Semester 2024-III Workshop No. 1 — Entrophy and Divide&Conquer

Nahin Jose Peñaranda Mejia 20231020032

Introduction

This analysis focuses on generating and processing DNA sequences of different lengths with different base probabilities (A, C, G, T) to examine sequence patterns and motifs. These sequences range in size from 1000 to 2000000 bases, and size from 5 to 100 nucleotide bases, allowing the study of motif occurrence and its relationship to sequence properties such as base probability distribution, motif size, and Shannon entropy. The system generates sequences based on certain base probabilities given by person and evaluates the presence of specific motifs (Combinations with same occurrence). The number of occurrences of each motif was counted and doing the same with the Shannon entropy (a measure of sequence randomness) was calculated to assess whether the sequence had high or low complexity.

$$H = -\sum p(x)\log p(x)$$

Complexity Analysis

Complexity in this context refers to the variability of the sequences based on the probability distribution of the four bases (A, C, G, T) and the presence of recurrent motifs.

In this sense, sequence variability determined by the probability distribution of the four bases and the existence of repeating motifs is referred to as complexity. Sequences with uniform randomness, like 0.25-0.25-0.25-0.25, have equal odds for each base, but skewed base probabilities provide more predictable, repetitive themes, like "AAAA."

For instance, the motif "AAAA" appeared 2,858 times in one sequence and 2,370 times in another, with sequences of size 10,000 and base probabilities 0.4-0.2-0.2-0.2. Given that there is a greater likelihood of "A" occurring, the motif appears more frequently, which implies a lower level of complexity in the sequences. Lower motif frequencies were seen in sequences with more balanced probability or higher motif size thresholds, which suggested increased complexity and decreased predictability.

Chaos Analysis

Chaos in DNA-like sequences can be quantified using Shannon entropy, which measures the disorder or unpredictability of the system. A sequence with low entropy is more ordered and predictable, while high entropy reflects greater randomness.

In sequences with low base probability diversity (e.g., 0.5-0.2-0.2-0.1), specific motifs like "AAAAAAA" appeared frequently, but with low entropy, indicating a more ordered system. In contrast, sequences with more balanced probabilities showed higher entropy, as reflected by the reduced occurrence of specific motifs.

According to the results, sequences that had been filtered for high entropy and had Shannon entropy enabled often showed fewer motif occurrences. For example, without entropy filtering, the sequence with probabilities of 0.3-0.1-0.3-0.3 and size 10,000 contained 180 motif occurrences, however the identical sequence with entropy filtering revealed only 132 motif occurrences. This shows that the system becomes more chaotic when Shannon entropy is applied, which reduces motif frequency and increases sequence diversity.

Results

Number of sequence	Size of sequence	Probability of bases(A-C-G-T)	Motif	Motif Size	Motif Occurrence	With Shannon Entropy	Time to find motif(ms)
1000	5	0.25-0.25-0.25-0.25	CAAGA	5	5	false	235
1000	5	0.25-0.25-0.25-0.25	ATGAT	5	4	true	147
10000	20	0.4-0.2-0.2-0.2	AAAA	4	2858	false	2466
10000	20	0.4-0.2-0.2-0.2	AAAA	4	2370	true	1721
100000	20	0.2-0.3-0.3-0.2	GCCGGG	6	1201	false	17063

100000	20	0.2-0.3-0.3-0.2	GCCGGG	6	1101	true	19140
10000	10	0.3-0.1-0.3-0.3	AAAGT	5	180	false	2005
10000	10	0.3-0.1-0.3-0.3	ATTGA	5	132	true	1209
10000	60	0.4-0.1-0.1-0.4	ATTT	4	14732	false	2099
10000	60	0.4-0.1-0.1-0.4	ATTT	4	13528	true	1538
1000000	40	0.1-0.2-0.2-0.5	TTTTTTT	7	138056	false	177767
1000000	40	0.1-0.2-0.2-0.5	TTTTTTT	7	98367	true	143264
10000	40	0.5-0.2-0.2-0.1	AAAAAAA	7	1426	false	1946
10000	40	0.5-0.2-0.2-0.1	Null	7	0	true	0
10000	40	0.2-0.3-0.3-0.2	CCGGCCG	7	95	false	2381
10000	40	0.2-0.3-0.3-0.2	Null	7	0	true	0
1000	10	0.2-0.3-0.3-0.2	CGG	3	219	false	261
1000	10	0.2-0.3-0.3-0.2	CGG	3	199	true	229

Discussion result

The results show a clear correlation between base probability, motif size and motif frequency. Sequences with a higher probability of certain bases show a higher occurrence of the motif, especially if the motif consists of repeated characters from the main base (eg "AAAA"). Application of Shannon entropy consistently reduced the occurrence of motifs, showing an increase in sequence diversity and a decrease in system predictability. This is particularly evident in sequences with a base probability of 0.4–0.2–0.2–0.2, where the occurrence of the “AAAA” motif is significantly drop when entropy is used. Computational cost (measured as time to find a motif) increases with sequence size and motif complexity. Larger sequences such as 1,000,000 bases take much longer to process, especially for more frequent motifs such as "TTTTTTTT"(This test was compiled in a PC with i3,1005g1). This shows that both sequence length and motif frequency are important factors affecting computational complexity.

Conclusions

To conclude, our study shows that the complexity and predictability of DNA sequences are mostly determined by motif size and base probability. Sequences with more balanced probabilities or larger motif sizes show more diversity and complexity, whereas sequences with skewed base probabilities tend to yield more frequent, predictable motifs. To decrease motif recurrence and create more chaotic systems, Shannon entropy is a good metric for increasing sequence randomization.