

CSE182 lecture 2 questions: sequence alignment

Vineet Bafna

October 4, 2007

The questions are open ended, but should help you understand lectures better. Do these questions make sense? Are they helpful in following the lecture? Constructive feedback is appreciated.

1 Introduction:

As many biological molecules (DNA/protein/RNA) can be expressed as strings, sequence databases are the most common form of biological databases, and so the most common query you could ask is:

Problem 1: *Given a database of sequences, and a query sequence, find all sequences in the database that are similar to the query.*

Blast is the prototypical tool for this query, and we will spend a few lectures discussing the internals of the tool, which combines ideas from molecular biology, algorithms, and statistics.

2 Questions:

1. Locate NCBI on the web and run Blast with a query. Can you make sense of the output? What are the different variants of Blast?
2. The Blast page has a number of search parameters. These include choice of a database, indel penalties, scoring matrix, E-value, word-match, and so on. What do they mean? We will return to this in subsequent lectures.
3. What is an alignment?
4. What does percent identity, and percent similarity mean in the Blast output? What are the + signs in the alignment?
5. Consider the DNA strings AGATGGCCCCATCG and CGGTCCCCCGATGG. Compute the best global alignment of the two strings assuming match score of 1, and mismatch/indel penalty of -1 . Compute the best local alignment, and the best global alignment. Are the two different?
6. The slide 20 (Base case) talks about initializing the S matrix for global alignment. How would you do this for Local alignment?
7. Compute the best global alignment of ACACAACGG and AAAACG using affine gap costs. Use $+1$ for match, -1 for mismatch, and gap extension, -3 for gap-open. Compare your results for the case when gap open penalty is 0.
8. **Repeat Detection:** When the human genome was sequenced, about 40% of it was found to be repetitive. In other words, there were strings that occurred multiple times (with minor mutations/indels) in the genome. Suppose you want to discover all repetitive regions on the human genome. Describe an approach to computing this. Present a rough calculation on how much memory your computer needs to have to allow this computation.
9. When you are computing sequence alignments for assembling DNA, local and global alignments are not quite what you need. Instead you need to align a *prefix* of a string to a *suffix* of another string. Design an algorithm to compute optimal prefix suffix alignments between two strings.
10. What is the genetic code? How does Blast align a DNA string to a protein string (we will do this in subsequent lectures)?