HERD: the Highest Expected Reward Decoding for HMMs with Application to Recombination Detection

Michal Nánási, Tomáš Vinař, Broňa Brejová

Faculty of Mathematics, Physics and Informatics, Comenius University, Mlynská Dolina, 842 48 Bratislava, Slovakia

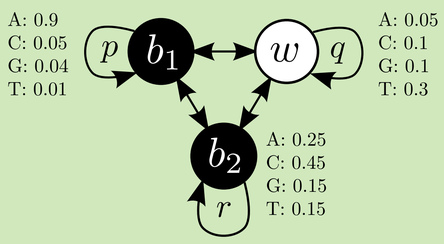
mic@compbio.fmph.uniba.sk
http://www.compbio.fmph.uniba.sk/herd

Introduction

Hidden Markov models (HMM) are an important tool for modeling biological sequences and their annotations. By annotating sequences, we mean assigning a label to each symbol according to its meaning or function. In our paper [Nanasi et al., 2010] we introduced a new decoding method for finding annotations and its application to detecting viral recombination in the HIV genome.

Hidden Markov models

- Hidden Markov model (HMM) defines a probability distribution over annotations of sequence X.



An example of a simple HMM with three states (b_1, b_2, w) and two colors (black and white).

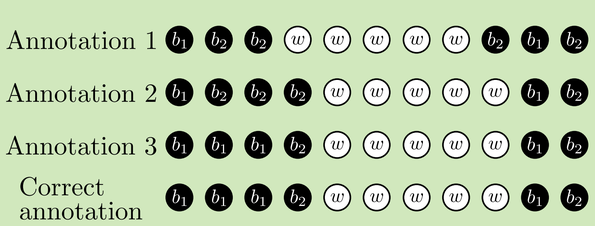
- A coloring function λ assigns a color to each state corresponding to its meaning. The probability of annotation Λ is the sum of state paths π with state colors $\lambda(\pi) = \Lambda$
- Example: state paths wb_1w and wb_2w have same annotation white, black, white

Decoding strategies

We express decoding strategies in terms of gain functions. Gain function $R(\Lambda, \Lambda')$ is a similarity (reward) between the annotation Λ and the correct annotation Λ'. We seek an annotation with the highest expected reward [Hamada et al., 2009]:

$$E_{\Lambda'|X} = \sum_{\Lambda'} R(\Lambda, \Lambda') \cdot P(\Lambda'|X) \quad (1)$$

- Viterbi algorithm finds the most probable state path that generates given sequence X. It is equivalent to gain function R_V that assigns score $+1$ if the state paths are same and 0 otherwise
- Posterior decoding assigns to each symbol of X the most probable state or color. The corresponding gain function R_P assigns $+1$ to every correctly annotated symbol
- Finding the most probable annotation corresponds to gain function R_A that assigns $+1$ if the annotation is correct and 0 otherwise. Finding such annotation is NP-hard [Brejova et al., 2007]



Example of three annotations with different rewards $R_*(\Lambda, \Lambda')$.

Highest Expected Reward Decoding (HERD)

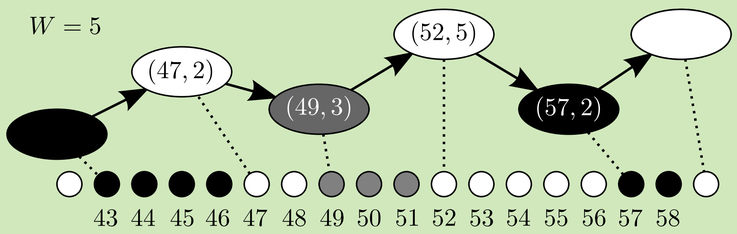
We propose a new gain function which can be optimized efficiently. Our gain function is appropriate when is hard to find exact boundaries. For each boundary between two colors in Λ we check whether Λ' has the same boundary within distance less then W. If so, the boundary gets reward $+1$, otherwise it gets reward $-\gamma$.

Example of an annotation with reward $2 - 2\gamma$

Algorithm

We have designed an effective algorithm for optimizing our objective.

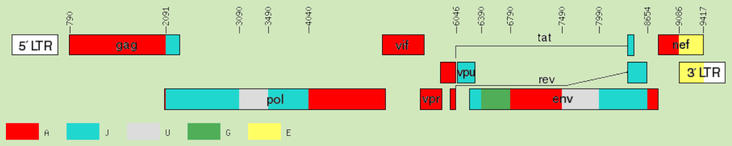
- First phase: compute posterior probabilities for all boundaries
- Second phase: construct directed acyclic graph in which every path from start vertex to end vertex represents one possible annotation, the weight of the path is the expected reward
- Third phase: find the path with the highest weight
- Running time: $O(n|E|W)$ where n is the length of the sequence, $|E|$ is the number of transitions in the HMM, W is parameter of the reward function



Part of one path in the graph constructed in the algorithm (top) and its corresponding annotation (bottom).

Detection of Viral Recombination

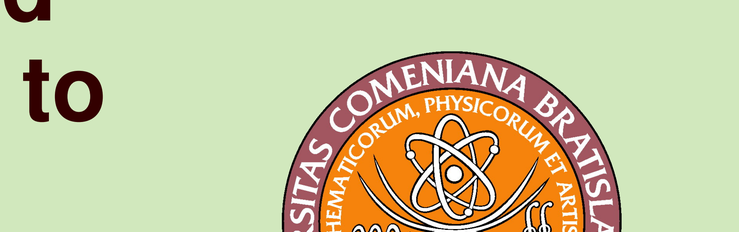
We considered the HIV (Human Immunodeficiency Virus) genome [Robertson et al., 2000]. The HIV virus mutates often and currently is divided into several subtypes. Some HIV genomes are a mosaic recombination of viruses from two or more different subtypes [Robertson et al., 2000]. Our goal is to distinguish parts of the viral sequences that have origin in different viral subtypes.



Source: http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/breakpoints.html Circulating recombinant form CRF11 contains genetic material from 4 subtypes of HIV-1 M group.

Jumping HMM

- Jumping HMM (jpHMM) [Schultz et al., 2009] models recombination in DNA sequences
- jpHMM consists of several profiles, each profile for one virus family
- Searching for exact recombination points is difficult since sequences from different families are similar. HERD overcomes this difficulty



Simplified structure of jpHMM. Each color corresponds to one subtype, each column to one symbol of DNA sequence.

Experiments

We compare our HERD algorithm with Viterbi algorithm previously used to detect viral recombination in HIV genome [Schultz et al., 2009]. Both algorithm use the same HMM.

Algorithm	% labels correct	Feature sp.	Feature sn.	Avg. dist.
Sequences with inter-subtype recombination				
HERD	95.7%	63.1%	58.9%	2.4
Viterbi	95.4%	53.4%	47.9%	1.8
Sequences with intra-subtype recombination				
HERD	91.6%	46.5%	41.9%	2.7
Viterbi	88.0%	32.8%	26.1%	2.7

Experimental results on 350 artificial HIV recombinants with average length 1200bp and recombination every 300bp. HERD was used with parameter $W = 10, \gamma = 1$. A feature is a block of labels of the same color and it is correctly predicted if its boundaries are misplaced by at most 10bp.

Acknowledgement: This research is funded by European Community FP7 grants IRG-224885 and IRG-231025 and VEGA grant 1/0210/10.

References

[Brejova et al., 2007] Brejova, B., Brown, D. G., and Vinar, T. (2007). The most probable annotation problem in HMMs and its application to bioinformatics. *Journal of Computer and System Sciences*, 73(7):1060–1077.

[Hamada et al., 2009] Hamada, M., Kiryu, H., Sato, K., Mituyama, T., and Asai, K. (2009). Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, 25(4):465–473.

[Nanasi et al., 2010] Nanasi, M., Vinar, T., and Brejova, B. (2010). The Highest Expected Reward Decoding for HMMs with Application to Recombination Detection. In *Combinatorial Pattern Matching, (CPM 2010)*, volume 6129 of *Lecture Notes in Computer Science*, pages 164–176. Springer.

[Robertson et al., 2000] Robertson, D. L. et al. (2000). HIV-1 nomenclature proposal. *Science*, 288(5463):55–56.

[Schultz et al., 2009] Schultz, A.-K., Zhang, M., Bulla, I., Leitner, T., Korber, B., Morgenstern, B., and Stanke, M. (2009). jpHMM: improving the reliability of recombination prediction in HIV-1. *Nucleic Acids Research*, 37(Web Server issue):W647–651.