Albert Einstein

Eine neue Bestimmung der Moleküldimensionen

Diss. ETH No. ?

PESHO IVANOV

OPTIMAL SEQUENCE ALIGNMENT USING A*

DISS. ETH NO. ?

OPTIMAL SEQUENCE ALIGNMENT USING A*

A dissertation submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

PESHO IVANOV
Dipl., Eidgenössisches Polytechnikum

born on 29 May 1989
citizen of Bulgaria

accepted on the recommendation of

Prof. Dr. Martin Vechev, examiner
Prof. Dr. Gunnar Rätsch, co-examiner
Prof. Dr. Veli Mäkinen, co-examiner
Prof. Dr. Paul Medvedev, co-examiner

2022

DOI: 10.3929/ethz-a-

With gratitude to my high-school physics teacher Svilen Rusev. Thank you for opening the world of science to me in an intuitive and visual way.

ABSTRACT

Sequence alignment is the process of detecting similarities between biological sequences, such as DNA, RNA and proteins. For the last half a century, sequence alignment has been of central importance for molecular biology. Applications include evolutionary biology, genome assembly, read mapping, variation detection and computational methods are crucial for the correctness of the analyses of the vast amounts of biological data. Can we use the the A* shortest path algorithm to find optimal alignments fast?

This thesis explores two variations of the alignment problem: *reads mapping* and *pairwise alignment*, also known as semi-global and global alignment. Biological sequences do not generally align perfectly due to biological differences and technical errors. Given two sequences, the desired alignment is a position-to-position correspondence between two sequences which minimizes the edit costs (substitutions, insertions or deletions). This task is closely related to calculating *edit distance*. Practical alignment algorithms are desired to ① find accurate alignments, ② apply to a wide range of data, and ③ use little time and memory.

Existing aligning algorithms are either optimal but quadratic or fast but appximate. Even though the resulting alignment is linear. This gap has motivated the development of various faster but approximate algorithms. Moreover, theoretical results suggest that it is that in general, alignment is not solvable in strongly subquadratic time.

We consider a principled alignment formulation based on shortest paths and demonstrate that the A* shortest path algorithm can be used to outperform current methods. Unlike existing methods, A* enables an *informed search* based on information from the unaligned sequence suffixes, thus radically improving the empyrical runtime scaling (up to linear) in the average case while providing optimality guarantees. On real data, this approach reaches orders of magnitude of speedup compared to existing approaches.

An optimal alignment can naturally be represented as a shortest path in an alignment graph (equivalent to the DP table). In order to find such a shortest path with minimal exploration, we instantiate the A* algorithm with a novel problem-specific heuristic function based on the unaligned parts of the sequences. This additional information is a problem-specific heuristic function and it heavily determines the efficiency of the search. For any explored state by A*, this heuristic function should compute a lower

bound on the remaining path length, or more specifically, the minimal cost of edit operations needed to align the remaining sequences.

Seed heuristic. In practice, while achieving polynomial speed ups on real data. It ① provides optimality guarantees according to edit distance, ② scales to long and noisy sequences, and ③ scales subquadratically with sequence length. To scale to large reference sequences, we extend the graph with a trie index. To scale to long queries, we introduce design an admissible *seed heuristic*, which is provably-optimal also efficient to compute. To scale to high error rates, we design

Many tools do not have a well-stated problem they optimize.

Probabilistic approach. Focus on the metric. Extend to MSA, local, affine. Prorotype implementations No asymptotics.

ZUSAMMENFASSUNG

Deutsche Zusammenfassung hier.

ACKNOWLEDGEMENTS

I would like to thank …

CONTENTS

NOTATION

FREQUENTLY USED SYMBOLS

E	energy
m	rest mass
p	impulse

PHYSICAL CONSTANTS

c	speed of light in vacuum, $c = 299\,792\,458\,\mathrm{m\,s^{-1}}$

(CODATA 2014 [**codata**])

1 INTRODUCTION

It is better to be wrong than to be vague.
— Freeman Dyson

The number of possible alignments grow exponentially with length. The usual underlying question to finding "correct" alignments. Regarding the precision of alignment, one is usually interested in base-to-base (aka letter-to-letter) correspondence between the sequences, even though for some applications a less detailed solution is sufficient: only the similarity between sequences or the location where a read maps to a reference. Exact alignment is only useful for very short sequences (often kmers), and for all other cases the optimized metric may be hamming distance, edit distance (unit costs), Levenshtein distance, affine costs, convex and concave costs, general costs and others.

Depending on the the number of aligned sequences, there is pairwise alignment and multiple sequence alignment (MSA). Depending on the parts of the sequences that are aligned to each other, we differentiate global, local and various semi-global alignemnts. There are generalizations to sequence-to-sequence alignment, including aligning to nonlinear structures, such as directed acyclic graphs, DAGs, general graphs and others. These structures are nowadays becoming more common as a compressed form of representing a set of references to which a sequence can be aligned. Often, one best alignment is sufficient but finding several best (top-K) alignments. In the context of read mapping, a set of reads is aligned to the same reference sequence so an indexing procedure is often useful for the performance.

We specifically consider the mapping of a set of reads to a general graph, and the global pairwise alignment.

Existing optimal algorithms are based on dynamic programming (DP) and run in quadratic time (assuming that the number of errors is proportional to the length)

we employ the A* algorithm which is an *informed search* algorithm. TODO: a case for the informed algorithms

2

SAMPLE CHAPTER

The true logic of this world is in the calculus of probabilities.

— James C. Maxwell

Maxwell1865 derived some very useful equations for electromagnetic fields:

$$\nabla \cdot \vec{D} = \rho \tag{2.1}$$

$$\nabla \cdot \vec{B} = 0 \tag{2.2}$$

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{2.3}$$

$$\nabla \times \vec{H} = \vec{j} + \frac{\partial \vec{D}}{\partial t} \tag{2.4}$$

The energy–momentum relation, **??**, is one of *my* important results:

$$E^2 = m^2c^4 + (pc)^2 \tag{2.5}$$

Write units like this: 5 µm.

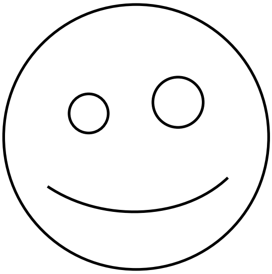

Figure 2.1: A lovely face.

3

SUMMARY

I dream my painting and I paint my dream.
— Vincent van Gogh

Summary here.

A

APPENDIX

Here be dragons.

CURRICULUM VITAE

PERSONAL DATA

Name	Albert Einstein
Date of Birth	March 14, 1879
Place of Birth	Ulm, Germany
Citizen of	Switzerland

EDUCATION

1896 – 1900	Eidgenössisches Polytechnikum, Zürich, Switzerland *Final degree:* Diploma
1895 – 1896	Aargauische Kantonsschule (grammar school) Aarau, Switzerland *Final degree:* Matura (university entrance diploma)
– July 1894	Luitpold-Gymnasium (grammar school) Munich, Germany

EMPLOYMENT

June 1902 –	Technical Expert, III Class *Federal Office for Intellectual Property*, Bern, Switzerland

PUBLICATIONS

Articles in peer-reviewed journals:
Conference contributions: