MAGUS: Multiple sequence Alignment using Graph clustring

Abstract:

Estimating large multiple sequence alignments (MSAs) is a fundamental bioinformatics problem. Divide-and-conquer is an effective strategy that has been shown to increase the scalability and accuracy of MSA estimation in proven methods such as SATé and PASTA. A sequence dataset is split into disjoint subsets in these divide-and-conquer techniques. A sequence dataset is split into disjoint subsets, alignments are calculated on the subsets using base MSA methods (e.g. MAFFT), and then the alignments are combined into an alignment on the whole dataset.

MAGUS, or Multiple Sequence Alignment using GraphclUStering, is a modern method for calculating large-scale alignments. MAGUS is similar to PASTA in that it follows almost the same steps (starting tree, similar decomposition technique, and MAFFT to compute subset alignments), but then merges the subset alignments using the Graph Clustering Merger. In this paper, we introduce a new approach for integrating disjoint alignments. On a diverse variety of biological and virtual datasets, our research reveals that MAGUS outperforms PASTA in terms of precision and speed on big datasets, while matching it on smaller datasets.

Introduction:

Multiple sequence alignment (MSA) is a fundamental step in many bioinformatic pipelines, but estimating MSAs on large sequence datasets is difficult, especially for datasets with low sequence identification and high rates of insertions and deletions (indels).Methods for dealing with massive datasets have been created.

divide-and-conquer: they decompose the input sequence dataset into subsets, coordinate each subset using a base alignment approach of choice, and combine the alignments together (treating them as constraints) into an alignment on the entire dataset. Divide-and-conquer has many advantages, including scalability of massive datasets, but the greatest gain is increased precision. Most alignment methods, including top-performing methods like MAFFT, either degrade in precision with high rates of evolution and sequence number, or actually cannot operate on large datasets.

Divide-and-conquer prevents this by using these techniques instead on smaller, more closely connected subsets of sequences.Since the alignments on such subsets are usually very precise, they are treated as constraints. As a result, the opportunity to combine constraint alignments has a huge impact on the precision of the final alignment. The alignment will then be used to approximate a tree. PASTA improved on SATé-II by changing the merger stage, while SATé-II improved on PASTA by changing the decomposition approach. PASTA is the state-of-the-art for this family of divide-and-conquer MSA pipelines as a result of these improvements, which increased precision and scalability. In addition, UPP (Nguyen et al., 2015) which is intended to achieve good alignments in the face of sequence length heterogeneity, uses PASTA to compute a ‘backbone' alignment on sequences considered ‘full length,' represents the backbone alignment using an ensemble of profile Hidden Markov Models, and then aligns the remaining sequences to the ensemble. As a result, though PASTA (like most other methods) is

appropriate for datasets with little to no sequence length heterogeneity; it is often used in methods for aligning datasets with varying sequence lengths. MAGUS, or Multiple Sequence Alignment using Graph clUStering, is a new approach for computing large-scale MSAs. MAGUS begins with the same steps as PASTA's first version, and has the following structure. PASTA uses a fast technique (similar to UPP, except using a random set of sequences for the backbone without regard for sequence length) and FastTree2 to compute an alignment. divides the sequences into disjoint subsets by removing edges until all subsets are less than the limit permitted (default 200), computes alignments on the subsets (default MAFFT), and then merges the alignments [merging pairs of restriction alignments and then completing the merger using transitivity]. PASTA then computes a tree on the merged alignment using FastTree, and can continue by iterating; by default, PASTA iterates three times. The two main ways that MAGUS differs from PASTA is that it combines the disjoint alignments using the Graph Clustering Merger using Opal]. We investigate MAGUS on a diverse variety of sequence databases, including both simulated and biological datasets. On almost all datasets, we found that MAGUS provides more reliable alignments with a single iteration than default PASTA (which requires three iterations); additionally, MAGUS is quicker because it only requires one iteration.

Related works:

Methods for dealing with massive datasets have been created.: Garriga E. et al. (2019) Large multiple sequence alignments with a root-to-leaf regressive method. Nat. Biotechnol., 37, 1466–1470,

Lassmann T. (2019) Kalign 3: multiple sequence alignment of large datasets. Bioinformatics, 36, 1928–1929.

PASTA is designed to produce good alignments in the presence of sequence length heterparity:

Nguyen N-p. et al. (2015) Ultra-large alignments using phylogeny-aware profiles. Genome Biol., 16, 124. MagUS is a new method to compute large-scale MSAs. It uses a fast technique similar to UPP.(Price M.N. et al. (2010) FastTree 2—approximately maximum-likelihood trees for large alignments.)

To compute a tree on the alignment, divides the sequences into disjoint subsets by removing edges until all subsets are less than:

Wheeler T.J., Kececioglu J.D. (2007) Multiple alignment by aligning alignments. Bioinformatics, Edgar R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity