Learning Significant Alignments: An Alternative to Normalized Local Alignment

Eric Breimer and Mark Goldberg*

Computer Science Department, Rensselaer Polytechnic Institute, 110 Eight Street, Troy NY 12180, USA, breime@cs.rpi.edu

Abstract. We describe a supervised learning approach to resolve difficulties in finding biologically significant local alignments. It was noticed that the $O(n^2)$ algorithm by Smith-Waterman, the prevalent tool for computing local sequence alignment, often outputs long, meaningless alignments while ignoring shorter, biologically significant ones. Arslan *et. al.* proposed an $O(n^2 \log n)$ algorithm which outputs a *normalized local alignment* that maximizes the degree of similarity rather than the total similarity score. Given a properly selected normalization parameter, the algorithm can discover significant alignments that would be missed by the Smith-Waterman algorithm. Unfortunately, determining a proper normalization parameter requires repeated executions with different parameter values and expert feedback to determine the usefulness of the alignments. We propose a learning approach that uses existing biologically significant alignments to learn parameters for intelligently processing sub-optimal Smith-Waterman alignments. Our algorithm runs in $O(n^2)$ time and can discover biologically significant alignments without requiring expert feedback to produce meaningful results.

1 Background

Local sequence alignment is an essential technique for identifying similarity between biological sequences [6,8,10]. The Smith-Waterman algorithm [15] is considered the standard tool for computing local sequence alignment. However, it was noticed (see, [2,4]) that the algorithm has two essential flaws. It often combines two or more segments of high similarity and aligns internal segments that are not related; the *mosaic effect*, see Fig. 1 (ii). Occasionally, it finds long alignments with a high score and misses shorter ones with a higher degree of similarity; the *shadow effect*, see Fig. 1 (i).

A number of attempts have been made to correct the flaws of the Smith-Waterman algorithm, including unsuccessful approaches that were abandoned [11,14], approaches that are computationally expensive [12,17], and approaches that require sensitive heuristics [18,3]. Further attempts were made to consider length in the computation of alignments [5,13,16]. The most recent and successful

* Supported by a grant from Rensselaer Polytechnic Institute.

M.-S. Hacid et al. (Eds.): ISMIS 2002, LNAI 2366, pp. 37–45, 2002.

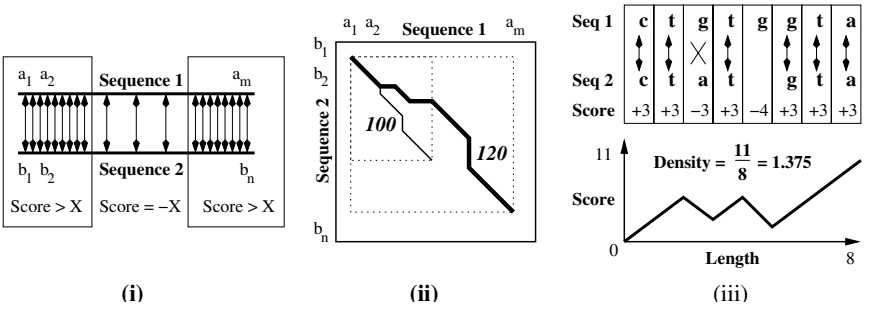

Fig. 1. (i) Mosaic effect: Two high scoring segments are joined into a single alignment that is not biologically significant. (ii) Shadow effect: Sub-optimal alignments with high degree of similarity are ignored in favor of longer alignments. (iii) Alignment Density: Alignment score divided by the length of the alignment.

approach was proposed by Arslan et. al. [4]. This approach seeks to maximize the *normalized score* S_N defined by

$$S_N = \frac{S}{l+N},$$

where S is the conventional alignment score, l is the sum of the lengths of the two aligned segments, and N is the *normalization parameter* used to control the degree of normalization. Arslan et. al. [4] designed an $O(n^2 \log n)$ algorithm for computing normalized local alignment (*nla*) that uses the fractional programming technique developed in [9]. Unfortunately, the algorithm is very sensitive to the value of N. If N is too small, the algorithm is indistinguishable from an exact matching algorithm, whereas if N is too large, the algorithm suffers from the same negative side effects of the Smith-Waterman algorithm. The useful range for N is input-dependent and the relationship between the input and the appropriate value of N is generally unclear [4]. Thus, applying the algorithm requires *guessing* a preliminary value for N, and obtaining meaningful alignments may require repeated execution of the algorithm with a varying value of N. Repeatedly executing the *nla*-algorithm and manually evaluating the results is tedious and time consuming. For large scale applications, a more automated and efficient process is still needed.

For the remainder of this paper, an alignment that captures biologically significant similarity is called a *motif*. In practice, motifs are identified by expert biologists. Coincidental alignments without biological significance are called *padding*. The training data consists of pairs of sequence segments where all motifs are known. Thus, given any alignment, we can identify the specific segments that are motifs and those that are padding. Degree of similarity, also called *density*, refers to the alignment's score divided by the alignment's length (see Fig. 1 (iii)). Alignment density, although similar to normalized score, does not include a normalization parameter and defines length to be the number of aligned symbol pairs.