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 single 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.

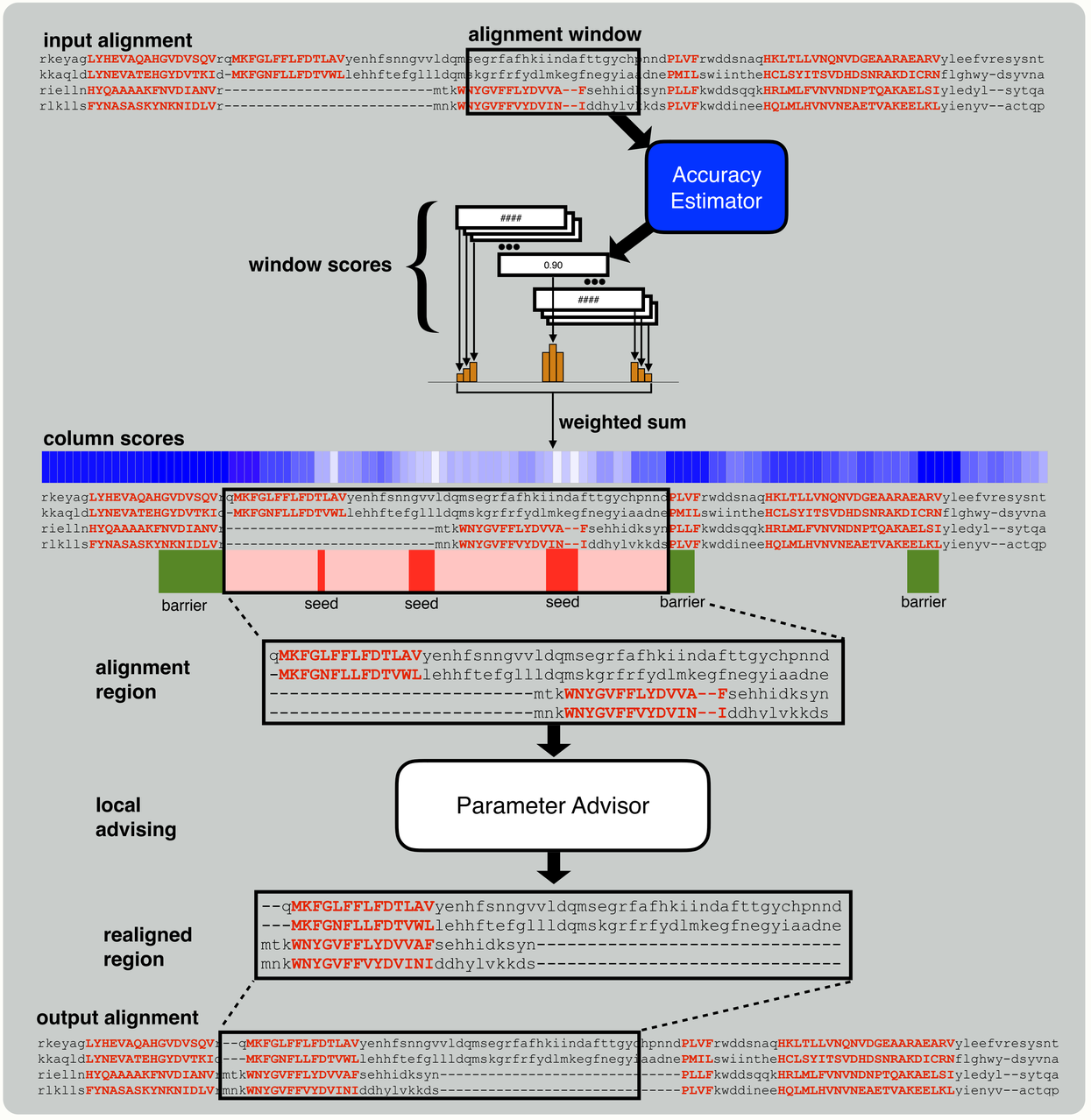
Window scores are generated using our accuracy estimator. A sliding **alignment window** induces a subalignment, which we score using the `Facet` alignment accuracy estimator. Each column of the alignment participates in several sliding windows, but is the center of a unique window.

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 central window for the column.

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.

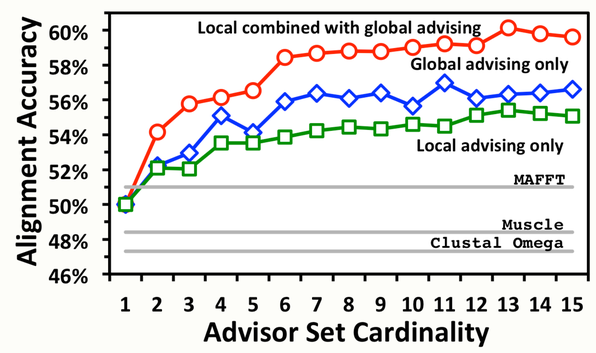
Local advising uses the parameter advising process described below (see “Parameter Advising") to find the aligner parameter choice that yields the realignment of highest estimated accuracy.

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.



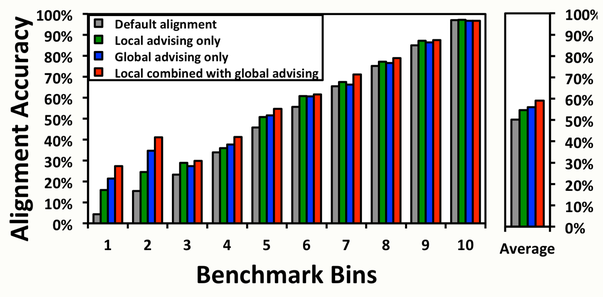
Experimental Results

Advising accuracy versus set cardinality



The cardinality of the advisor set used for both local and global advising is on the horizontal axis, while average accuracy is on the vertical axis. Accuracy is averaged over *difficulty bins* to correct for an overabundance of easy-to-align benchmarks, where each benchmark is binned by the accuracy of the default `Opal` [4] alignment. The average accuracy of alignments produced using **local advising only** on the default `Opal` alignments is shown in green. Accuracy achieved using **global advising only** with the `Opal` aligner is shown in blue. The **local combined with global advising** curve shows the accuracy of using global parameter advising on alignments that have been improved using adaptive local realignment. The grey lines show the accuracy of other popular aligners using default parameter choices and our weighting.

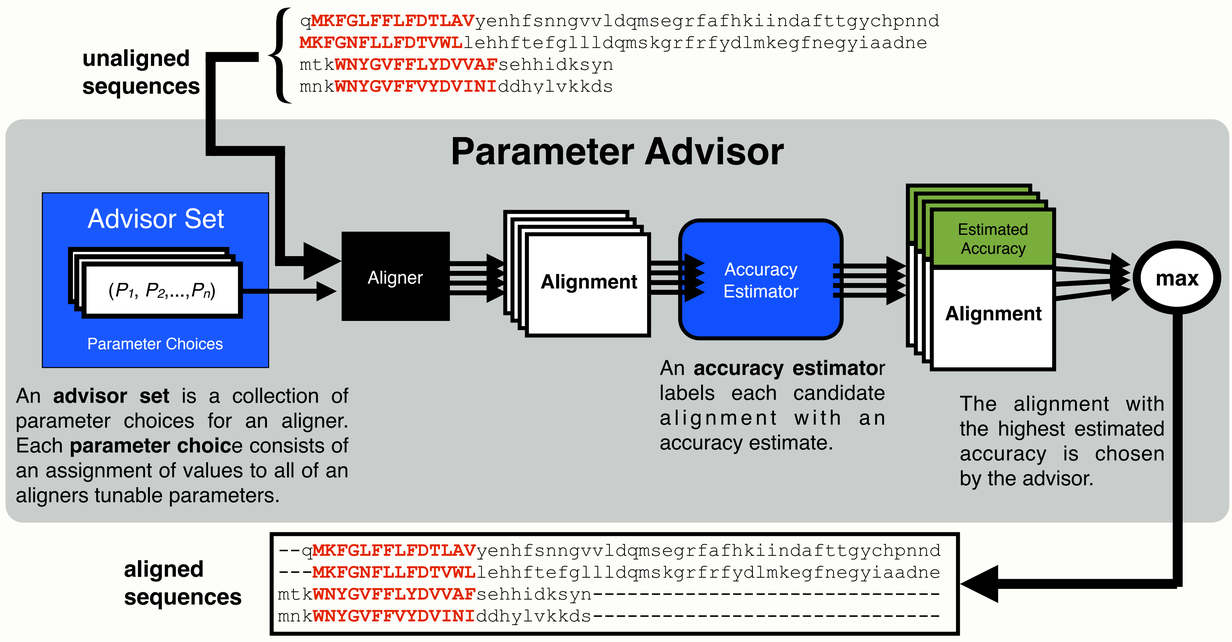
Advising accuracy within bins



The horizontal axis shows all ten benchmarks bins (described in the caption above), and vertical bars show the accuracy averaged over just the benchmarks in that bin. The rightmost chart shows the average across all bins (same as the plot above). The accuracy of the `Opal` **default alignment** is shown in grey, the accuracy of using **local advising only** on the default `Opal` alignment in green, **global advising only** using `Opal` in blue, and **local combined with global advising** (as described earlier) in red. All of the advising was performed using an advising set of cardinality 10.

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).

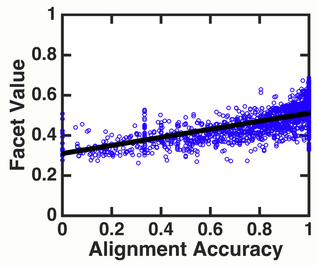


Accuracy Estimation

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` (**F**eature-based **Ac**curacy **Es**timator) 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 `Opal` aligner, which implements local and global advising using `Facet`, and example advisor sets, is available at:

facet.cs.arizona.edu

References

(1) **DeBlasio, D. and Kececioglu, J.** Boosting alignment accuracy through adaptive local realignment. *Submitted,* 2016.

(2) **DeBlasio, D.F., Wheeler, T.J., and Kececioglu, J.D.** Estimating the accuracy of multiple alignments and its use in parameter advising. Proceedings of the *International Conference on Research in Computational Molecular Biology (RECOMB*), April 2012.

(3) **Kececioglu, J and DeBlasio, D.** Accuracy estimation and parameter advising for protein multiple sequence alignment. *Journal of Computational Biology*, March 2013.

(4) **Wheeler, T.J., and Kececioglu, J.D.** Multiple alignment by aligning alignments. Proceedings of the 15th *ISCB Conference on Intelligent Systems for Molecular Biology, Bioinformatics,* July 2007.

Research supported by NSF Grant IIS-1217886, and
NSF IGERT in Comparative Genomics DGE-0654435.