2015-DE-01

Important Parts

Information in the human genome is essentially encoded by the four components Adenin, Cytosin, Guanin, and Thymin. Often, genes are characterized as sequences of characters A, C, G, and T, like CAGGAGGAT.

Such sequences may be very long. Researchers are looking for their important parts, which must occur at least twice in a sequence. The importance of a part is characterized by its *value*, which is computed as follows:
part length + number of its occurrences in the sequence.

Then, the most important part of a sequence is the important part with the highest value. For example: The most important part of the sequence CAGGAGGAT is AGGA. Its value is 6: AGGA is 4 characters long, and it occurs 2 times in the sequence. G is less important, because its value is 5: it occurs 4 times, but is only 1 character long.

What is the most important part of the sequence: CATAGTAGTACA ?
Enter the solution here (as a sequence of capital letters):

The correct answer is:

TAGTA is the correct answer. Its importance value is 7 (length: 5; frequency: 2), and there is no other important part with a value ≥ 7.
There are several other subsequences with importance value 6, though: TAGT and AGTA (length: 4, frequency: 2) as well as A (length: 1, frequency: 5).

The solution can be found by looking at all occurrences of each character, one by one, and see how far you can get finding more than one equal continuations. Both of the two Cs, for instance, can twice be continued to CA, but that's it. The five As can twice be continued to AGTA, the three Ts can twice be continued to TAGTA, and the two Gs can twice be continued to GTA. Because none of these continuations occurs more often than twice, the longest of them (TAGTA) is most important among them. It is also more important than any of the single characters in this example.

It's Informatics!

The discovery of the genetic code has been a breakthrough in biology. Moreover, it lead to a fundamental change of perspective: Life could be considered as being determined by the information that is encoded in the genome sequence. From this perpective, genetics essentially is about encoding and processing information. Starting in the 1990s, researchers from both biology and informatics began to co-operate, and the area of bioinformatics started to grow. Since then, many new ways of representing and processing biological information with computers have been developed. This has had severe impacts on biology, but also on pharmacy and medicine. The analysis of (genetic) sequences is only one aspect of bioinformatics.

Attributes

Category: INF
Age Group / Difficulty: TBD

Interactivity: open text
colour-blind-proof: yes

Authorship

2015-04-14 Proposal: Wolfgang Pohl (DE)

Wording

sequence (of characters)

(most) important part

Comments

2015-04-14 Wolfgang Pohl (DE):
This task has been derived from a problem that was set in the German CS contest "Bundeswettbewerb Informatik" (contest 33, round 2).

2015-05-13 Wolfgang Pohl (DE):
The question string and, accordingly, the explanation section was modified due to review comments. The correctness of the explanation section has been verified by a program.

Files

2015-DE-01-eng.html (this file)
2015-DE-01-eng.pdf (pretty print)

License
