Computationally Efficient Estimation of the Error Rates of Hidden Markov Model Results

Lee A. Newberg
lee.newberg÷wadsworth÷org

Center for Bioinformatics
Wadsworth Center
Albany, NY 12201 USA

Dept. of Computer Science
Rensselaer Polytechnic Inst.
Troy, NY 12180 USA

The Players

sequence database scan
false positives rate
false negative rate
hidden Markov model (HMM)
hidden Boltzmann model (HBM)
Viterbi algorithm
forward algorithm
hypothesis test
importance sampling

I am searching a huge database; I need *p*-values to be under 10^{-12}. What are the *p*-values of the sites your HMM/HBM-based software finds?

Umm … I can do that for you. I shall simulate some random sequences and let you know what fraction of them give a score as extreme as those sites. I should need … oh, I would need a few trillion sequences for *p*-values of 10^{-12}, … *uh oh!*

Importance Sampling Superhero to the rescue.

I can do it with a few hundred simulated sequences!

What? How??

The Set-Up

The statistical significance for some score s_0 is defined to be

$$p(s_0) = \sum_D \Pr(D|B)\Theta(s(D) \geq s_0)$$

where the sum is over sequences *D*, Pr(*D*|*B*) is the probability of the sequence under some background model *B*, the score *s*(*D*) is from the software, and Θ is a function that is 1 if its argument is true or 0 if it is not.

The Hook

But if *T* is some other model we can also write

$$p(s_0) = \sum_D \Pr(D|T) f(D)$$

where

$$f(D) = \frac{\Pr(D|B)\Theta(s(D) \geq s_0)}{\Pr(D|T)}$$

We can sample sequences according to the model *T* and average their corresponding *f*(*D*) values. This is called *importance sampling*. If *T* is well chosen, only a few hundred sequences are needed for a good estimate.

The Tale

What model *T* should I use?

Toward calculating *Pr(D|T)*, we use a HMM/HBM forward algorithm with all the HMM/HBM-software transition and emission probabilities raised to some power 1/*T*. Specifically, we define the model for parameter *T* as

$$\Pr(D|T) \propto \Pr(D|B)\mathrm{HMM}(D|p^{1/T})$$

We compute the normalization factor

$$Z(T) = \sum_D \Pr(D|B)\mathrm{HMM}(D|p^{1/T})$$

as we would compute HMM($D|p^{1/T}$), but using the *mean* emission probability of an emitter *E*

$$\sum_d \Pr(d|B) p_E(d)^{1/T}$$

in lieu of any specific emission probability $p_E(d)^{1/T}$ for letter *d*.

To sample, we perform a stochastic backtrace through the *Z*(*T*) calculation. We sample the path as usual, and at each encounter with each emitter *E* we sample a letter *d*, with probability

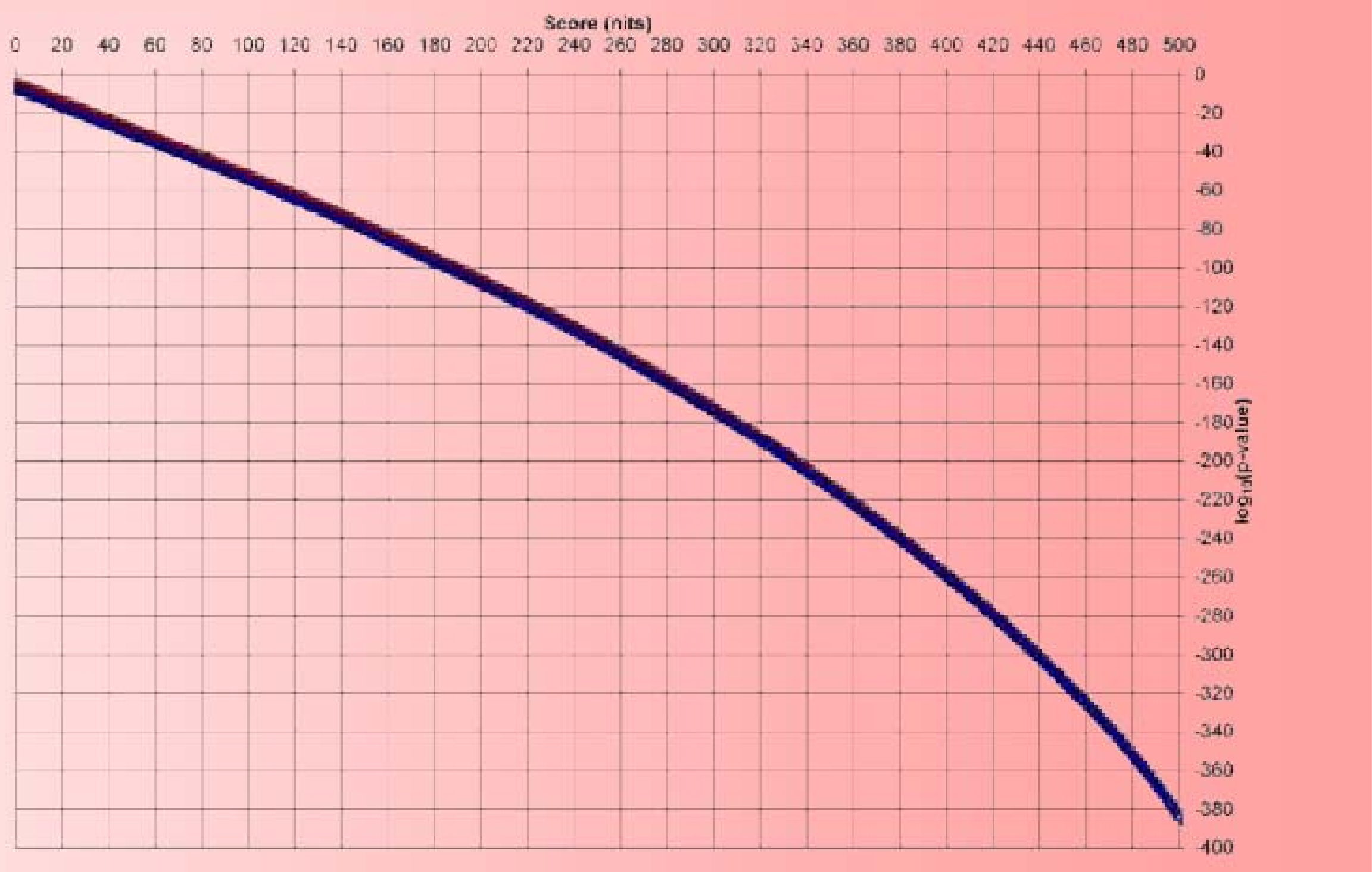
$$\frac{\Pr(d|B) p_E(d)^{1/T}}{\sum_{d'} \Pr(d'|B) p_E(d')^{1/T}}$$

For each sequence thus sampled, we compute

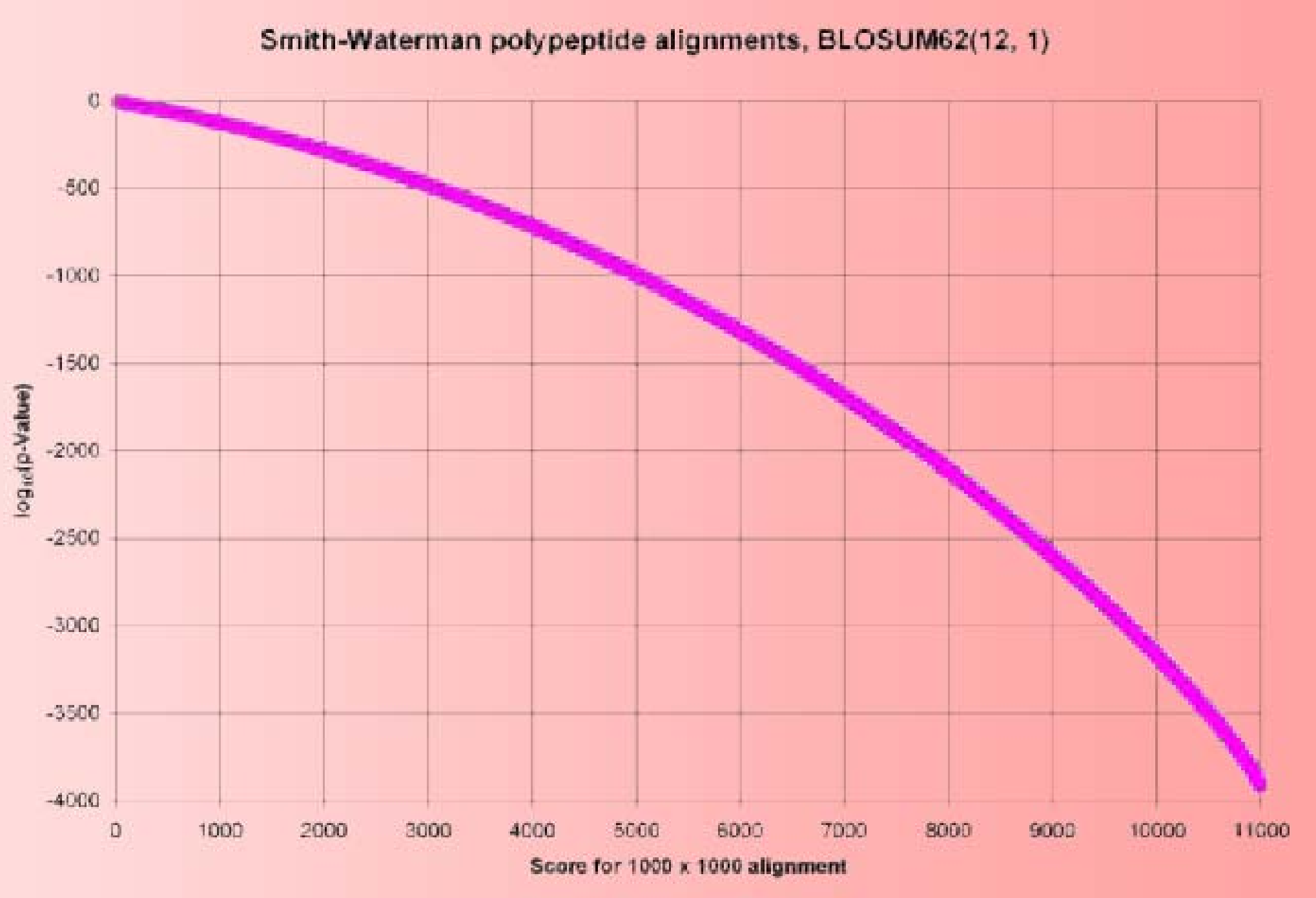
$$f(D) = \frac{Z(T)\Theta(s(D) \geq s_0)}{\mathrm{HMM}(D|p^{1/T})}$$

The Wire

Statistical significance of HMMER profile-HMM scores for a random module model of length M=100, scanning a sequence of length L=200.



Statistical significance of Smith-Waterman alignments using BLOSUM62 with 12 and 1 for insertion start and extension penalties.
You can compute *p*-values to 10^{-4000}??



The Sting

- It works for maximum (Viterbi) scores, forward scores, ….
- Can do statistical sensitivities too.
- It works better for extreme scores.
- There are heuristics for choosing *T*.
- **http://bayesweb.wadsworth.org/alignmentSignificanceV1**

The Thanks

Thanks to Sean Eddy for access to his HMMER 3.0 source code and to him and Chip Lawrence for helpful advice. This research was supported by US NIH NHGRI K25-HG003291 and US DOE BER DE-FG02-09ER64756.