ECS129

Assignment: Option 4 (Programming)

A Metric for Protein Sequences

Due: Friday, March 13th, 2020

1 A new method for computing the similarity between two protein sequences

The concept is distance is sequence space is not clearly established. Protein sequence alignment programs provide multiple scores, from a raw score obtained with the dynamic programming method used to align the two string of letters representing the sequences based on a given substitution matrix, to an E-value to measure the statistical significance of the alignment. None of those scores however are actual metric (i.e. distance measures): it was never established that they satisfy the triangular inequality. In addition, they all depend on some parameters that can be seen as arbitrary. The main problem is often related to the possible presence of gaps in the alignment.

Recently, Smale and colleagues introduced a new method for comparing protein sequence [1] that alleviate many of the problems mentioned above. This method does not generate an alignment between two sequences: as such, it does not need to consider gaps. This method is based on the concept of kernels, and it is proven to define an actual metric on the space of protein sequence.

The method works as follows [1]. Let $\mathcal{A}$ be the set of amino acids ($|\mathcal{A}| = 20$.

Step 1. Definition of a kernel $K^1 : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$.

Using the formulation of BLOSUM62, we can define a kernel $Q : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$. Q can be seen as the raw data of BLOSUM62. Let p be the marginal probability defined on $\mathcal{A}$ by Q. That is,

$$p(x) = \sum_{y \in \mathcal{A}} Q(x, y) \tag{1}$$

We define the BL62 matrix as:

$$BL62(x, y) = \frac{Q(x, y)}{p(x)p(y)} \tag{2}$$

Then a kernel $K^1 : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ is given by:

$$K^1(x, y) = BL62(x, y)^{\beta} \tag{3}$$

where β is a parameter.

Step 2. Let $\mathcal{A}^1 = \mathcal{A}$ and define $\mathcal{A}^{k+1} = \mathcal{A}^k \times \mathcal{A}$ for any $k \in \mathbb{N}$. We say $s = (s_1, \ldots, s_k)$ is a k-mer if $s \in \mathcal{A}^k$ for some $k \in \mathbb{N}$ with $s_i \in \mathcal{A}$. We define:

$$K_k^2(u, v) = \prod_{i=1}^{k} K^1(u_i, v_i) \tag{4}$$

where u, v are two k-mers, $u = (u_1, \ldots, u_k)$ and $v = (v_1, \ldots, v_k)$. K_k^2 is a kernel on the set of all k-mers.

In simpler words, given two subsequences u and v of length k, $K_k^2(u, v)$ is the dot product between those two subsequences.

Step 3. Let $f = (f_1, \ldots, f_m)$ be an amino acid sequence, and let $|f|$ be its length. We write $u \subset f$ whenever u is of the form $u = (f_{i+1}, \ldots, f_{i+k})$ for some $1 \leq i+1 \leq i+k \leq m$. Let g be another amino acid chain. We define:

$$K^3(f, g) = \sum_{\substack{u \subset f, u \subset g \\ |u| = |v| = k \\ \text{all} \quad k = 1, 2, \ldots}} K_k^2(u, v). \tag{5}$$

Step 4. We define the correlation kernel $\hat{K}^3$ normalized from the kernel K^3 as:

$$\hat{K}^3(f, g) = \frac{K^3(f, g)}{\sqrt{K^3(f, f) K^3(g, g)}} \tag{6}$$

Step 5. The correlation kernel $\hat{K}^3$ defines a dot product on the space of sequences. The corresponding distance is defined by:

$$d_K(f, g) = \sqrt{2(1 - \hat{K}^3(f, g))} \tag{7}$$

We note that the distance d_K does not rely on gap penalties, nor does it generate an alignment between the two sequences that are compared.

2 Problem

For this option, I am asking you to:

- Implement this method. Your program should read in:
  - The matrix BL62

  - The parameter β
  - Two sequences S_1 and S_2

  The program will then output the distance between the two sequences, as defined by equation (7).

- Test your program on the the three sequences provided (seq1.fa, seq2.fa, seq3.fa). The expected values with $\beta = 0.01$ are: d(Seq1,Seq2) = 0.278, d(Seq1,Seq3) = 0.0245, and d(Seq2,Seq3)=0.2799.

- Write a summary report. There is no need to send a lengthy write-up, but it should definitely include an introduction, results and analysis, a conclusion, and references to published work, if needed.

References

[1] W.-J. Shen, H.-S. Wong, Q.-W. Xiao, X. Guo, and S. Smale. Towards a mathematical foundation of immunology and amino acid chains. 2012.