# Biological sequence annotation with hidden Markov models

Michal Nánási
*UK, Bratislava, Fakulta matematiky, fyziky a informatiky*

Hidden Markov models (HMM) are an important tool for modeling biological sequences and their annotations. By annotating sequences we mean assignment labels for each symbol according to its function. For example, in gene finding we want to distinguish between regions of DNA that encodes proteins from non-coding sequence. Hidden Markov model defines a probability distribution over all annotations of sequence $X$.

Commonly used algorithm for HMM decoding is the Viterbi algorithm. Viterbi algorithm finds the most probable annotation for subset of HMMs. In general, the sequence annotation is NP-hard and Viterbi algorithm is used as heuristic algorithm. Recently it has been shown that other decoding methods have better result than Viterbi in specific applications. We propose new decoding method that allows uncertainty in region boundaries.

Our method is based on a framework of maximum expected boundary accuracy decoding. Boundary is the change of a label in annotation sequence at particular place. We define our objective function $R$ in terminology of gain functions . In particular, let $A$ be an annotation and $A'$ be the correct annotation. Every boundary at position $i$ in $A$ will get reward $+1$ if in $A'$ is the same boundary at position $j$ and $|i - j| < W$. Otherwise, that boundary will get reward $-\gamma$.

Our goal is to find the annotation maximizing the expected reward. Expected reward of annotation $A$ of sequence $X$ is $\sum_{A'} R(A, A') \cdot P(A'|X)$. We call our method the Highest Expected Reward Decoding (HERD). The time complexity of HERD algorithm is $O(nWC|E| + nC^2W^2)$ where $n$ is the length of sequence, $C$ is number of different labels and $W$ is parameter from gain function.

We evaluate this approach on the problem of detecting viral recombination in HIV genome and compare it with existing tool called jumping HMM which uses the Viterbi algorithm. HERD has slightly better performance in terms of correctly labeled symbols, and also is significantly better with respect to feature specificity and sensitivity. The feature is block of label of same color and it is correctly predicted if its boundaries are misplaced by at most 10 symbols.

This paper is large subset of author's master thesis.