A Supervised Learning Approach for Detecting Significant Local Alignments

Eric A. Breimer,[1] **Mark K. Goldberg**[2]

Keywords: local sequence alignment, supervised learning, mosaic effect

1 Introduction

It was observed (see [1], [2]) that the Smith-Waterman algorithm for local sequence alignment has two essential flaws: it often finds long alignments with a high score and misses shorter ones with a higher degree of similarity (the *shadow effect*); and it often combines two or more segments of high similarity and aligns internal segments that are not related (the *mosaic effect*).

Arslan *et. al.* [2] proposed an $O(n^2 \log n)$ algorithm that outputs a *normalized local alignment* which maximizes the degree of similarity (alignment score divided by alignment length) rather than the total similarity score. Given a properly selected normalization parameter, the new algorithm eliminates both the shadow and mosaic effects. Unfortunately, determining a proper normalization parameter requires repeated executions with different parameter values, and also expert feedback to determine the usefulness of the alignments.

We use the supervised learning approach to construct an $O(n^2)$ algorithm that eliminates the mosaic effect while requiring no expert feedback to produce meaningful alignments. Given inputs comprised of pairs of sequences with known *motifs*, *i.e.*, an alignment that captures a biologically significant similarity as defined by an expert, our learning algorithm determines parameter values needed to extract motifs from a few top-scoring, possibly sub-optimal alignments. The expectation is that the algorithm evolved from learning will align and extract such motifs from unseen input sequences. We provide an automatic framework for using existing motifs to tune the post-processing of sub-optimal alignments.

2 Learning Approach

We use a modification of the Smith-Waterman algorithm similar to that proposed by Barton in [3] to output all non-overlapping maximal scoring alignments to see if a known motif is discovered, *i.e.*, contained within a sub-optimal alignment. The first parameter learned is the number of top scoring alignments that must be out-putted in order to guarantee that a motif is found. Our experiments show that if a motif is discovered at all, it will be among the top k scoring alignments, where k is the number of expected motifs. We can use this information to limit the number of computed alignments to improve the efficiency of the evolved algorithm and of future training.

Within top sub-optimal alignments, the degree of similarity (also called alignment density) in the motifs is greater than the degree in the *padding*, *i.e.*, segments of an alignment that do not represent biologically significant similarity. Given an alignment, it is computationally efficient to discriminate motifs from paddings. This can be done by measuring the growth of the similarity within different sampling intervals along the alignment. The objectives of the learning algorithm

[1]Computer Science Department, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180-3590. E-mail: breime@cs.rpi.edu

[2]Computer Science Department, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180-3590. E-mail: goldberg@cs.rpi.edu

are to determine the threshold separating the motifs from padding and the length of the sampling interval needed to detect the threshold in a given alignment.

An appropriate length for the sampling interval must be carefully selected. Small segments of the padding may have high density; similarly, small segments of the motif may possess low density. A large sampling interval may cover two or more motifs, which hinders the ability to identify the start and end of individual motifs. By sampling and plotting segment densities using different interval lengths, we can compute the minimum interval length that adequately discriminates motifs from padding. (see Fig. 1).

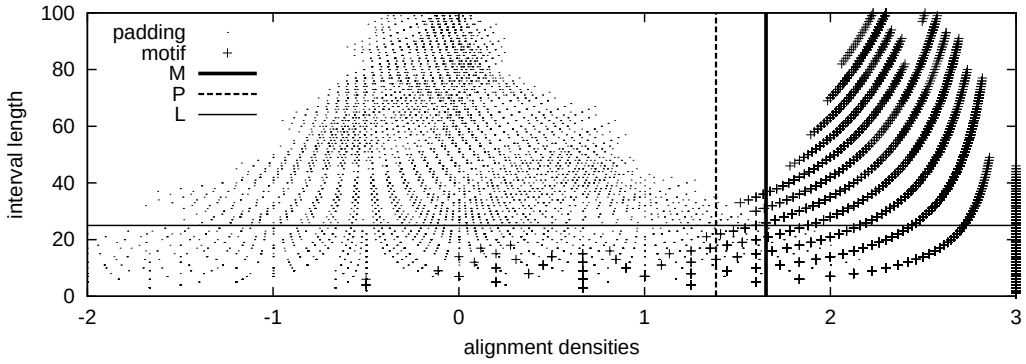

Figure 1: Alignment density distribution for training input. L is the minimum interval length. P and M are the density thresholds for identifying padding and motif segments, respectively.

L is defined as the minimum interval length such that all motif points above L are greater than density M and all padding points above L are less than density P. In the presence of noise, these thresholds can be relaxed so that a small percentage of outlying points can be ignored.

Using the interval length and density thresholds, we apply a post-processing algorithm to label segments of the sub-optimal alignments as either motif or padding. Given alignments with known motifs, [3] we learn thresholds for accurately labeling the training data and then we cross-test on unseen input. Our experiments show that the evolved algorithm is effective in discriminating motifs from padding and correcting all instances of the *mosaic effect* in the unseen samples. The running time of the algorithm is $O(n^2)$. We believe that our approach is a starting point for the design of more automatic and adaptive alignment algorithms.

References

[1] Altschul, S., Erickson, B. 1988. Significance levels for biological sequence comparison using nonlinear similarity functions. *Bulletin of Mathematical Biology* 50:77-92.

[2] Arslan, A., Eğecioğlu, Ö., Pevzner, P. 2001. A new approach to sequence comparison: normalized sequence alignment. In: *Proceeding of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB 2001),* Montreal: pp. 2-11.

[3] Barton, G. 1993. An efficient algorithm to locate all locally optimal alignments between two sequences allowing for gaps. *Computer Applications in the Biosciences* 9:729-734.

[4] Smith, T., Waterman, M. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* 147:195-197.

[3] We use the alignment of ATP-binding cassette sub-family B (MDR/TAP) between human and mouse.