Ensemble Multiple Sequence Alignment via Advising

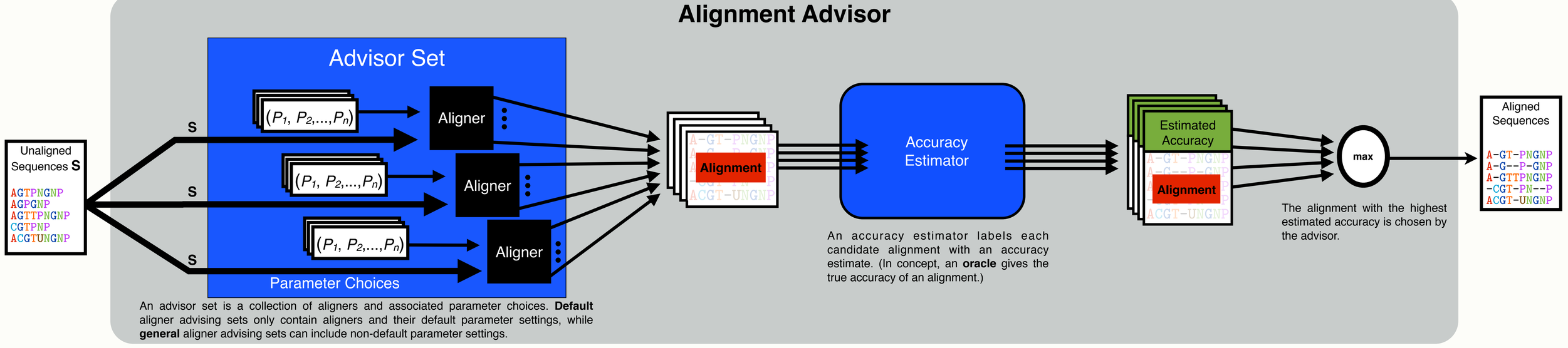
Dan DeBlasio and John Kececioglu
Department of Computer Science, The University of Arizona

Overview

The accuracy of multiple sequence alignments computed by an aligner for different settings of its parameters, as well as alignments computed by different aligners using their default settings, can differ markedly. **Parameter advising** is the task of choosing a parameter setting for an aligner so as to maximize the accuracy of the resulting alignment. We extend parameter advising to **aligner advising**, which chooses among a set of aligners to maximize accuracy. In the context of aligner advising, **default** advising selects from a set of aligners that are using their default settings, while **general** advising chooses both the aligner and its parameter setting.

We apply aligner advising for the first time to obtain a true **ensemble aligner**, that combines a collection of aligners and parameter settings to yield a new more accurate aligner. Through experiments on benchmark protein sequence alignments, we show that parameter advising for a fixed aligner gives a significant boost in accuracy over simply using its default setting, for the most accurate aligners currently used in practice. Furthermore, for ensemble alignment, default aligner advising gives a further boost in accuracy over parameter advising for any single aligner, and furthermore general aligner advising improves beyond default advising. Our new ensemble aligner that results from general aligner advising, when evaluated on standard suites of protein alignment benchmarks, and selecting from a set of four or more choices, is significantly more accurate than the best single default aligner.

Alignment Advisor



An advisor set is a collection of aligners and associated parameter choices. **Default** aligner advising sets only contain aligners and their default parameter settings, while **general** aligner advising sets can include non-default parameter settings.

An accuracy estimator labels each candidate alignment with an accuracy estimate. (In concept, an **oracle** gives the true accuracy of an alignment.)

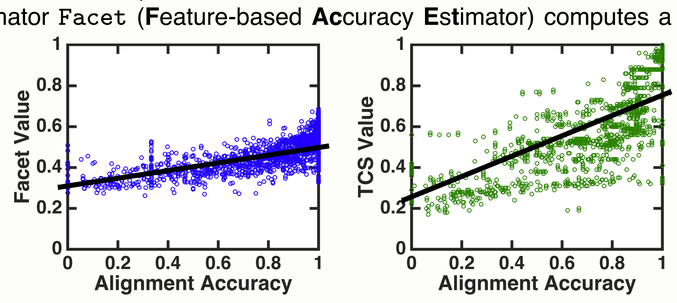
The alignment with the highest estimated accuracy is chosen by the advisor.

Accuracy Estimator

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.

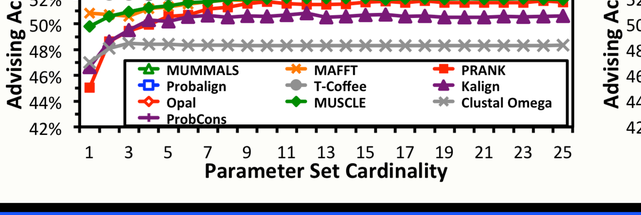
Given a computed alignment, an **accuracy estimator** outputs a real number whose value should correlate with the alignment's true accuracy. Our estimator `Facet` (**F**eature-based **Ac**curacy **Est**imator) computes a accuracy estimate that is a linear combination of efficiently-computable **feature functions** (see [5,6]).

The plots to the right show the correlation of `Facet` and `TCS` (**T**ransitive **C**onsistency **S**core [1]) with alignment accuracy, for alternate alignments of standard benchmarks.



Advisor Sets

An advisor can only be as good as the best alignment in its advisor set. Finding an optimal advisor set for a fixed estimator is **NP-complete**. For advisor sets of size k, we have shown that a greedy approach yields an (ℓ/k)**-approximation algorithm** for any constant ℓ**.** The greedy algorithm starts with an optimal parameter set of size ℓ, and repeatedly augments it with the parameter whose addition yields the highest advising accuracy. For small cardinalities ℓ, an optimal set can be found using exhaustive search. An **oracle** advising set is one that is optimal for an oracle advisor that knows the true accuracy of an alignment. Optimal oracle sets can be found even for very large cardinalities (see [3,4]).



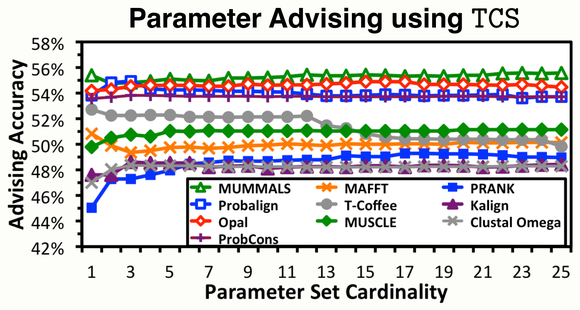
Parameter Advising

Parameter advising is the task of choosing a parameter setting for an aligner so as to maximize the accuracy of the resulting alignment. For 10 popular aligners we test the accuracy of a parameter advisor using both the `Facet` and `TCS` accuracy estimators. (The "Ensemble Aligner" section specifies how the aligners were selected.) For these aligners, we enumerated the Cartesian product of reasonable settings of their tunable parameters.

The figures below show the accuracy of advising across cardinalities, using both the `Facet` (left) and `TCS` (right) accuracy estimators on greedy advising sets.

Parameter Advising using `Facet`

Parameter Advising using `TCS`



Ensemble Aligner

The **default** aligner advisor uses a universe consisting of 17 of the most popular aligners, shown in the table to the right.

For **general** aligner advising, the universe consisted of a total of 863 parameter and aligner combinations. The 10 aligners for which we enumerated parameters were selected by:

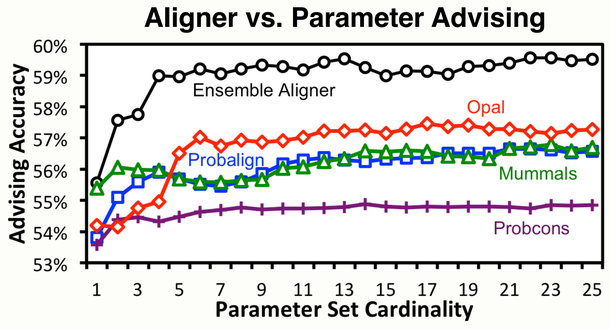
1. Finding an optimal oracle set of size $k = 5$ (`Kalign`, `MUMMALS`, `Opal`, `Probalign`, and `T-Coffee`).
2. Adding four aligners that are used extensively in the literature (`Clustal Omega`, `MAFFT`, `MUSCLE`, and `ProbCons`).
3. Constructing greedy sets for default aligner advising. These greedy sets contained all of the aligners already chosen above, with the addition of the `PRANK` aligner.

While default and general aligner advising eventually achieve the same maximum accuracy, general aligner advising does so at a smaller cardinality.

In the figures below, we use the term "aligner advising" to refer to general aligner advising.

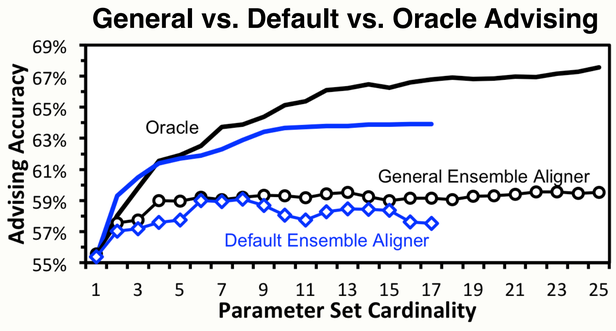
Aligners	
`Clustal`	Thompson, Higgins, and Gibson, 1994
`Clustal2`	Larkin, Blackshields, Brown, Chenna, *et al.*, 2007
`Clustal Omega`	Sievers, Wilm, Dineen, Gibson, Karplus, *et al.*, 2011
`DIALIGN`	Subramanian, Kaufmann and Morgenstern, 2008
`FSA`	Bradley, Roberts, Smoot, Juvekar, Do, *et al.* 2009
`Kalign`	Lassmann and Sonnhammer, 2005
`MAFFT`	Katoh, Kuma, Toh, and Miyata, 2005
`MUMMALS`	Pei and Grishin, 2006
`MUSCLE`	Edgar, 2004
`MSAProbs`	Liu, Schmidt, and Maskell, 2010
`Opal`	Wheeler and Kececioglu, 2007
`POA`	Lee, Grasso, and Sharlow, 2002
`PRANK`	Loytynoja and Goldman, 2005
`Probalign`	Roshan and Livesay, 2006
`ProbCons`	Do, Mahabhashyam, Brudno, and Batzoglou, 2005
`Sate`	Liu, Warnow, Holder, Nelesen, Yu, *et al.* 2011
`T-Coffee`	Notredame, Higgins, and Heringa, 2000

Aligner vs. Parameter Advising



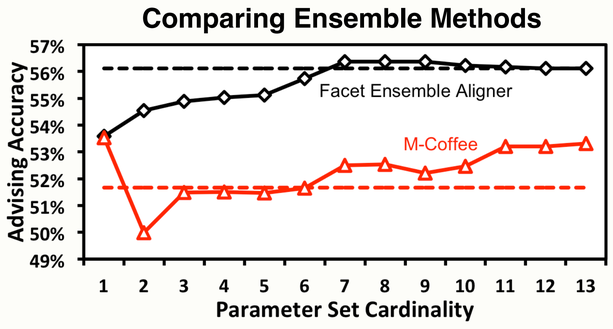
The horizontal axis shows the greedy set cardinality and the vertical axis is the advising accuracy for general aligner advising (black), and the four most accurate aligners using parameter advising. Aligner advising achieves a 2% boost in accuracy over parameter advising.

General vs. Default vs. Oracle Advising



The horizontal axis shows the advising set cardinality, and the vertical axis is the advising accuracy for general aligner advising (black), and default aligner advising (blue), on greedy advising sets (circles/diamonds) and the oracle (no marks). Notice that while default aligner advising plateaus at a similar advising accuracy, general aligner advising achieves this accuracy at a lower cardinality.

Comparing Ensemble Methods



The horizontal axis shows the advising set cardinality and the vertical axis is the advising accuracy, for aligner advising (black), and the M-Coffee aligner [7] (red). The dashed lines show ensemble accuracy using the default set of 6 aligners included in M-Coffee. Using the `Facet` estimator and aligner advising achieves a 4% boost in accuracy.

Future Work

- Extend advising from protein to DNA sequences
- Develop new feature functions that correlate more closely with true accuracy
- Expand the universe by enumerating more parameter choices for all aligners
- Include other popular aligners

Research supported by NSF Grant IIS-1217886

References

(1) **Chang J.M., DiTommaso P., Notredame C.** TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Molecular Biology and Evolution*, 2014.

(2) **DeBlasio, D. and Kececioglu, J.** Ensemble Multiple Sequence Alignment via Advising. Submitted to *ACM Conference on Bioinformatics, Computational Biology and Health Informatics (BCB)*, 2015.

(3) **DeBlasio, D and Kececioglu, J.** Learning Parameter-Advising Sets for Multiple Sequence Alignment. *ACM/IEEE Transactions on Computational Biology and Bioinformatics.* In press, 2015 (early access online).

(4) **DeBlasio, D and Kececioglu, J.** Learning parameter sets for alignment advising. Proceedings of *ACM Conference on Bioinformatics, Computational Biology and Health Informatics (BCB)*, September 2014.

(5) **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.

(6) **Kececioglu, J and DeBlasio, D.** Accuracy Estimation and Parameter Advising for Protein Multiple Sequence Alignment. *Journal of Computational Biology*, March 2013.

(7) **Wallace IM, O'Sullivan O, Higgins DG, Notredame C.** M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research*, March 2006.