Facet: a feature-based accuracy-estimation tool for protein multiple sequence alignments

Dan DeBlasio and John Kececioglu

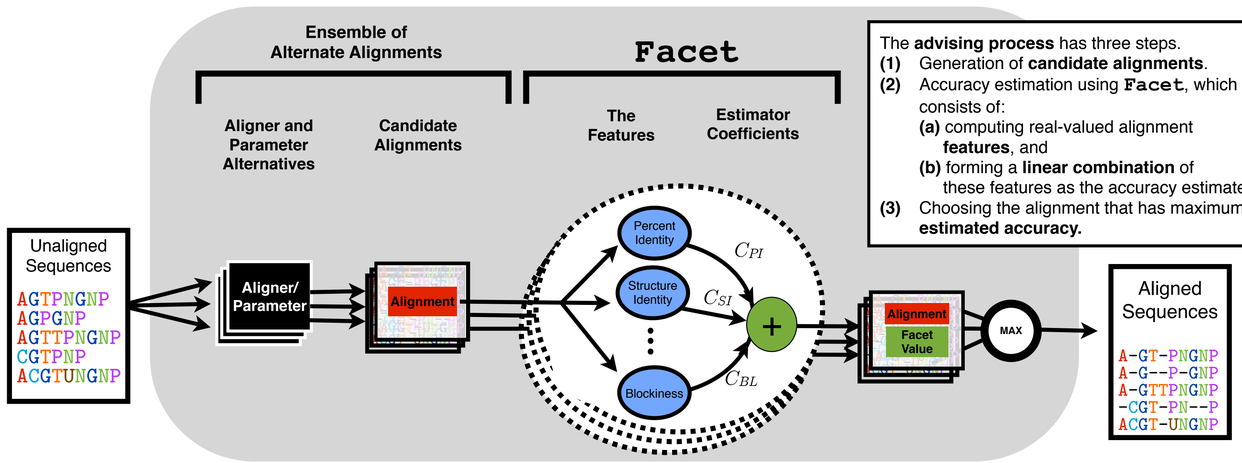
Department of Computer Science, The University of Arizona

Overview

Selecting an aligner — and parameter values for the aligner's scoring function — to obtain a quality alignment of a specific set of sequences can be challenging. Different aligners and different parameter values can produce vastly different alignments of the same sequences. In principle, a user could simply try various aligners and parameter settings, and choose the resulting alignment that is the most *accurate* (the alignment that best agrees with the correct alignment of the sequences), except that in practice the accuracy of an alignment typically cannot be measured (since the correct alignment is not known). We overcome this obstacle by combining efficiently-computable, real-valued *features* of an alignment into an accuracy *estimator* that is suitable for choosing both aligners and parameter settings.

Facet ("Feature-based Accuracy Estimator") is an easy-to-use, open-source utility for estimating the accuracy of a protein multiple sequence alignment. Facet can be readily applied to both *parameter advising* (choosing good parameter values) and *aligner advising* (choosing a good aligner). For the accuracy estimator, which is linear in the alignment features, the tool provides optimized default coefficients that are best on average (coefficients may also be specified manually), and can be run as a stand-alone tool, or included in any pre-existing Java application. For boosting alignment accuracy by parameter advising, the Facet website provides optimal pre-computed parameter sets (namely, substitution matrices and affine gap penalties).

Facet is available at **facet.cs.arizona.edu**



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.**

Applying Facet to parameter advising and aligner advising improves alignment accuracy by as much as 27% on the most challenging benchmarks.

Citation

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.

Example

Facet can be run as a stand-alone program, by executing a shell script that invokes the Java application, or by calling the Facet method from within the user's Java code. The input to the shell script is three files: a multiple sequence alignment file in FASTA format, a secondary structure prediction file, and the corresponding structure probability file. Secondary structure must first be predicted for the input sequences using PSIPRED (configuration and formatting scripts are provided). In the example below, Facet scores are computed for three alternate alignments (align1.fa, align2.fa, align3.fa) of the same sequences (seqs.fa).

Only predict structure once

Facet values on 'standard out'

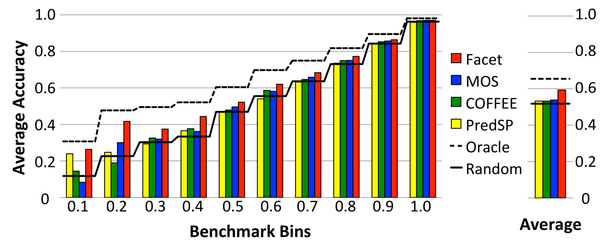
```
$./PSIPRED_wrapper.pl seqs.fa > seqs_struc 2> seqs_prob
$./FACET.sh align1.fa seqs_struc seqs_prob
align1.fa    0.565
$./FACET.sh align2.fa seqs_struc seqs_prob
align2.fa    0.868
$./FACET.sh align3.fa seqs_struc seqs_prob
align3.fa    0.342
```

Including Facet into existing code can be done by a single call to the Facet.estimator method. The FacetAlignment object encapsulates the sequence alignment and structure prediction (and takes arrays specifying the aligned sequences, the structure prediction and the structural probabilities).

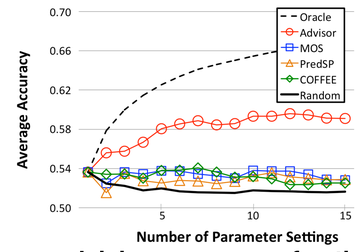
```
FacetAlignment align1 = new FacetAlignment(alignedSeqs1,strucPred,strucProb);
FacetAlignment align2 = new FacetAlignment(alignedSeqs2,strucPred,strucProb);

if(Facet.estimator(align1) > Facet.estimator(align2))
  return alignedSeqs1;
else
  return alignedSeqs2;
```

Results

Parameter advising

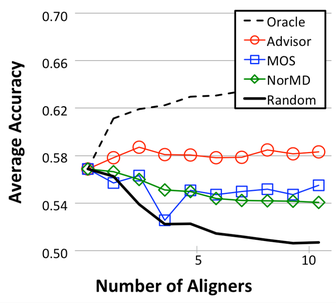


Average advising accuracy for estimators from the literature. The benchmarks are divided into bins based on the true accuracy of the alignment computed with Opal using the single best parameter setting. Each of these benchmarks is then realigned with Opal using an optimal ensemble of 10 parameter settings. The average true accuracy of the alignment chosen using various estimators is shown.



Average advising accuracy of estimators when varying the cardinality of the parameter ensemble. The average true accuracy of the alignment chosen by an estimator, averaged over all benchmark bins, using an optimal parameter ensemble of a given cardinality, is shown.

Aligner advising



Average advising accuracy of estimators when varying the cardinality of the aligner ensemble. The average true accuracy of an estimator, averaged over all bins, when using an optimal aligner ensemble of a given cardinality, is shown. The aligners most often used in the optimal ensemble are: T-Coffee (Notredame, *et al.* 2000), Kalign (Lassman 2005), Mummals (Pei and Grishin 2005), Probalign (Roshan and Liesay 2006), ClustalW (Larkin, *et al.* 2007), and Opal (Wheeler and Kececioglu 2007).

Estimators from the literature
COFFEE (Notredame *et al.* 1998), NORMD (Thompson, *et al.* 2001), MOS (Lassman and Sonnhammer 2005), and PredSP (Ahola *et al.* 2008).

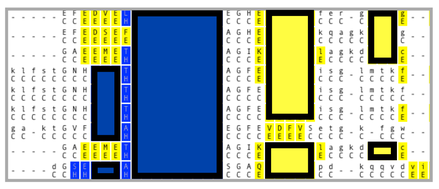
Accuracy estimation method

The features

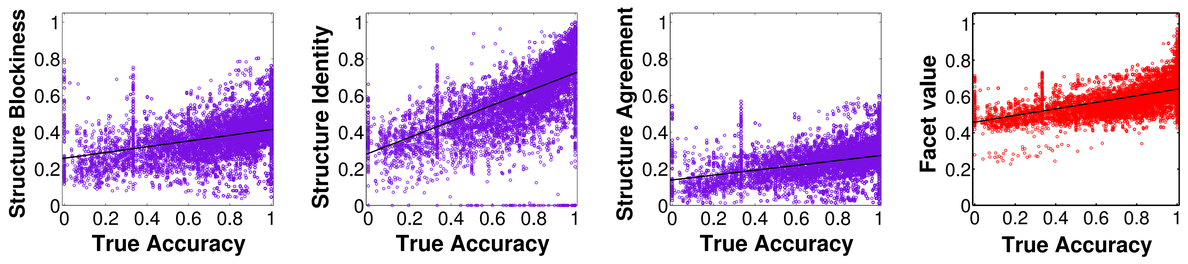
The real-valued features used by Facet measure characteristics of alignments that ideally correlate with true accuracy. The set of features contains sequence-based measures such as **precent identity**, **information content**, and **gap frequency**, as well as several secondary-structure-based measures. The structure-based measures tend to be the strongest features for identifying high-accuracy alignments.

Protein secondary structure is a labeling of the residues in the sequence 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 (right).

From http://www.ebi.ac.uk/training/online



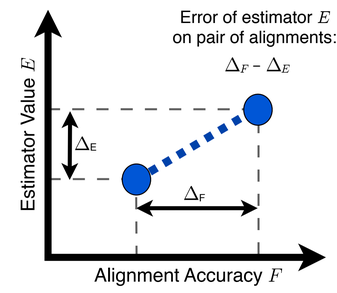
Each feature has a positive correlation with true accuracy when measured on candidate alignments, but no single feature is sufficient by itself for a good estimator. The most informative feature (with the largest coefficient) is **Secondary Structure Blockiness**, which finds a packing of alignment *blocks* (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 packing of blocks (the bold rectangles) and below is a scatter plot of Blockiness versus alignment accuracy. Each point in the scatter plot represents one alignment, with its associated Blockiness value and true accuracy.



In addition to Blockiness, features that have high coefficients are **Gap Open Density**, **Secondary Structure Agreement** (the probability that an aligned residue pair has the same structure, averaged over all pairs), **Secondary Structure Identity**, and **Average Substitution Score** (BLOSUM62 substitution averaged over aligned residue pairs). Scatter plots of Secondary Structure Identity and Secondary Structure Agreement are shown above, along with the value of the Facet estimator.

Estimator coefficients

The Facet estimator value is a linear combination of feature values whose optimal coefficients are found by solving a *linear programming problem*. When used for advising, an estimator *ranks* alignments; the linear program finds optimal coefficients that minimize the error for this task.



Given a training set of example alignments, we consider how Facet ranks each pair of alignments. On each pair (A,B), we want the Facet estimator E to match the difference in true accuracy F. The *error* e_{AB} is the amount by which Facet underestimates this difference. The optimal coefficients C_{PI}, C_{SI},...,C_{BL} minimize the total error.

$$\text{minimize} \quad \sum_{(A,B)} w_{AB}\, e_{AB}$$

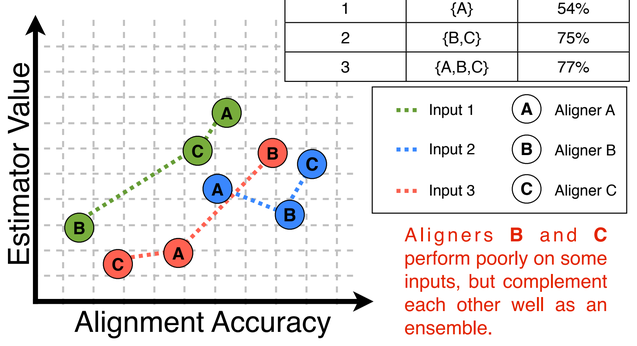
$$\text{subject to} \quad e_{AB} \geq \max \left\{ \begin{array}{l} 0, \\ \Big(F(B) - F(A)\Big) \\ -\Big(E(B) - E(A)\Big) \end{array} \right\}$$

In the optimization problem on the left, a pair of example alignments is weighted by w_{AB}, so that each benchmark bin has the same total weight. $E(A)$ is a linear expression in the coefficients C_{PI},C_{SI},...,C_{BL}. $F(A)$ is a constant for a given alignment A. Consequently, the optimization problem is a linear program in the coefficient variables and error variables e_{AB}.

Alignment ensemble

Choosing the ensemble of parameters or aligners that will produce the candidate alignments for advising is crucial. 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 solve an *integer linear program* to find the optimal ensemble that provides the best candidate alignments for advising. Using CPLEX, we can find optimal ensembles up to cardinality 15, drawn from a universe of over 2,000 parameter settings.

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



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

Acknowledgements

Research supported by the NSF IGERT Grant in Comparative Genomics DGE-0654435 and NSF Grant IIS-1217886. Travel funding to ISMB/ECCB 2013 was generously provided by the National Science Foundation.