ABSTRACT

Sequence alignment is the process of detecting similarities between sequences. Since genetic sequences were first sequenced half a century ago, sequence alignment is a basic task in molecular biolog, with applications in evolutionary biology, genome assembly, variation detection, and others. An ongoing transition from using genomes to using pangenomes motivates a rethinking of the classic alignment algorithms. The variety of applications combined with the growing amount of genetic data motivate the development of fast and accurate alignment algorithms.

Existing alignment algorithms are either optimal but quadratic or fast but approximate. This thesis proses an elegent approach to alignment based on the A* algorithm, which is both heuristically fast and provably optimal. It has been shown that alignment is likely not solvable in strongly subquadratic time in the general case. The goal we pursue throughout this thesis is to apply the A* approach to as many types of data as possible, while remaining fast and optimal.

We consider two types of alignment: *semi-global*, for mapping a set of DNA sequences to a pangenome reference; and *global*, for calculating the edit distance distance between two sequences. In order to handle various data dimensions, we propose several techniques and empyrically study their runtime scaling: a trie index enables sublinear scaling with the reference size, *seed heuristic* enables near-linear scaling with sequence length, and inexact seed matching and match chaining enable scaling to high error rates. Owing to the superior scaling of the A* approach, our prototypical implementations run orders of magnitude faster than existing optimal approaches even on long erroneous sequences.

We foresee a multitude of future directions for advancing A* for sequence alignment, including other types of alignment, generalizing the edit distance metric, relaxing the optimality guarantee, theoretical analyses of the performance, and more efficient implementations for production use.