m + g^2}, \quad (1)$$

$$p_{ij}^{(2)} = \frac{h_{ij}^{(n)} m + g^2/2 + g\sqrt{h_{ij}^{(n)}(1-h)m + g^2/4}}{m + g^2}, \quad (2)$$

де  $h_{ij}^{(n)}$  — частота події  $A_{ij}^{(n)}$  в  $m$  випробуваннях.  $N$  кількість інтервалів  $I_{ij}^{(n,m)} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$  ( $N = n(n-1)/2$ ) і  $L$  - кількість інтервалів  $I_{ij}^{(n,m)}$ , які містять ймовірності  $p_{ij}^{(n)}$ .

$h^{(n,m)} = \rho(x, x') = \frac{L}{N}$  будемо називати *p*-статистикою.



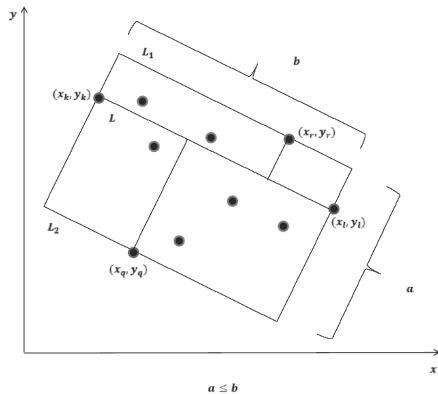
# Модифікована $p$ -статистика

Нехай  $t(x_k)$  - кратість вибіркового значення  $x_k$

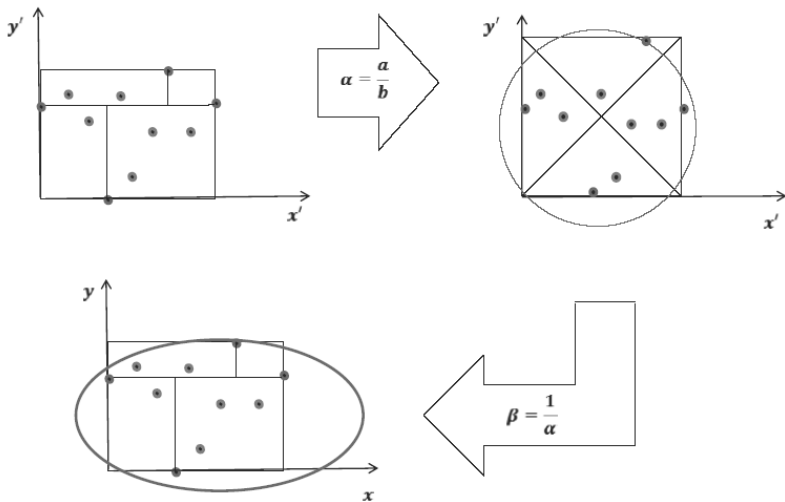
$$p_{ij} = p(A_{ij}) = p(x^* \in (x_{(i)}, x_{(j)})) \approx \gamma_i + \gamma_{i+1} + \dots + \gamma_{j-1} + \frac{j-i}{n+1},$$

$$\gamma_l = \gamma(x_{(l)}) = \frac{t(x_{(l)}) - 1}{n+1},$$

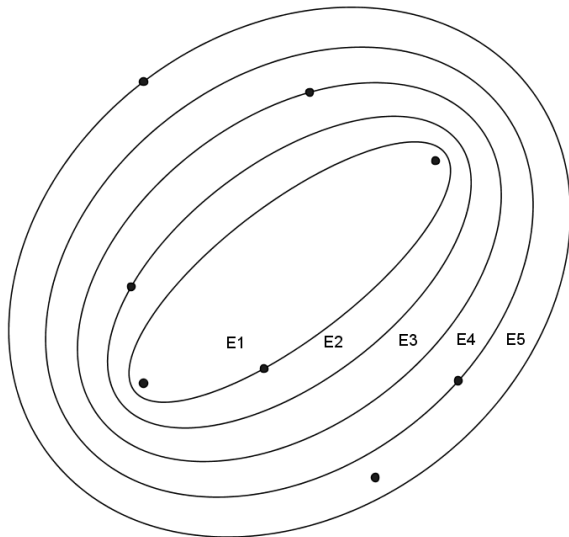
# Еліпси Петуніна



# Еліпси Петуніна



# Еліпси Петуніна



# Еліпси Петуніна

При побудові  $p$ -статистики варіаційному ряду вибірок  $\vec{x}_{(1)} \preceq \vec{x}_{(2)} \preceq \dots \preceq \vec{x}_{(n)}$  поставимо у відповідність послідовність вкладених еліпсоїдів  $E_{(1)} \subset E_{(2)} \subset \dots \subset E_{(n)}$ . Ймовірність того, що елемент  $\vec{x}$  із генеральної сукупності  $G$  задовольняє умові  $\vec{x}_{(i)} \preceq \vec{x} \preceq \vec{x}_{(j)}$ , рівна ймовірності потрапити між еліпсами  $E_{(i)}$  і  $E_{(j)}$ , тобто  $\frac{j-i}{n+1}$ . Ця умова дозволяє побудувати  $p$ -статистику для багатовимірною випадку.

Таким чином будуємо  $p$ -статистику як завжди, тільки подія  $A_{ij}^{(k)}$  буде полягати в тому, що  $x'_k$  попаде в область  $E_{(j)} \setminus E_{(i)}$ .

# Результати

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.009	0.009	0.0045	0.01365	0.01811
rat	-	-	0.00869	0.00633	0.00797	0.0088
rabbit	-	-	-	0.00619	0.00996	0.011
human	-	-	-	-	0.00595	0.00441
duck	-	-	-	-	-	0.01754
gorilla	-	-	-	-	-	-

Табл.:  $p$ -статистики при представленні ДНК методом 1

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.03221	0.03896	0.0177	0.07635	0.09232
rat	-	-	0.30299	0.07789	0.06678	0.05958
rabbit	-	-	-	0.05915	0.09386	0.08013
human	-	-	-	-	0.03564	0.03172
duck	-	-	-	-	-	0.72326
gorilla	-	-	-	-	-	-

Табл.:  $p$ -статистики при представленні ДНК методом 2

# Результати

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.01839	0.01839	0.00922	0.02752	0.04045
rat	-	-	0.03393	0.01004	0.02045	0.0182
rabbit	-	-	-	0.0105	0.02045	0.0182
human	-	-	-	-	0.00821	0.00912
duck	-	-	-	-	-	0.03174
gorilla	-	-	-	-	-	-

Табл.:  $p$ -статистики при представленні ДНК методом 3

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.20654	0.20122	0.18473	0.28257	0.32461
rat	-	-	0.15438	0.12881	0.22096	0.20542
rabbit	-	-	-	0.1298	0.2198	0.21226
human	-	-	-	-	0.16495	0.16271
duck	-	-	-	-	-	0.22192
gorilla	-	-	-	-	-	-

Табл.:  $p$ -статистики при представленні ДНК методом 4

# Результати

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.009	0.00902	0.0045	0.01367	0.01821
rat	-	-	0.00877	0.00635	0.00801	0.00882
rabbit	-	-	-	0.0062	0.01	0.01101
human	-	-	-	-	0.00595	0.00446
duck	-	-	-	-	-	0.01759
gorilla	-	-	-	-	-	-

Табл.:  $p$ -статистики при представленні ДНК методом 5

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.41829	0.46788	0.18442	0.35209	0.46493
rat	-	-	0.995	0.34451	0.50635	0.43534
rabbit	-	-	-	0.37162	0.52875	0.60944
human	-	-	-	-	0.8112	0.38868
duck	-	-	-	-	-	0.49541
gorilla	-	-	-	-	-	-

Табл.:  $p$ -статистики при представленні ДНК методом 6



# Результати

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.59803	0.56986	0.90713	0.88857	0.94818
rat	-	-	0.99778	0.56047	0.39723	0.40133
rabbit	-	-	-	0.42243	0.38003	0.3786
human	-	-	-	-	0.59813	0.77453
duck	-	-	-	-	-	0.95202
gorilla	-	-	-	-	-	-

Табл.:  $p$ -статистики при представленні ДНК методом 7

# Висновки

Алгоритм побудови  $p$ -статистик для послідовностей ДНК досить повільний. Тому не доцільно його застосовувати при послідовностях, довжина яких перевищує  $10^3$ .

Результати суперечать інтуїтивному уявленню, про зв'язок близькості геномних послідовностей і міжвидової близькості.

# Література



Chenglong Yu, Mo Deng, Stephen S.-T. Yau, DNA sequence comparison by a novel probabilistic method, Information Sciences 181 (2011) 1484–1492



Dorota Bielinska-Waz, Timothy Clark, Piotr Waz, Wiesław Nowak, Ashesh Nandy, 2D-dynamic representation of DNA sequences, Chemical Physics Letters 442 (2007) 140–144



Wei Deng and Yihui Luan, Hindawi Publishing Corporation, Analysis of Similarity/Dissimilarity of DNA Sequences Based on Chaos Game Representation, Abstract and Applied Analysis, Volume 2013, Article ID 926519, 6 pages, <http://dx.doi.org/10.1155/2013/926519>

# Література



Jure Zupan and Milan Randic, Algorithm for Coding DNA Sequences into “Spectrum-like” and “Zigzag” Representations, J. Chem. Inf. Model. 2005, 45, 309-313



Д. А. Ключин, Ю. И. Петунин, Непараметрический Критерий Эквивалентности Генеральных Совокупностей, основанный На Мере Близости Между Выборками, ДК 519.21



Д. А. Ключин, М. В. Присяжная, Многомерное ранжирование с помощью эллипсов Петунина, Журнал обчисл. та прикл. матем. № 4(114) 2013, стор. 1-7, УДК 519.71



Дмитро А. Ключин, Міра близькості між виборками, що містять атоми, Вісник Київського університету, Серія: фізико-математичні науки, 2005, 3, УДК 519.9