Adaptive Local Realignment via Parameter Advising

Dan DeBlasio and John Kececioglu
Department of Computer Science, The University of Arizona

Overview

Mutation rates can vary across the residues of a protein, but when multiple sequence alignments are computed for protein sequences, the same choice of values for the substitution score and gap penalty parameters is often used across their entire length. We provide for the first time a new method called adaptive local realignment that automatically uses diverse alignment parameter settings in different regions of the input sequences when computing protein multiple sequence alignments. This allows parameter settings to locally adapt across a protein to more closely match varying mutation rates.

We build on our prior work on global alignment parameter advising, which recommends an appropriate aligner parameter setting by ranking alternate alignments using the Facet accuracy estimator. Our new method takes a computed global alignment, and in each local region that has low estimated accuracy, generates collection of candidate realignments using a precomputed set of alternate parameter choices. If one of these alternate realignments has higher estimated accuracy than the original subalignment, the region is replaced with the realignment, and the concatenation of these realigned regions forms the new output alignment.

Adaptive local realignment significantly improves the quality of alignments over using the best default parameter choice. In particular, this new method of local advising, when combined with prior methods for global advising, boosts alignment accuracy by almost 23% over the best default parameter setting on the hardest-to-align benchmarks (and almost 5.9% over using global advising alone).

Adaptive Local Realignment

The input to adaptive local realignment [1] is an initial alignment and a parameter advising set. The input alignment could be obtained using global parameter advising on this set, or even from a default parameter setting.

Window scores are generated using our accuracy estimator. A sliding alignment window induces a subalignment, which we score using the Facet alignment accuracy estimator. Each column of the alignment participates in several sliding windows, but is the center of a unique window.

Column scores are computed as a weighted sum of the scores of windows in which a column participates. The weighting is a geometric distribution around the center of a unique window.

Alignment regions are determined by identifying columns of high score, which we call barriers, and columns of low score, which we call seeds. Columns which are not seeds or barriers may be realigned.

We define a realignment region by starting at a seed and expanding, left and right until we reach a barrier. Barriers and columns without a seed remain unchanged.

Local advising uses the parameter advising process described below (see "Parameter Advising") to find the aligner parameter choice that yields the realignment of highest estimated accuracy.

The output alignment is constructed by replacing the realignment regions with the new alignment found by local parameter advising, if this has higher estimated accuracy.

Parameter Advising

Parameter advising is the task of choosing a parameter setting for an aligner to yield a high-accuracy alignment (see [2,3]). A parameter advisor consists of two major components: (1) the advisor set of parameter choices used to generate candidate alignments, and (2) an advisor estimator that ranks alignments by estimated accuracy (see "Accuracy Estimator" to the right).

Accuracy Estimation

The accuracy of a multiple sequence alignment is measured as the fraction of substitutions from core columns of a reference alignment that are also present in the computed alignment output by an aligner. In practice, a reference alignment is not known (otherwise we would not be invoking an aligner), so accuracy values must be estimated.

Given a computed alignment, an accuracy estimator outputs a real number whose value should be positively correlated with the alignment's true accuracy. Our estimator Facet (Feature-based Accuracy Estimator) computes an accuracy estimate that is a linear combination of efficiently-computable feature functions (see [2,3]).

The plot below shows the correlation of the Facet accuracy estimator with alignment accuracy, for alternate alignments of standard benchmarks.

Availability

More information about adaptive local realignment, including a modified version of the opal1 aligner, which implements local and global advising using Facet, and example advisor sets, is available at: facet.cs.arizona.edu

References


Research supported by NSF Grant 09-21788, and NSF K2ER in Comparative Genomics DGE-0654435.