

A Supervised Learning Approach for Detecting Significant Local Alignments

Eric A. Breimer¹
Mark K. Goldberg²

Keywords: normalized local alignment, supervised learning

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 effect* and *mosaic effect*. 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 propose a supervised learning approach that yields an $O(n^2)$ algorithm that effectively eliminates the *mosaic effect* while requiring no expert feedback to produce meaningful alignments. We use input sequences with known *motifs* to train the algorithm to align and extract these *motifs* by learning parameters for processing sub-optimal alignments. The term *motif* refers to an alignment that captures a biologically significant similarity as defined by an expert or oracle. The expectation is that the learned algorithm will be able to align and extract such *motifs* from unseen input sequences. The fundamental difference between our approach and others is that we provide an automatic framework for using existing *motifs* to tune the post-processing of sub-optimal alignments.

2 Learning Approach

Given training data, 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 *motif* is discovered (e.g., contained within a sub-optimal alignment). We consider the number of top scoring alignments that must be outputted in order to guarantee that a *motif* is not missed. Our experiments show that if a *motif* is discovered, it will typically be among the top k scoring alignments, where k is the number of expected *motifs*. Through training, we determine if the discovery percentage reaches an asymptote and use this information to limit the number of sub-optimal alignments computed in the future.

Within sub-optimal alignments, the degree of similarity (also called alignment density) in the *motifs* is greater than the degree in the *padding*. The term *padding* refer to segments of an align-

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

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

ment that do not represent biologically significant similarity. Density thresholds can be used to discriminate *motifs* from *padding*. However, selecting the proper sampling interval to determine density is problematic. Very small segments of the *padding* often possess high density. Similarly, small segments of the *motif* may possess low density. A large sampling interval is also problematic because it may encompass two or more *motifs*. Thus, pinpointing the start and end of individual *motifs* is difficult. By sampling and plotting segment densities using different interval lengths, we can simultaneously detect the minimum interval length and density thresholds that adequately discriminates *motifs* from *padding*s (see figure 1).

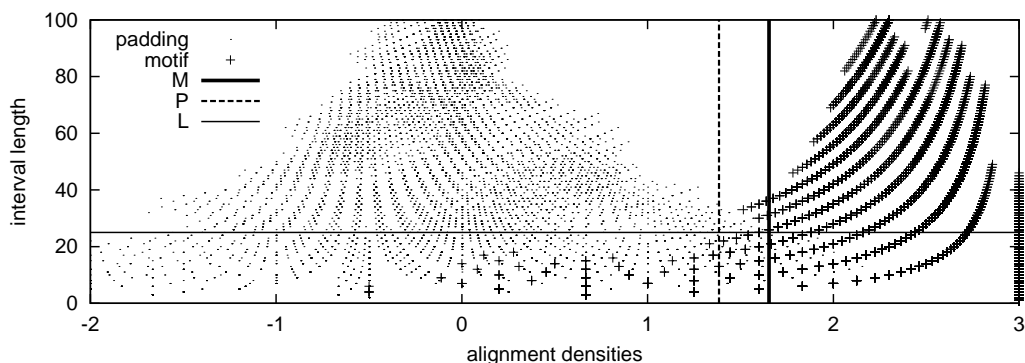


Figure 1: Alignment density distribution for training input. L is the minimum interval length. P and M are the density thresholds for identifying padding and motif segments, respectively.

Using these density thresholds and the interval length, we can apply a variety of different post-processing algorithms to extract potential *motifs* and discard *padding*s. After training on a set of *motifs*³ and cross-testing on unseen segments, the post-processing was able to detect and correct all instances of the *mosaic effect*. 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., Egecioglu, 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.

³The alignment of ATP-binding cassette sub-family B (MDR/TAP) between human and mouse