Inferring a nucleotide substitution model by simulating mutation with AlphaFold2

July 10, 2023

Abstract

Sequence Alignment algorithms rely on some form of substitution matrix, which can heavily influence their results. For DNA sequences, BLAST, the most widely used sequence alignment tool, uses a matrix which rewards matches with a +2 and penalizes mismatches with a -3. In this thesis, we try to explain why such a simple substitution matrix was chosen, illuminate the problem in breadth and derive and propose a new substitution matrix for DNA sequences.

Contents

1 Sequence Alignment & Substitution Matrices

This section will introduce the context for the thesis at large: sequence alignment and substitution matrices. The reader is supposed to be versed in biology, however, the thesis aims to reintroduce the them to the topic again.

The chapter is supposed to be sufficiently in-depth to illustrate the significance of the contributions without overburdening the reader.

1.1 What is Sequence Alignment?

(2 pages) Very shallow introduction

1.2 Why is Sequence Alignment useful?

(4 pages) This subsequence is supposed to give the reader an idea of the significance of sequence alignment in general, and also show (and/or cite) a few specific examples where sequence alignment has had a positive effect on the world.

1.3 How does Sequence Alignment work from a birdseye view?

(4 pages) A description of the low-level workings of Sequence Alignment algorithms, specifically how scoring works, the dynamic programming algorithms, and BLAST. MSA and other algorithms will be mentioned, but not fully explained.

1.4 What is a Substitution Matrix?

Segway into substitution matrices, their Markovian nature, and discuss why nucleotide subsitution matrices, albeit smaller, pose a greater challenge than amino acid substituion matrices.

1.5 What are the most important ways to derive substitution matrices for nucleotides?

(3 pages) This subsection gives insight into established ways of deriving substitution matrices.

2 A new nucleotide subsitution matrix

The goal of this section is to derive a new way of deriving a substitution matrix for nucleotides, while reiterating why this might be a good, or why it might be a bad idea.

2.1 Context for the new method

This is pretty difficult to plan, as I am unsure what the new method will be.

2.2 How do we know a substitution matrix is good? (Deriving some form of evaluation)

To claim that the new method is good, there needs to be some form of evaluation preferrably there are some established metrics for sequence alignment, or some

benchmark datasets, however, if there are no such things, it is up to us to invent some form of metric.

2.3 In-depth explanation of new method

Giving a formal explanation of the new method.

3 Experimentation and results

Doing some experiments and explaining them. Preferrably, also some visualizations.