Analiza szeregów czasowych. Dalekozasięgowe zależności w obrębie sekwencji nukleotydowych i aminokwasowych

Adrian Kania

Zakład Biofizyki Obliczeniowej i Bioinformatyki
Wydział Biochemii, Biofizyki i Biotechnologii, Uniwersystet Jagielloński

Podsumowanie

Etapy:

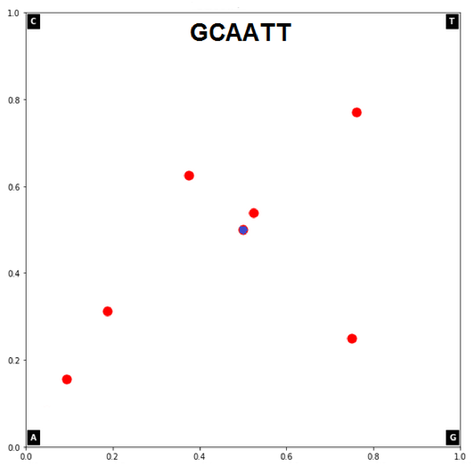
- Reprezentacja numeryczna sekwencji biologicznych (Chaos Game Representation)
- Analiza w oparciu o metody przeznaczone do szeregów czasowych (współczynnik Hursta)
- Klasyfikacja organizmów w oparciu o powyższe charakterystyki.

Chaos Game Representation

Niech $(x_1, y_1) = (0.5, 0.5)$ oraz $N_A = (0, 0)$, $N_T = (1, 1)$, $N_C = (0, 1)$, $N_G = (1, 0)$. Kolejne współrzędne określone są rekurencyjnie przez:

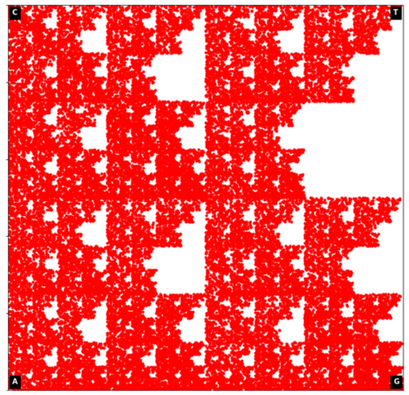
$$(x_{i+1}, y_{i+1}) = (x_i, y_i) - 0.5 \cdot ((x_i, y_i) - N_{i+1})$$

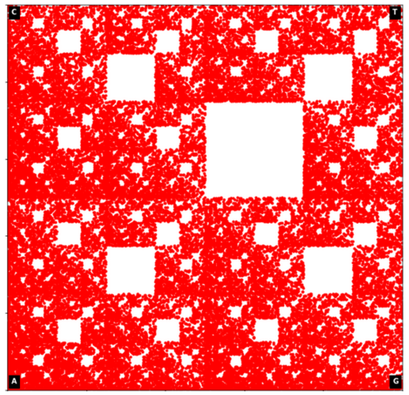
gdzie $N_{i+1} \in \{N_A, N_G, N_C, N_T\}$ jest nukleotydem na $i+1$-tej pozycji..



Rysunek 1:Konstrukcja CGR

Szukanie wzorców





Rysunek 2:Brak podsekwencji GT oraz AT odpowiednio.

Współczynnik Hursta

Dla dowolnego szeregu dyskretnego $\{u_i\}_{i=1...N}$ definiujemy:

częściową średnią:

$$< u >_n = \frac{1}{n} \sum_{i=1}^{n} u_i,$$

a następnie częściową różnicę:

$$u_j(n) = \sum_{i \leqslant j} (u_i - < u >_n),$$

Z kolei:

$$R(n) = \max_j u_j(n) - \min_j u_j(n),$$

oraz odchylenie standardowe:

$$S(n) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (u_i - < u >_n)^2}.$$

Współczynnik Hursta (H) zdefiniowany jest jako:

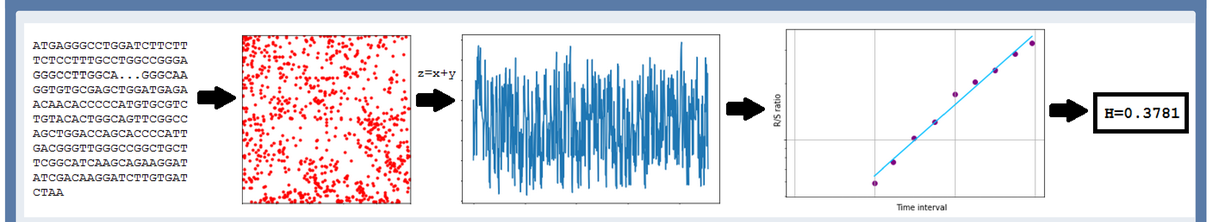
$$E[\frac{R(n)}{S(n)}] = Cn^H$$

gdzie $n \to \infty$, a C jest pewną stałą. Logarytmując obie strony otrzymujemy:

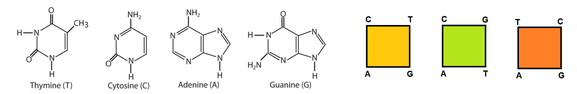
$$\log(\frac{R(n)}{S(n)}) = H \log n + C_1.$$

To oznacza, że H jest współczynnikiem nachylenia wykresu $\log(\frac{R(n)}{S(n)})$ vs $\log n$.

Procedura wyzaczania współczynnika Hursta na podstawie CGR



Gen SPARC -Klasyfikacja



Rysunek 3:Trzy wersje CGR użyte do wyznaczenia współczynnika Hursta.

Taxon	H1	H2	H3
owl	0.31	0.26	0.28
chicken	0.28	0.25	0.27
human	0.39	0.37	0.35
chimpanzee	0.38	0.37	0.34
dog	0.36	0.36	0.34
glassy fish	0.33	0.28	0.31
greater amberjack	0.34	0.29	0.32

Tabela 1:Współczynnik Hursta dla wybranych taksonów.

Literatura

1 S. Vinga, J. Almeida, Alignment-free sequence comparison - a review, Bioinformatics, 19 (2003), pp. 513-523.

2 Y. Changchuan, Encoding DNA sequences by integer chaos game representation, Journal of Computational BiologyVol. 26, No. 2, 2017.

3 L. Yihui, D. Wei, Artificial Intelligence and Data Mining: Algorithms and Applications, Hindawi Publishing Corporation, 2013.

Kontakt

Email: adrian15x.kania@uj.edu.pl