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 (Feature-based Accuracy Estimator) 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. The optimal advisor 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) an advisor set of parameter choices used to generate candidate alignments, and (2) an estimator that ranks alignments by estimated accuracy.

The Alignment Process with Parameter Advising

Parameter Estimation

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 best accurate features utilize protein secondary structure. Features that use structure include Secondary Structure Blockiness, Secondary Structure Identity and Secondary Structure Agreement. Correlation between these features and true accuracy are shown below.

Accuracy Estimation

The feature-based estimator for Facet computes a value that is a linear combination of efficiently computable function values. 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 \( \Delta E \) to the difference in true accuracy \( \Delta A \) for pairs of example alignments. This difference-fitted estimator performs well in practice.

Experimental Results

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 all benchmark sets, is shown. The optimal parameter choice achieves a weighted accuracy of 50%.

Future Work

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

References


Facet: a feature-based accuracy-estimation tool

Facet is available at facet.cs.arizona.edu

Including Facet into existing code can be done by a single call to theFacet 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).