Poster H-79

CONTRAlign 2: Discriminative Training for Multiple Sequence Alignment

Authors:
Chuong B. Do (*Stanford University*)
Samuel S. Gross (*Stanford University*)
Robert C. Edgar (*Stanford University*)
Kazutaka Katoh (*Kyoto University*)
Serafim Batzoglou (*Stanford University*)

Short Abstract: We present CONTRAlign 2, an extension of the CONTRAlign conditional random field pairwise aligner to consistency-based progressive multiple sequence alignment. CONTRAlign 2 improves upon its predecessor by incorporating an array of new sequence features designed to exploit the additional signal found in the alignment of multiple protein sequence sets.

Long Abstract:
Comparative analysis or structure prediction for sets of related protein sequences typically begins with identifying patterns of amino acid substitution via protein sequence alignment. While the evolutionary information obtained from alignments can provide insights into protein structure, constructing accurate alignments may be difficult when proteins share significant structural similarity but little sequence similarity. Indeed, for modern alignment tools, alignment quality drops rapidly when the sequences compared have lower than 25\% identity, the "twilight zone" of protein alignment.

In recent years, most alignment methods that have claimed improvements in alignment accuracy have done so not by proposing substantially new algorithms for alignment but rather by incorporating additional sources of information. For instance, when structures of some sequences are available, the 3DCoffee program uses pairwise alignments from existing threading-based (FUGUE) and structural (SAP and LSQman) alignment tools to guide sequence alignment construction. When homologous sequences are available and computational expense is of less concern, the PRALINE-PSI program uses PSI-BLAST--derived sequence profiles to augment the amount of evolutionary information available to the aligner. The SPEM program takes the additional step of heuristically incorporating PSIPRED predictions of protein secondary structure, a strategy also adopted in the latest version of PRALINE-PSI.

As these programs demonstrate, incorporating additional information can often yield considerable benefits to alignment quality. However, choosing parameters for more complex models can be difficult. In traditional dynamic-programming--based alignment programs, log-odds--based substitution matrices are estimated from large external databases of aligned protein blocks, and gap parameters are typically ``hand chosen'' to maximize performance on benchmark tests. When dealing with more expressive models, however, the high-dimensionality of the parameter space hinders such manual procedures. From the perspective of numerical optimization, the non-convexity of aligner performance as a function

of parameters makes hand-tuning difficult for alignment algorithms that rely on complicated ad hoc scoring schemes.

Furthermore, optimizing benchmark performance often leads to overfitting, a situation in which the selected parameters are nearly optimal for training benchmark alignments but work poorly on new test data. To combat overfitting, many machine learning studies make use of cross-validation, a technique in which an algorithm is trained and tested on independent data sets in order to estimate the ability of the method to generalize to new situations.

Previously, we presented CONTRAlign, an extensible and fully automatic framework for parameter selection and protein pairwise sequence alignment based on a probabilistic model known as a pair conditional random field (pair-CRF). In the CONTRAlign methodology, the user first defines an appropriate model topology for pairwise alignment. Unlike for ad hoc algorithms in which model complexity (and hence risk of overfitting) corresponds roughly with the number of free parameters in the model, the effective complexity of a CONTRAlign pair-CRF--based model is controlled by a set of regularization parameters, allowing the user to adjust the trade-off between model expressivity and the risk of overfitting. Given a set of gold standard partially labeled alignments, CONTRAlign uses gradient-based optimization and holdout cross validation to automatically determine regularization constants and a set of alignment parameters with good expected performance for future alignment problems.

In this paper, we describe CONTRAlign 2, an extension of the original CONTRAlign conditional random field pairwise aligner to consistency-based progressive multiple sequence alignment. Like pairwise CONTRAlign, CONTRAlign 2 can learn both substitution and gap parameters that generalize well to previously unseen sequences using as few as 20 multiple sequence alignments, as demonstrated through stringent cross-validation conditions. However, CONTRAlign 2 extends upon its predecessor by incorporating an array of new sequence features designed to exploit the additional signal found in the alignment of multiple protein sequence sets. With these improvements, CONTRAlign 2 achieves state-of-the-art accuracy for protein multiple sequence alignment, with significant gains over existing multiple alignment techniques for the case of low-identity sequences.