Overview

New technical advances in next-generation sequencing have provided biomedicalists 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 (feature-based accuracy estimator) 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

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 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.

Facet: A Feature-Based Accuracy Estimator

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(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 pick 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 \( C_{\alpha}, C_{\beta}, \ldots, C_{\gamma} \) minimize this error.

Results

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.