z5242692
Chenqu Zhao
COMP9101 (T2-2020)
Homework 3 - Q1

Traverse the original DNA sequence to count the numbers n_s, n_n, n_a, n_k, n_e of occurrences of each letter S, N, A, K, E. This takes $O(n)$ time.

Let $M = \min\{n_s, n_n, n_a, n_k, n_e\}$. According to the definition of the venom level, the largest possible venom level $x \leq M$. Suppose that this snake's venom level x is M. The expected sequence is $S * xN * xA * xK * xE * x$, of which letters must strictly follow the order and the number of occurrences. For example, K cannot appear on the left of N, and K cannot appear $x + 1$ times between A and E. Apply the greedy method to scan from the left of the original sequence, delete the unrelated letters (not in {S, N, A, K, E}) as well as the letters that violate the order or the number of the occurrences mentioned above. Terminate the scan if less than x appearances available for any letters in {S, N, A, K, E}.

If no termination happens, the scan will reach the rightmost of the sequence, and the remaining sequence is the same as the expected sequence. Thus, the venom level is equal to x. Otherwise, we need to modify the value of x and keep scanning recursively. We can change the value of x as what we do in a binary search. Let $x = \frac{\mathrm{M}}{2}$, and see if it matches the sequence. If so, let $x = \frac{3M}{4}$ and see if matches, otherwise let $x = \frac{5M}{8}$ and see, and so on. The result is the largest possible number of x whose expected sequence could be matched after the scan. If none of the expected sequences of x is matched, the venom level is 0. There are $O(\log n)$ scans and each scan takes $O(n)$.

In total, this algorithm takes $O(n) + O(n \log n) = O(n \log n)$.