Section 3: Sequence Alignment/Machine Learning

The first two readings introduce early methods for aligning DNA sequences, or identifying regions of similarity that may be a consequence of functional or evolutionary relationships. These are the efforts of independent research groups that devise methods for maximizing matching nucleotides, while minimizing gaps and gap length. The last reading gives a brief introduction to machine learning and how it has been applied to

Smith and Waterman describe an algorithm, now known as the Smith-Waterman algorithm, for determining similar regions between text strings. This can be applicable to both nucleic acids and proteins. It is essentially an extension of the Needleman-Wunsch algorithm, using a very similar substitution matrix scheme. However, the main difference is that no matrix cells are allowed to be negative, and gap-scoring is used. This algorithm operates on the order of O(n3), which is far too slow for practical large-scale application. However, it set the foundation for future works that developed more computationally efficient alternatives.

One such alternative is described in Altschul et al. Their titular algorithm, “basic local alignment search tool (BLAST),” is primarily used to compare a query sequence with a library of sequences and identify matches. It uses a heuristic algorithm that skips the full alignment procedure, such as the one in the Smith-Waterman algorithm. This sacrifices the guarantee of optimality for a vast increase in speed. It first locates short matches in a process known as seeding. Then, the algorithm makes local alignments and extends these in both directions while tracking the score.

Yip et al. provide a review of machine learning and its uses in genomics. I’m presenting on this today, and have a summary included in my powerpoint.

* Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. Journal of Molecular Biology, 215(3):403-10. PMID: 2231712.
* T.F. Smith and M.S. Waterman. (1981) Identification of common molecular subsequences. Journal of Molecular Biology,147(1): 195-7. PMID: 7265238.
* Yip, KY, Cheng, C, Gerstein, M (2013). Machine learning and genome annotation: a match meant to be?. Genome Biol., 14, 5:205.