

Relationships

One important concept in bioinformatics, and in science in general, is the idea of a relationship. That is, a rule \mathbb{R} that takes two objects, x and y , and gives back a *True* or *False* answer, represented by $x\mathbb{R}y$. For example, some relationships are:

- x is greater than y
- x is the double of y
- x is the father of y
- x is classmate of y
- x is on the same taxonomic genus as y

If a relationship \mathbb{R} is *reflexive*, *symmetrical* and *transitive*, then we say that \mathbb{R} is an *equivalence relationship*. That is

- **Reflexivity:** For all x , $x\mathbb{R}x$ is True.
- **Symmetry:** For all x and all y , $x\mathbb{R}y$ is the same as $y\mathbb{R}x$
- **Transitivity:** For all x , y and z , if $x\mathbb{R}y$ and $y\mathbb{R}z$, then $x\mathbb{R}z$

On the other hand, if the relationship is not symmetrical, and instead it is *anti-symmetrical*, then we say that \mathbb{R} is an *order relationship*. That is

- **Symmetry:** For all x and all y , if $x\mathbb{R}y$ is true, and $y\mathbb{R}x$ is true, then $x=y$, that is x and y are the same.

Please give a list of relationships, indicating if they are *equivalence*, *order* or not. The goal of this question is to enhance your observation skills, so we aim for *quantity* and *originality*. The score of the answer will be the number of relationships that are reported by only one person.

Distances

1. What is the difference between the *Hamming* distance and *Edit* distance?
-

The Hamming distance is the number of differences of the same length sequences but the Edit distance can be applied to the different length sequences to obtain distance scores according to insertions, deletions and substitutions.

2. Show that the relationship “distance between x and y is zero” is an equivalence relationship
-

We can show the distance between x and y is zero by the Hamming distance calculation is zero. Because in order to get a zero score in Hamming, they should be same length of sequences.

3. Show that the relationship “distance between x and y is small” is *not* an equivalence relationship
-

We can not show the distance between x and y is small by the Hamming distance. Because the Hamming distance can not have minus values. We need to use the Edit distance to show that distance between x and y is small by scoring gaps and mismatches

Alignment

1. What is the difference between *global* and *local* alignment?
-

Global alignment is an attempt to align end to end whole sequences but local alignment is finding local regions with the highest level of similarity between two sequences.

2. What is the difference between *semi-global* and *local* alignment?
-

Semi global alignment is a modification of global alignment that allows to make the gaps penalty-free on end or/and beginning of the sequences but the local alignment is finding local regions with the highest level of similarity between two sequences.

3. What is the ideal use case for *global alignment*?
-

The ideal use of case is to find all possible alignments having the highest score in end to end alignments.

4. What is the ideal use case for *semi-global alignment*?
-

The ideal use of case is to find all possible alignments having the highest score in end to end alignments including making the gaps penalty-free.

5. What is the ideal use case for *local alignment*?

The ideal use of case is to find all possible alignments having the highest score in the local regions that have highest level of similarity between two sequences.

6. Two bacterial strains of the same species have the same genes but they may be in different order. How would you test this hypothesis?

In order to find the conserved genes in different strains, we prefer to use local alignment because the genome structure and size of strains may be different than each other and we need to find highest level of similarities in local regions.

7. How can you calculate the percentage of nucleotides conserved between the same two bacterial strains of the previous question?

According to the aligned regions' scores, we can determine the percentage of conserved nucleotides between two strains.

Scoring

1. What is the role of λ in the scoring of alignments?

The role of lambda is normalizing the scoring of alignments, it provides talking in same language.

2. In DNA, why the *match* score is positive, and *mismatch* score is negative?

Matching is what we want and positive effect for our scores but mismatch is expensive and negative effect for our scores because matching presents similarity, mismatch presents unsimilarity.

3. In proteins, why some substitution scores are positive and others are negative? What is the biological interpretation of positive scores?

Positive scores indicate that the given amino acid substitution occurs more frequently in the alignment than expected by chance, while negative scores indicate that the substitution occurs less frequently than expected.

4. Why gap score have two parts: existence and extension?

According to evolutionary meanings genomic molecules are transporting as in parts, in order to determine scores according to this condition, they have to cost different than each other,so existence gaps are more expensive than extensive gaps for scoring.

5. Why we need that the gap score be lower than the substitution score?

According to biological causes, gap is more abundant in evolutionary progress, mutations on nucleotides(substitutions) are more expensive than gaps for use to score.

Computational cost

1. What is *computational cost*?

Computational cost is a concept that focuses on the amount of computing resources needed for particular kinds of tasks. The resource usage that depends on the input size.

2. What is the computational cost of the Smith-Waterman algorithm?

The algorithm compares two sequences by computing the similarity score by means of dynamic programming.

3. What is an *heuristic*?

It is a technique designed for solving a problem more quickly when classic methods are too slow, or for finding an approximate solution when classic methods fail to find any exact solution.

4. What is the strategy that BLAST uses to speed up the local alignment?

BLAST uses heuristics to produce results quickly.

5. What is the trade-off of the *word size* parameter in BLAST?

This search strategy offers a tradeoff between speed and sensitivity; smaller word-sizes result in greater sensitivity at the expense of speed while larger word-sizes optimize BLAST for speed.
