



# TUTORIAL: A LINEAR PROGRAMMING APPROACH TO MULTIPLE SEQUENCE ALIGNMENT

Fern Hunt

Mathematical and Computational Sciences Division 8910  
National Institute of Standards and Technology  
hunt@email.nist.gov

The advent of large scale sequencing of DNA has triggered a massive accumulation of DNA sequence data and data about proteins- the products of the genes encoded by DNA. Methods for applying this information to fields such as law enforcement, biotechnology, medicine or to gain crucial understanding of the biological significance and functionality of genes and proteins depend on a technique known as sequence alignment. An alignment is an array that displays the degree of relatedness of a set of sequences (either DNA or protein) through matching and equivalent substitutions and the insertion of gap characters that are used to maximize the degree of similarity. The quality of the alignment is evaluated by assigning a total “score” or “cost” and the goal is to attain the “best” alignment. The most widely used algorithms for constructing alignments are based on dynamic programming. Unfortunately a direct implementation of these methods is not suitable for aligning a large number of long sequences. In our talk we will discuss an alternative formulation of the alignment problem based on Markov decision theory. Here one seeks to minimize an average or expected cost subject to data-derived constraints. In this setting the problem is equivalent to a linear programming problem which can be solved efficiently.