Inferring a nucleotide substitution model by simulating mutation with AlphaFold2

July 25, 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, potentially tailored to specific protein families.

Contents

1 Notes & preparation

1.1 Previous work & literature review

There has been previous work aiming to solve the same problem. States, Gish, and Altschul [7] first a comparison of the default BLAST parameters and a specific model with transversion specific parameters – however, their paper is not publicly available. Hamada et al. [3], whilst focussing on different sequencing technologies, give a method for generating a substitution matrix for specific species and sequencing technologies using expected counts and a Markov model approach. Jain et al. [4] use another EM approach to specifically find better alignment parameters for MinION sequencing technology. The above use simulated data for training and evaluation.

Generally, the most notable thing is: for nucleotides, there seem to be only simulated test data (via synthetic evolution or synthetic mutation) – there are no reliable non-synthetic, manually curated datasets. Simulated evolution, at least when done very naivley, can be perfectly reconstructed, however, it can not truly reflect evolution. A good overview of the current simulated datasets can also be found in section 5 of [2].

I also did a lot of research on AlphaFold. Most notably, it turns out that there are studies showing that AlphaFold2 can fold mutated protein sequences [5] as well as studies claiming it cannot [6, 1] - the former using a different metric of protein divergence.

2 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.

2.1 What is Sequence Alignment? (2 pages)

(2 pages) Very shallow introduction

2.2 Why is Sequence Alignment useful? (4 pages)

(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.

2.3 How does Sequence Alignment work from a birdseye view? (4 pages)

(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.

2.4 What is a Substitution Matrix? (2 pages)

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

2.5 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.

3 A new nucleotide subsitution matrix (around 8 pages)

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.

3.1 Context for the new method

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

3.2 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.

3.3 In-depth explanation of new method

Giving a formal explanation of the new method.

4 Experimentation and results (5 pages)

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

References

[1] Gwen R. Buel and Kylie J. Walters. "Can AlphaFold2 Predict the Impact of Missense Mutations on Structure?" In: *Nature Structural & Molecular Biology* 29.1 (Jan. 2022), pp. 1–2. ISSN: 1545-9985. DOI: 10.1038/s41594-021-00714-2. (Visited on 07/25/2023).

[2] Jiannan Chao, Furong Tang, and Lei Xu. "Developments in Algorithms for Sequence Alignment: A Review". In: *Biomolecules* 12.4 (Apr. 2022), p. 546. ISSN: 2218-273X. DOI: 10.3390/biom12040546. (Visited on 07/25/2023).

[3] Michiaki Hamada et al. "Training Alignment Parameters for Arbitrary Sequencers with LAST-TRAIN". In: *Bioinformatics* 33.6 (Mar. 2017), pp. 926–928. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw742. (Visited on 07/14/2023).

[4] Miten Jain et al. "Improved Data Analysis for the MinION Nanopore Sequencer". In: *Nature methods* 12.4 (Apr. 2015), pp. 351–356. ISSN: 1548-7091. DOI: 10.1038/nmeth.3290. (Visited on 07/24/2023).

[5] John M. McBride et al. *AlphaFold2 Can Predict Single-Mutation Effects on Structure and Phenotype*. June 2023. DOI: 10.1101/2022.04.14.488301. (Visited on 07/20/2023).

[6] Marina A. Pak et al. *Using AlphaFold to Predict the Impact of Single Mutations on Protein Stability and Function*. Sept. 2021. DOI: 10.1101/2021.09.19.460937. (Visited on 07/25/2023).

[7] David J. States, Warren Gish, and Stephen F. Altschul. "Improved Sensitivity of Nucleic Acid Database Searches Using Application-Specific Scoring Matrices". In: *Methods* 3.1 (Aug. 1991), pp. 66–70. ISSN: 1046-2023. DOI: 10.1016/S1046-2023(05)80165-3. (Visited on 07/18/2023).