Eukaryotic Gene Prediction Maximizing Posterior Accuracy

Lizzy Gerischer and Mario Stanke
Institute for Mathematics and Computer Science
University of Greifswald
{lizzy.gerischer , mario.stanke}@uni-greifswald.de

Problem

Probabilistic models such as HMMs are often used to model unknown variables X, e.g. a gene structure or an alignment that need to be predicted or constructed. A popular decoding approach is to use the MAP (Maximum *a posteriori*) estimator $\hat{x}_{\text{MAP}}$ that maximizes

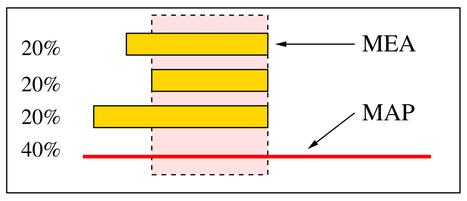
$$E\left[\mathbb{I}_{\{x=X\}}\right] \text{ with respect to } x$$

and can, in the case of HMMs, be computed with the Viterbi algorithm. For some applications it might be more appropriate to *maximize* another target: the *expected accuracy* (MEA). $\hat{x}_{\text{MEA}}$ estimates the *similarity* of an arbitrary x to the actual but unknown value of X

$$E\left[a(x, X)\right] \text{ with respect to } x.$$

Do *et al* have applied this principle in the multiple sequence aligner ProbCons [2]. We here pursue such a maximum expected accuracy approach in the gene prediction tool AUGUSTUS [1].
In gene prediction the unknown variable is the correct gene structure G of a DNA or RNA sequence. MAP tries to find a gene structure that maximizes the probability of being exactly correct. However,

- many uncertainties lead to low probabilities of even the most likely gene structures,
- this approach does not take similarity considerations into account.



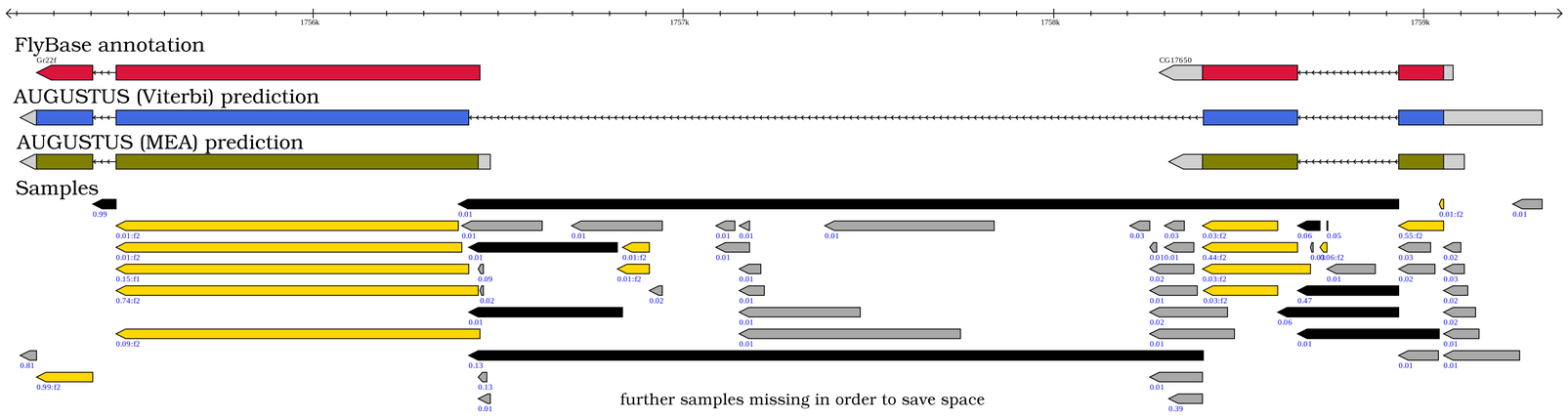
A toy example illustrates the intuition: The MAP estimator chooses the option with probability of 40% (no gene) although the occurrence of a gene is more likely (60%). The alternatives derive from the uncertainty of the start codon position.

Results

		Drosophila		human	
		Viterbi	MEA	Viterbi	MEA
gene	sn	42.61%	44.07%	13.21%	14.62%
	sp	52.15%	60.43%	6.29%	10.26%
transcript	sn	32.41%	33.74%	8.19%	9.06%
	sp	52.15%	60.43%	6.29%	10.26%
exon	sn	69.87%	71.17%	69.57%	68.13%
	sp	75.41%	85.64%	47.68%	63.84%
base	sn	92.36%	85.43%	80.72%	77.07%
	sp	92.08%	96.65%	56.80%	70.14%

Evaluations with `eval` on complete *D. melanogaster* chr. 2L and human chr. 21
sn = sensitivity
sp = specificity

We observe major improvements of the specificity in both species.



4.5 kb region of chromosome 2L of *Drosophila melanogaster* created with GBrowse. The red track is an annotation taken from `FlyBase`, the blue track a prediction by AUGUSTUS with the MAP approach and the green track a prediction with the MEA method.

Method

We introduce an alternative to the Viterbi algorithm with the gene finder AUGUSTUS [1] that **M**aximizes the **E**xpected **A**ccuracy (MEA).
Measuring the accuracy of a gene structure g requires the correct gene structure G. Since G is not given we calculate the *expected* accuracy instead

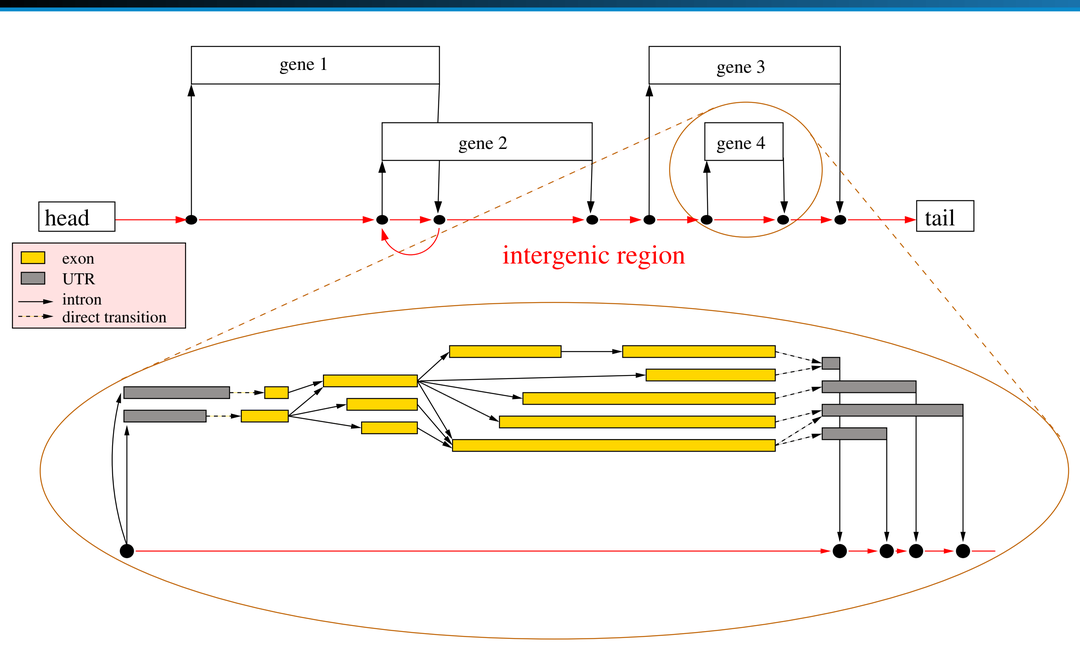
$$E\left[a(g,G)\right] = \sum_{g'} P(G=g'|s)\cdot a(g,g') \approx \sum_{g'} \frac{1}{m} \underbrace{\sum_{i=1}^{m} \mathbb{I}_{\{g_i=g'\}}}_{\text{number of sampled } g'} \cdot a(g,g') = \frac{1}{m} \sum_{\substack{g' \\ g' \text{ sampled}}} a(g,g'),$$

where g' goes over all possible gene structures, s is the DNA or RNA sequence, $P(g'|s)$ is estimated by its sample frequency (sampling algorithm), m the number of sample iterations and a is an accuracy criterion defined as

$$a(g,g') = \sum_{\text{exon } e_g \text{ in } g} \mathbb{I}_{\{g' \text{ contains } e_g\}} + \sum_{\text{intron } i_g \text{ in } g} \mathbb{I}_{\{g' \text{ contains } i_g\}}$$

We maximize the expected accuracy by transferring it to a shortest path problem in a graph.

Graph Representation



A path $p = (v_1, ..., v_n)$ through the *MEA exon graph* $M = (V, E)$ with nodes $v_i \in V$ and edges $(v_i, v_j) \in E$ is a possible gene structure g. The optimal path maximizes the sum of node and edge scores s which is equivalent to maximizing the posterior accuracy. A modified Bellman-Ford algorithm was implemented.

$$\begin{aligned} \textit{weight}(p) &= \sum_{i=1}^{n-1} (s(v_i) + s(v_i, v_{i+1})) = \sum_{i=1}^{n-1} \left(\frac{1}{m} \sum_{j=1}^{m} \mathbb{I}_{\{g_j \text{ contains } v_i\}} + \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}_{\{g_j \text{ contains } (v_i, v_{i+1})\}} \right) \\ &= \frac{1}{m} \sum_{j=1}^{m} a(g, g_j) \end{aligned}$$

Further Studies

- To further improve the accuracy values a custom training of the HMM parameters might be useful.
- When predicting genes in several related species, we use a MEA exon graph for each species (work in progress).

References

[1] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack and B. Morgenstern. AUGUSTUS: *ab initio* prediction of alternative transcripts In *Nucleic Acids Research, 2006*

[2] C.B. Do, M.S.P. Mahabhashyam, M. Brudno and S. Batzoglou. ProbCons: Probabilistic consistency-based multiple sequence alignment. In *Genome Research, 2005*

Source Code

The source code is available at

`http://bioinf.uni-greifswald.de/augustus/`