Overview

Facet 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 beat 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).

Applying Facet to parameter advising and aligner advising improves alignment accuracy by as much as 25% on the most challenging benchmarks.

Citation

Accuracy estimation method

The features

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 the 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 tests, using an optimal parameter ensemble of a given cardinality, is shown.

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 10, drawn from a universe of over 2,000 parameter settings.

Acknowledgements
Research supported by the NSF IGERT Grant in Comparative Genomics DOE-0654435 and NSF Grant IIS-1217286. Travel funding to ISMB/ECBB 2013 was generously provided by the National Science Foundation.
Correlation of Estimators with Accuracy

Each scatter plot shows the value of an estimator versus true accuracy for all alignments in the 15-parameter test set.

Rank of Chosen Alignment

The average rank of the alignment chosen by each advising method. The bar chart is on the 10-parameter set across accuracy bins. The line plot is the rank averaged over all accuracy bins, using an optimal parameter ensemble of a given carnality.