ABSTRACT

Sequence alignment is the task of finding similarities between sequences. It has been a central building block in molecular biology since DNA, RNA and protein sequences were first obtained half a century ago. Sequence alignment is applied to research and medicine, with applications in evolutionary biology, genome assembly, oncology and many others. The analyses of the growing amounts of genomic data require algorithms with high accuracy and speed. Moreover, the ongoing transition from single genome references to pangenomes (representative for a whole population) motivates novel alignment algorithms.

We consider the two problems: *semi-global* alignment of a set of DNA reads to a pangenome graph reference (possibly containing cycles), and *global* (end-to-end) alignment of two sequences. Recent theoretical results conclude that strongly subquadratic algorithms are unlikely to exist for the worst case. Moreover, the runtime and memory of existing optimal algorithms scale quadratically even for similar sequences. An open problem is to develop an algorithm with linear-like empirical scaling on inputs where the errors are linear in n. In the present thesis we introduce an approach that aims to solve this problem in order to develop practical optimal alignment algorithms.

Modern optimal aligners perform uninformed search – they neglect information from the not-yet-aligned parts of the sequences. We exploit this information in a principled framework where an alignment with minimal edit distance is equivalent to a shortest path in an alignment graph, and the remaining information about the sequences is captured in a heuristic function that estimates the length of a remaining shortest path. The classic shortest path algorithm A* uses such a heuristic to direct the search, and yet it finds provably shortest paths given that the heuristic is *admissible*, i.e. it always return a lower bound on the edit distance of the remaining suffixes. Curiously, previous attempts to apply A* to multiple sequence alignment (MSA), did not result in practical algorithms even for pairwise alignment. In the present thesis we investigate how to do that. In the shortest path formulation, we (i) suggest a novel highly-informed admissible heuristic, (ii) design efficient algorithms and data structures for computing the heuristic, (iii) prove their optimality, (iv) implement the presented algorithms, and (v) compare their performance and scaling to other optimal algorithms. Our approach is provably optimal according to edit distance, its runtime empirically scales subquadratically (and sometimes even near-linearly) with the output size, which translates to orders of magnitude of speedup compared to state-of-the-art optimal algorithms.

Throughout this thesis, we demonstrate how to encompass various dimensions of the input complexity while preserving the speed and optimality. First, we demonstrate how to apply a trie index to scale the alignment runtime sublinearly with the reference size. Then, we introduce a novel seed heuristic for $A^\star$ which enables aligning of long sequences (up to 100 Mbp) near-linearly with their length. Last, we extend the seed heuristic to a general chaining seed heuristic that encompases inexact matching, match chaining, and gap costs to raise the tolerated error rate (to 30% for synthetic data and 10% for real data). Interestingly, our seed heuristic challenges the omnipresent seed-(chain)-extend paradigm to sequence alignment which aims to connect long and high-quality seed matches. We, instead, use the information of the lack of matches to dismiss suboptimal alignments, which leads to optimizing seeds for being short and not having many matches.

These first steps to scalable optimal alignment using $A^\star$ give rise to a number of research directions: approaching other alignment types, more general optimization metrics, relaxed optimality guarantees, performance analyses, improved heuristics, applications outside of biology, and more performant algorithms and implementations.