# Gene Prediction with a Conditional Random Field Maximizing Expected Accuracy

Prof. Mario Stanke
(Universität Greifswald)

February 1, 2013

Many bioinformatics tasks are solved by defining a probability for every element of a discrete solution space and by then devising an algorithm that searches - exactly or approximately - for an element with highest probability. Gene prediction is a common task in biological sequence analysis, but not yet satisfactorily solved for higher organisms. The elements of the solution space, here gene structures, can be defined as a classification of each base of the input DNA sequence into one of a few categories. We construct a probability distribution over all possible gene structures through a linear-chain conditional random field and find a most likely gene structure with the Viterbi algorithm. However, we recently found that on the same model an algorithm that maximizes another criterion - the expected accuracy - gives better performance than the Viterbi algorithm. An outlook to a more complex non-linear model for the simultaneous prediction of genes in many related species will also be given.