Sequence Alignment Non-Gap Penalties and Gains

**Abstract**

Sequence alignment is an important component of genome comparison or assembly. When the quality of alignment is important, a global sequence alignment is often used, ensuring that the canonical scoring function is globally maximized. One of the most commonly used algorithms is the Needleman-Wunsch algorithm, which uses a dynamic programming approach to reduce the time complexity to O(N²) [1]. The scoring consists of three components: matches, mismatches, and gaps. The most commonly used functions for each of the three are linear for matches, linear for mismatches, and affine for gaps. There have been several papers discussing the use of more complex gap penalty functions such as logarithmic or affine logarithmic [2]. These have been motivated by empirical results, although turned out to be inappropriate for alignment [3]. Despite this focus on gap functions, few papers acknowledge the use of non-linear match gains and mismatch penalties (here referred to as non-linear match and mismatch functions, NLM). This is due in part to the large complexity of testing different coefficients for these functions, which is a cubic increase in simulation time complexity compared to just altering the gap penalty function. Still, it is possible to perform the Needleman-Wunsch algorithm with arbitrary scoring functions without an increase in time complexity, as is implied by previous research [4,5].

In this paper, I hypothesized that NLM functions would have a significant positive effect on scoring outcome. A small subset of all possible NLM functions were explored with simple coefficients. The results indicate that an opening penalty for sequential mismatches is beneficial to score by discouraging mismatches with no effect on gaps, but complex match functions decrease the accuracy by encouraging more gaps to form.