Learning Advisors for Multiple Sequence Alignment

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

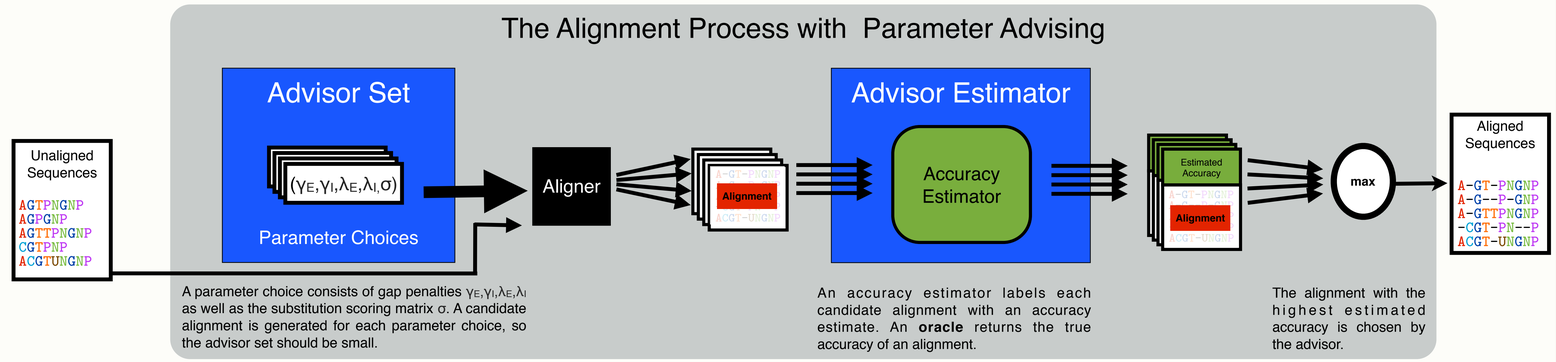
Overview

While the multiple sequence alignment output by an aligner strongly depends on the parameter values used for its alignment scoring function (i.e. choice of gap penalties and substitution scores), most users rely on the single default parameter setting. A different parameter setting, however, might yield a much higher-quality alignment for a specific set of input sequences. The problem of picking a good choice of parameter values for a given set of input sequences is called parameter advising. A **parameter advisor** has two ingredients: (i) a **set** of parameter choices to select from, and (ii) an **estimator** that estimates the accuracy of a computed alignment; the parameter advisor then picks the parameter choice from the set whose resulting alignment has highest estimated accuracy.

Our estimator `Facet` (**F**eature-based **Ac**curacy **E**stimator) is a linear combination of real-valued feature functions of an alignment. We assume the feature functions are given as well as the universe of parameter choices from which the advisor's set is drawn. For this scenario we define the problem of learning an optimal advisor by finding the best possible set, or estimator, for a collection of training data of reference alignments. Learning optimal advisors is **NP-complete.** For the advisor sets problem, we develop a fast approximation algorithm that finds near optimal sets. For the advisor estimator problem, we have an efficient method for finding the coefficients for the estimator that performs well in practice.

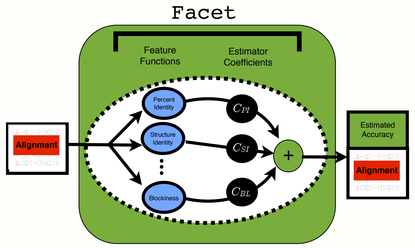
Parameter Advising

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.



Accuracy Estimation

Given a computed alignment, the advisor *estimator* outputs a real number. This value should correlate with the true accuracy of the alignment. Our estimator `Facet` (**F**eature-based **Ac**curacy **E**stimator) computes a value that is a linear combination of efficiently-computable *feature functions.*

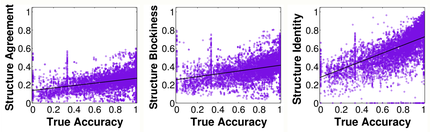


Feature Functions

Features used in `Facet` are

- real-valued functions of an alignment,
- efficiently computable, and
- correlate positively with true accuracy.

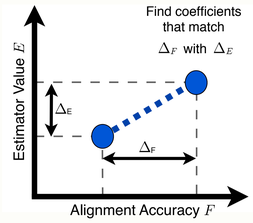
The set of features contains sequence-based measures such as **Percent Identity** and **Gap Frequency**. The most accurate features utilize protein secondary structure. Features that use structure include **Secondary Structure Blockiness, Secondary Structure Identity** and **Secondary Structure Agreement**. The correlation between these features and true accuracy are shown below.



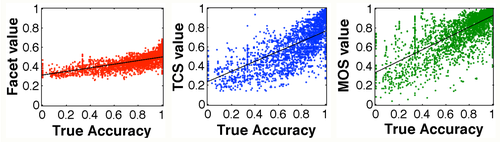
Estimator Coefficients

The problem of finding optimal coefficients for advising, given a set of features and parameter choices, is **NP-complete.**

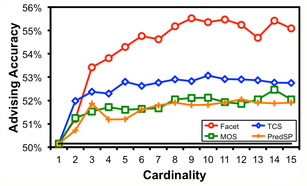
Since the estimator is used to rank alignments by estimated accuracy, we find coefficients for `Facet` by matching the difference in estimator value E to the difference in true accuracy F for pairs of example alignments. This difference-fitted estimator performs well in practice.



Experimental Results



Correlation between estimators and true accuracy. To work well for advising, an accuracy estimator should have small spread and a large slope. `Facet` combines feature functions that trend well with accuracy, achieving better spread than other estimators, while retaining positive slope.



Average advising accuracy of estimators on sets of varying cardinalities. The average true accuracy of the alignment chosen by an estimator from the greedy advisor set, averaged over weighted benchmarks, using 12-fold cross validation, is shown. The optimal default parameter choice achieves a weighted accuracy of 50%.

Learning Parameter Sets

The Advisor Set Problem

Given

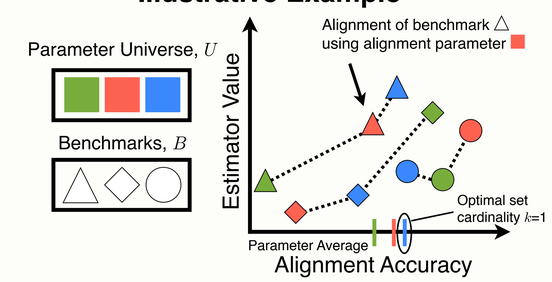
- a universe U of alignment **parameter choices,**
- a collection B of training **benchmarks,** each consisting of a set of input sequences and a reference alignment,
- an upper bound k on the **set cardinality,**
- an **accuracy estimator** E,

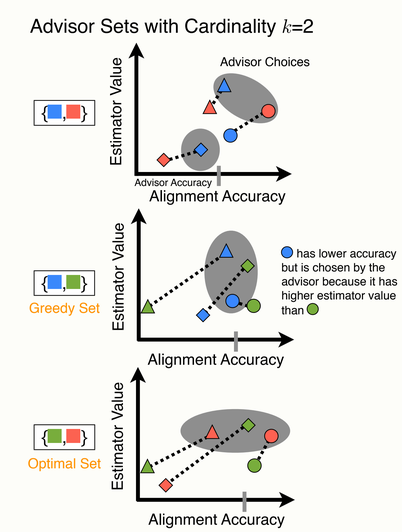
find an advisor set $P \subseteq U$, with $|P| \leq k$, that maximizes the average true accuracy of the alignments chosen by the advisor on the training benchmarks.

Theoretical Results

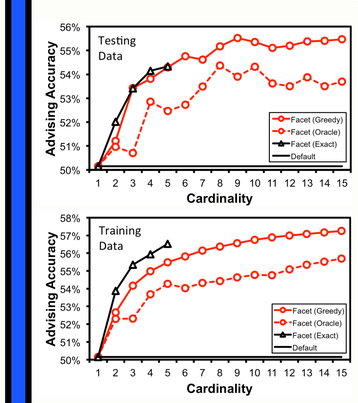
- Finding an optimal advisor set for a fixed estimator is **NP-complete**.
  - For small cardinalities, an optimal set can be found by exhaustive search.
- A near-optimal set can be efficiently found by an $\frac{\ell}{k}$**-approximation algorithm**.
  - Algorithm is given an optimal set of size $\ell < k$.
  - Actually outperforms the optimal set on testing benchmarks.
  - Approximation algorithm is **greedy**, and runs in polynomial time.
- An optimal advisor set for an oracle can be efficiently found in practice. We call this an **oracle set**.

Illustrative Example

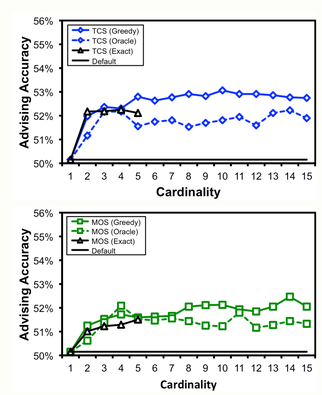




Experimental Results



Average advising accuracy of `Facet` on advisor sets of varying cardinalities, found by different methods. The average true accuracy of the alignment chosen by an estimator, averaged over weighted benchmarks, on both testing and training benchmarks, using 12-fold cross validation, is shown. In addition to the optimal *exact set* of a given cardinality, we show the *greedy set* found by the approximation algorithm, and the *oracle set.* The benchmarks are weighted so the optimal default parameter choice achieves an accuracy of 50%. Note that the accuracy of the exact set and the greedy set are similar on the testing benchmarks.



Average advising accuracy of `TCS` and `MOS` on advisor sets of varying cardinalities, found by different methods. The average true accuracy of the alignment chosen by an estimator, averaged over weighted benchmarks, on testing benchmarks, using 12-fold cross validation, is shown. In addition to the optimal *exact set* of a given cardinality, we show the *greedy set* found by the approximation algorithm, and the *oracle set.* The benchmarks are weighted so the optimal default parameter choice achieves an accuracy of 50%. Note that the advising accuracy of the exact set is lower than the greedy set.

Software

`facet.cs.arizona.edu`

`Facet` is implemented in `Java`, and is able to be run as a stand-alone program, or integrated into an existing application.

Parameter sets, alignment benchmarks, and scripts are available for download at the website above.

Research supported by NSF Grant IIS-1217886
Travel funding to ISMB 2014 was generously provided by ISCB

Future Work

- Extend advising from protein to DNA sequences
- Extend parameter choices to include the selection of aligner
- Develop new feature functions that correlate more closely with true accuracy

References

Ahola V. Aittokallio T. Vihinen M. Uusipaikka E. **Model-based prediction of sequence alignment quality.** Bioinformatics. 2008;24:2165–2171. (`PredSP`)

Chang J.M., DiTommaso P., Notredame C. **TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction**. Mol. Biol. Evol. 2014;31.

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

Lassmann T. Sonnhammer E.L.L. **Automatic assessment of alignment quality.** Nucleic Acids Research. 2005;33:7120–7128. (`MOS`)

Thompson J.D. Plewniak F. Ripp R. Thierry J.C. Poch O. **Towards a reliable objective function for multiple sequence alignments.** J Mol Biol. 2001;314:937–51 (`NORMD`)

Penn O. Privman E. Landan G. Graur D. Pupko T. **An alignment confidence score capturing robustness to guide tree uncertainty. Molecular Biology and Evolution**. 2010a;27:1759–1767.(`Guidance`)