Global Measures of Uncertainty

Long Overdue in Computational Molecular Biology

Lee A. Newberg[1,2] Bobbie-Jo M. Webb-Robertson[3]
Lee Ann McCue[3] Charles E. Lawrence[4]

[1]Wadsworth Center, New York State Department of Health

[2]Department of Computer Science, Rensselaer Polytechnic Institute

[3]Computational Biology and Bioinformatics, Pacific Northwest National Laboratory

[4]Division of Applied Mathematics, Brown University

ISMB / ECCB, July 2, 2009

Hypothesis Testing vs. Credibility Limits

- **Question:** Smith-Waterman alignment with $E = 10^{-40}$. It's a good alignment right?
- **Answer:** No, there is a reasonable chance that sizable alignment blocks are wrong.

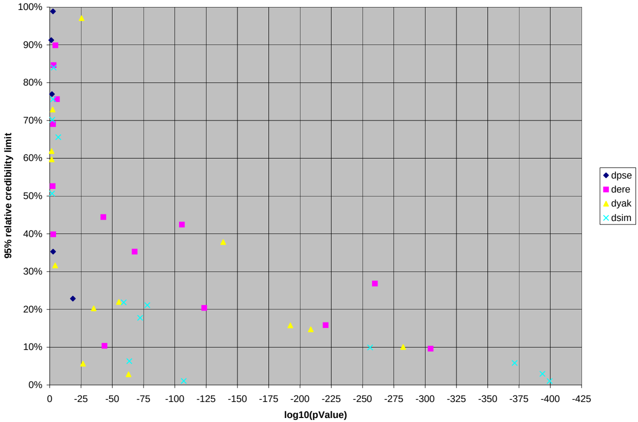
E-Value and p-Value Are for Hypothesis Testing

E, p are small when random data is unlikely to do as well.

Credibility Limits (a.k.a. Bayesian Confidence Limits)

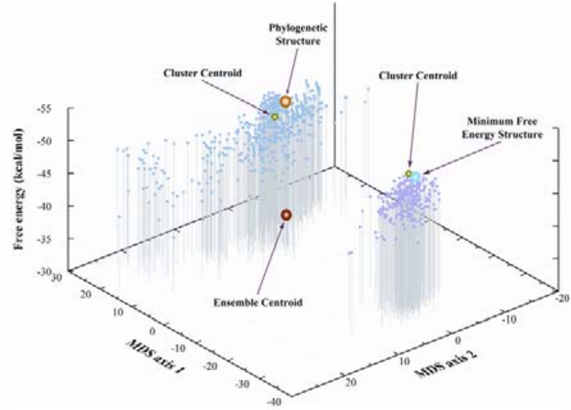
How many differences must be permitted to capture 95% of the posterior probability?
95% credibility limit is tight if most good solutions are similar.

Smith-Waterman Alignment



- Individual cases are bad even at superb *p*-values.
- *E*-values, *p*-values are a poor proxy for credibility.

5S RNA Secondary Structure



- No single structure represents the ensemble well.
- Minimum Free Energy isn’t the best representative.

Discrete High-Dimensional Inference

Much of computational biology is discrete high-D inference:

- Sequence alignment . . . . which residues are matched?
- RNA secondary structure . . . . which bases pair?
- Network inference . . . . which edges included?
- Nucleosome occupancy . . . . at which sequence positions?

Solution spaces are immense yet we often choose a point estimate solution.

Today's goal: Compute a global measure of representativeness of a point estimate.

Uncertainty of individual features (*e.g.*, bases pairings) — valuable and important but not our goal.

Algorithms for Discrete High-Dimensional Inference

Many problems are tackled with dynamic programming:

Hidden Markov Models

- Sequence alignment: HMMER
- Protein folding: HMMSTR / ROSETTA

Partition Function Computations

- RNA secondary structure: Sfold

Viterbi / Maximum Score / Minimum Energy

- Seq. Alignment: Smith-Waterman, Needleman-Wunch
- RNA secondary structure: Mfold

Collectively, *Hidden Boltzmann Models*

Computing / Estimating Credibility

1. If Viterbi: Set solution space probability distribution.
2. Distribution of differences from point estimate via either:

Sampling via HBM Stochastic Backtrace

- Draw 1000 samples
- Compare to point estimate

Fourier Computation

- Exaclty computes probability for each count of differences
- Runtime slowdown $=$ number of differences possible. (With parallel processors, same as unmodified algorithms.)
- Memory-usage: same as unmodified algorithm

3. d is “x% credibility limit” if x% of ensemble is distance $\leq d$.

Distance Distribution for Sequence Alignment

Set Solution Space Probability Distribution

For sequences x and y, set probability of an alignment A with score $s(x, y, A)$ to be:

$$\Pr[A|x, y] \propto \exp\left(\lambda s(x, y, A)\right)$$

for some parameter $\lambda > 0$, *e.g.*, $\lambda = \ln(10)/5$.

Modify algorithm: Add scores → multiply exponentiated scores, "$\max s_i$" → "$\sum \exp(\lambda s_i)$"

Choose an Approach

For a 3000 nt × 3000 nt alignment, Fourier is plenty fast. We get the full, exact distribution of the number of pairing differences.

Fourier Computation

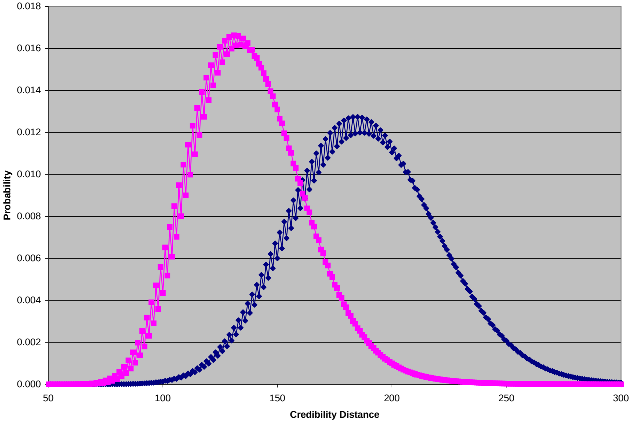
Computing the distribution for differences from a point estimate:

Algorithm Outline

- For each $\omega \in \left\{\cos\left(\frac{2\pi k}{n}\right) + i \sin\left(\frac{2\pi k}{n}\right), k = 0, \ldots, n-1\right\}$ (nth roots of unity) do
  - Run a modified HBM algorithm: If an HBM transition or emission implies d differences then multiply by ω^d.
- Fourier transform the n results.

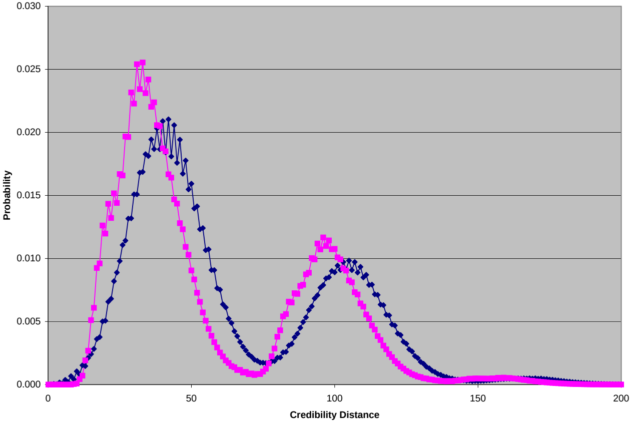
Note: Each ω can be run on a separate processor.

Number of Pairing Differences: Centroid vs. Viterbi



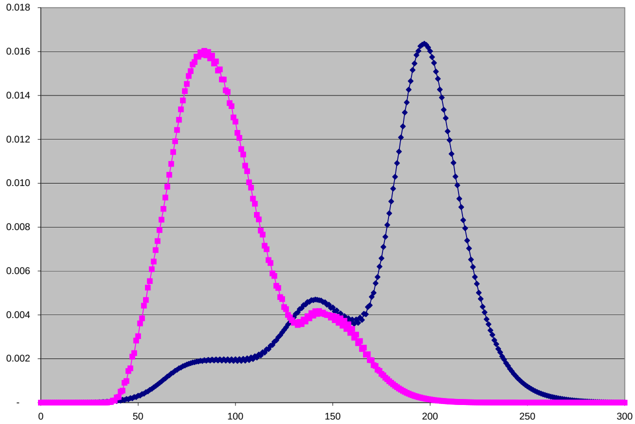
Example #1: Human (1769 nt) × Mouse (1575 nt).
Viterbi=1123 bp, Centroid=1099 bp.

Number of Pairing Differences: Bimodal



Example #2: Human (1691 nt) × Mouse (2219 nt).
Viterbi=214 bp, Centroid=205 bp.

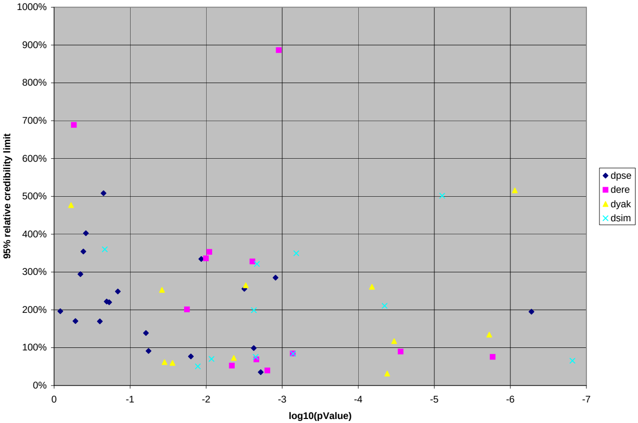
Number of Pairing Differences: Rich Structure



Example #3: Human (1677 nt) × Mouse (1666 nt).
Viterbi=450 bp, Centroid=438 bp.

95% Relative Credibility vs. Weak *p*-Value

$$\textit{relative}\ \text{credibility limit} = \frac{\text{credibility limit}}{\text{\# pairings in Viterbi alignment}}$$



Take-Home Points

- For discrete high-dimensional inferences, point estimates should be regarded with suspicion.
- *E*-values, *p*-values don't indicate credibility well.
- Credibility distributions can be calculated / estimated with reasonable efficiency.
- The 95% credibility limit is a global measure of representativeness of a point estimate.
- Centroids almost always beat Viterbi by this measure.

References

Sampling: `http://dx.doi.org/10.1371/journal.pcbi.1000077`
Fourier: `http://dx.doi.org/10.1089/cmb.2008.0137`
Author: `http://www.rpi.edu/~newbel/`

Poster U2: Estimating *p*-values for arbitrary HMMs / HBMs.