Improving the Quality of Protein Sequence Alignments by Estimating their Accuracy

Dan DeBlasio and John Kececioglu

Department of Computer Science, University of Arizona

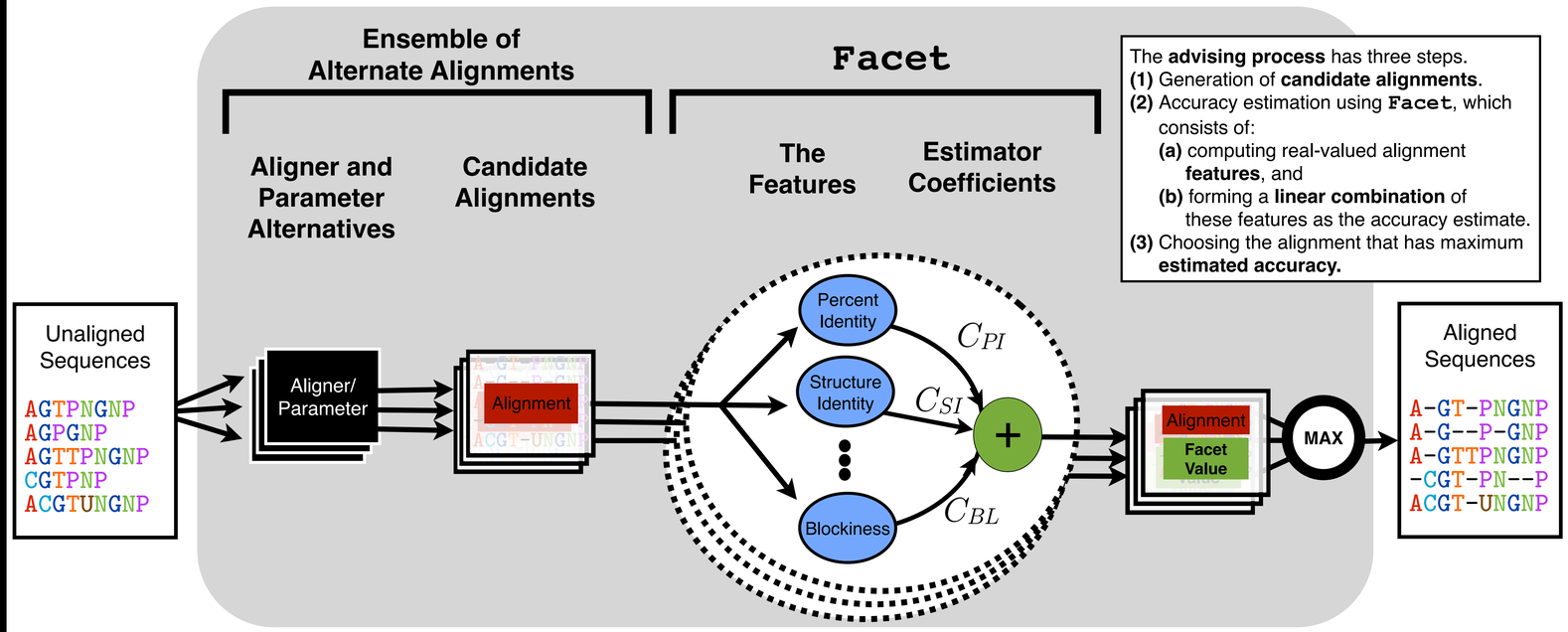
Overview

New technical advances in next-generation sequencing have provided biologists with massive amounts of DNA and protein data. A non-trivial step in the analysis of such data is aligning similar sequences for comparative studies. Each alignment tool offers different strengths and weaknesses. Aligners often have many user-specified parameters that can greatly affect the accuracy of the computed alignment, and users often rely on the default parameter setting. Researchers are forced to either use this default setting, or spend considerable time finding a suitable alternative. For a set of input sequences to align, our tool **Facet** (**f**eature-based **ac**curacy **est**imator) selects a good aligner and a good parameter setting. Facet does this by combining alignment features into an accuracy estimator. These independent features are informed by our knowledge of how proteins evolve and fold. Using Facet to choose a parameter setting improves alignment accuracy by up to 27% over the best default setting.

Facet is freely available at

facet.cs.arizona.edu

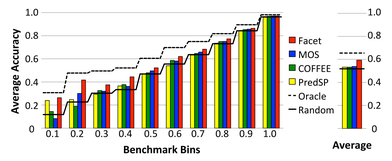
Advising



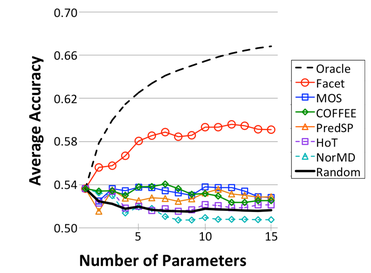
The **advising process** has three steps.
(1) Generation of **candidate alignments**.
(2) Accuracy estimation using **Facet**, which consists of:
(a) computing real-valued alignment **features**, and
(b) forming a **linear combination** of these features as the accuracy estimate.
(3) Choosing the alignment that has maximum **estimated accuracy.**

Results

Parameter Advising

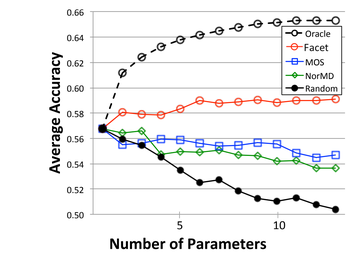


Average advising accuracy of competing estimators. The benchmarks are divided into bins based on the accuracy of their alignment when using the single best parameter choice. Each of these benchmarks is then realigned using an ensemble of 10 parameter settings. The figure shows the average accuracy of the alignment chosen using the competing estimators for each bin (left) and over all bins (right).



Average accuracy of alignments chosen using competing estimators when varying the parameter ensemble cardinality. The graph shows the accuracy of an estimator, averaged over all bins, when using a parameter ensemble of a given cardinality.

Aligner Advising



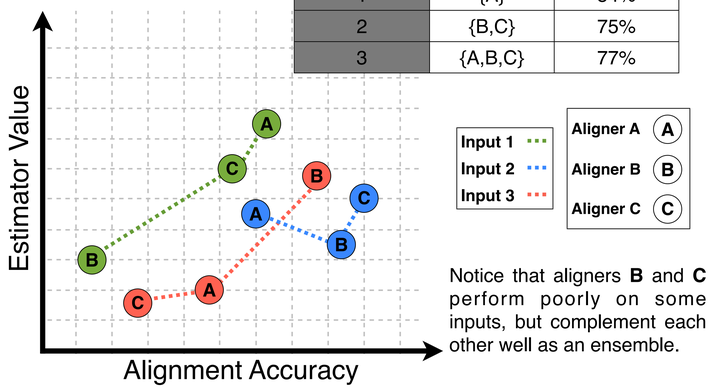
Average accuracy of alignments chosen using competing estimators when varying the aligner ensemble cardinality. The graph shows the accuracy of an estimator, averaged over all bins, when using an aligner ensemble of a given cardinality.

Alternate Alignments

Choosing the ensemble of parameters or aligners that will produce the candidate alignments for advising is very important. If the candidate alignments for an input are all poor, the chosen alignment will also be poor. The cardinality of the ensemble should be small to reduce the time for generating the candidates. Given an input cardinality *k*, we use an **integer linear program** to find the optimal ensemble that provides the best candidate alignments for advising. An ensemble can be optimized either for an **oracle**, which always returns the true accuracy, or a given estimator.

The figure shows an example of the problem, for three sequence inputs and an ensemble of three parameter settings or aligners. We show the estimator value versus the true accuracy of the alignment produced on each input by each member of the ensemble. Colors identify the sequence inputs and labels identify the ensemble member.

Ensemble cardinality	Ensemble members	Average accuracy
1	{A}	54%
2	{B,C}	75%
3	{A,B,C}	77%



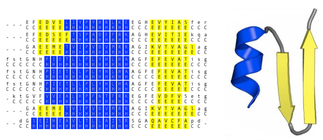
Notice that aligners **B** and **C** perform poorly on some inputs, but complement each other well as an ensemble.

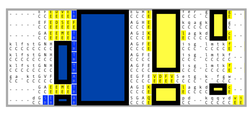
Facet: A Feature-Based Accuracy Estimator

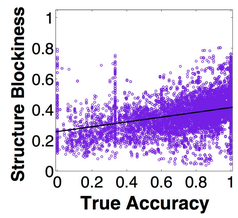
The Features

The real-valued features used by **Facet** measure characteristics of alignments that ideally correlate with true accuracy. The set of features contain sequence-based measures, such as percent identity, information content, and gap frequency, and secondary-structure-based measures. The structure-based features tend to be the most indicative of high-accuracy alignments.

Protein secondary structure is a labeling of the sequence residues by one of three structure types: α-helix (blue), β-sheet (yellow) and coil (grey). The figure shows an alignment labeled by its predicted structure *(left),* and a schematic of the folded structure[1] *(right).*





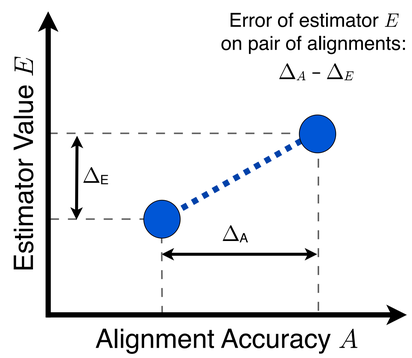


Each feature has a positive correlation with true accuracy when measured on candidate alignments, but no single feature is a good estimator on its own. The most informative feature (the one with the highest coefficient) is **Secondary Structure Blockiness**, which finds a covering of an alignment by *blocks* (contiguous columns on a subset of rows with the same structure type) that maximizes the number of pairs of aligned residues in the blocks. The figure on the left shows a covering by blocks (as bold rectangles) and the correlation of Blockiness with alignment accuracy. Each point in the scatter plot represents one alignment, with its associated Blockiness value and true accuracy.

Estimator Coefficients

The **Facet** value is a linear combination of feature values whose optimal coefficients are found using a **linear** or **quadratic program**. When used for advising, an estimator will *rank* alignments, and we want to set its coefficients to minimize the error for this task.

For a set of example alignments, we examine every pair of alignments to find out if Facet is ranking them correctly. On each pair, we want the Facet estimator to match the difference in accuracy. The *error* is the amount by which Facet underestimates this difference. The optimal coefficients $C_{PI}, C_{SI}, \ldots, C_{BL}$ minimize this error.



Reference:
J. Kececioglu and D. DeBlasio, **Accuracy Estimation and Parameter Advising for Protein Multiple Sequence Alignment**, Journal of Computational Biology 20(4), pp. 259-279, 2013.

Research supported by the NSF IGERT Grant in Comparative Genomics DGE-0654435

Footnote:
1. Figure from http://www.ebi.ac.uk/training/online, used under the creative commons license