cpt_s 350

Homework 1

11641327 Yu-Chieh Wang

2020/1/18

1. Problem is a language.

   (1) Given: a number n and two primes p and q,

   Question: is it the case that $n = p \cdot q$?

   $L : \{\langle n, p, q \rangle : n = p \cdot q\}$.

   (2) Given: a number n,

   Question: is it the case that $n = p \cdot q$ for some primes p and q?

   $L : \{\langle n \rangle : \exists p, q \text{ where } n = p \cdot q\}$.

   (3) Given: an NFA A and a word w,

   Question: Does A accept w?

   $L : \{\langle A, w \rangle : \text{w be accepted by A}\}$.

   (4) Given: an NFA A,

   Question: Is there a word w such that A accepts w?

   $L : \{\langle A \rangle : \exists w \text{ be accepted by A}\}$.

2. Big O

(1) Function $2n^3 - 18n$ is $O(n^3)$ and also it is $O(n^4)$ but it is not $O(n^2 \log n)$.

a.

Proof : $2n^3 - 18n \leq 4n^3$ for almost all n.

b.

Proof: Assume $2n^3 - 18n$ is $O(n^2 \log n)$.

Then, $\exists C > 0, n_0 > 0$ such that $2n^3 - 18n \leq C \cdot n^2 \log n$ for all $n \geq n_0$.

Second, try to solve the equation : $\frac{2n^3}{n^2 \log n} - \frac{18n}{n^2 \log n} \leq C$ for all $n \geq n_0$.

Next, $\frac{2n}{\log n} - \frac{18}{n \log n} \leq C$ for all $n \geq n_0$.

Finally, when $n \to \infty$, the equation becomes $+\infty \leq C$ which is wrong.

(2) Function $3n^2 2^{2n}$ is $2^{O(n)}$.

Proof : $3n^2 2^{2n} \leq 2^{4n}$

Give $\log_2$ to each side, then we can get : $\log_2 3n^2 2^{2n} \leq \log_2 2^{4n}$.

Next, we solve the equation : $\log_2 3n^2 + log_2 2^{2n} \leq \log_2 2^{4n}$,

$\to \log_2 3n^2 + 2n \leq 4n$

$\to \log_2 3n^2 \leq 2n$ for almost all n.

3. 3D structure —Design an Algorithm M to compute a similarity metric between two protein molecules.

   A. Why such an algorithm M could help to develop new medical drugs?

      This algorithm can bring more benefits to scientists than using the microscope and the naked eye to identify. First of all, It can accurately mark the difference between old and new medicines, making it easier for scientists to improve the effectiveness of medicines when developing new medicines. Second, It can compare two different brands of medicines to help scientists find ingredients that really help cure the disease. Next, when two drugs are detected to be very similar, scientists can choose to market drugs with fewer side effects. Finally, scientists can use the results of this algorithm to determine whether their brand drugs are likely to be plagiarized.

   B. How to represent a protein molecule by a data structure?

      For each protein, there is a array to store all its compositions; for example, $C_6H_{12}O_6$, CO_2 or H_2O. In addition, each cell of the array has a link list to store a chemical formula. In this link list, each node has three values, the first one is the a string to store the name of chemical element (C, H, or O), the second one is a number which represents how much of the element is in the chemical formula, and the last value points to the position of the next node.

   C. Two definitions of the similarity metric between two protein molecules

      There are two methods : Full similarity and partial similarity for each chemical formula. First, the full similarity means we only count the same formula between two protein. For example, if we have $C_6H_{12}O_6$ in both protein, it counts. However,

if we have CO_2 in one protein and CO in the other protein, it doesn't count. Then, we count a probability by how many the same formula do they have. The second one is partial similarity which means that we count by how many similar elements. For instance, there are 100% similarity between H_2O and H_2O, but only 80% similarity between CO_2 and CO. The way we count the probability is as follow:

$$\frac{(The\ number\ of\ the\ same\ elements\ that\ two\ protenis\ have)x2}{The\ sum\ of\ all\ elements\ of\ two\ proteins}$$

Therefore, in this case, CO_2 has three elements, and CO has two elements. Then, we can say the sum of all elements of two proteins is 5. In addition, both of them have the same elements CO, so we can say the number of the same elements that two proteins have is 2. As the result, we get the probability : $\frac{2x2}{5} = 80\%$.

Pros and Cons between the two methods

Full similarity method looks more logical because it counts by chemical formulas, but the partial similarity counts by elements which are not complete compositions. However, from another perspective, elements can be overlapped which means that chemical formulas can be disassembled to synthesize other chemical formulas. Therefore, it may be the point that two proteins have the same elements.

Two algorithms M for the two definitions.

Full similarity:

input: two arrays of two proteins

run two array to see how many the same chemical formulas F do they have.

get the number of chemical formulas N that two proteins have.

return (F*2)/N.

Partial similarity:

input: two arrays of two proteins

run one array to get the its all elements and the number of them.

then, run the other array to do the same things.

next, sum the number of all elements of two proteins N.

finally, count how many elements do both proteins have F,

and then return (F*2)/N.