Centroid Methods for High-Dimensional Inference in Computational Biology

Lee A. Newberg[1,2], William Thompson[3], Sean Conlan[4], Thomas M. Smith[1,2], Lee Ann McCue[5], Charles Lawrence[3]

[1]The Wadsworth Center, New York State Department of Health, Albany, NY 12201
[2]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180
[3]Center for Computational Molecular Biology & Div. of Applied Mathematics, Brown University, Providence, RI 02912
[4]Columbia University, New York, NY 10032, USA
[5]Pacific Northwest National Laboratory, Richland, WA 99352

Motivation & Introduction

Advances in data collection technologies have generated increasingly large data sets available for analyses. While the emergence of such large data sets seems to imply more precise parameter estimation, paradoxically just the opposite is becoming common. This paradox arises because these technologies have created opportunities to draw inferences on previously unanswerable high dimensional questions. Optimization has been the workhorse of prediction and inference procedures in computational biology for virtually its entire history It has been the basis for the solution of most of its most important problems, e.g. optimal alignments, minimum free energy structure (MFE) prediction, and maximum *a posteriori* (MAP) motif finding. However, recently, the findings of Ding *et al*. (2005) and Newberg *et al.* (2007) have led us to question this apparoach. We propose, instead, a predictor that is centered in the posterior weighted ensemble of solutions. We suggest as a representative solution one that has as many components as possible in common with the posterior weighted ensemble. Such an estimator minimizes the Hamming risk and is called the *centroid.* Statistical decision theory shows that optimization based estimators minimize risk only under a zero/one loss function. This has little merit in discrete high-D spaces since the most likely solution rarely covers more than a minuscule fraction of the posterior space.

The use of the centroid as a representative solution in computational biology is not new. It was introduced by Miyazawa (1994) as a reliable alignment method for sequence pairs. For RNA secondary structure prediction, Ding *et al.* (2005) show that MFE predictions produced 43% more prediction errors and have lower sensitivity than alternative *centroid* estimators. Here, we show that for motif detection, in both simulated and real data, the centroid estimator combined with a full phylogenetic model helps improve specificity, sensitivity, and positive predictive value compared to MAP based solutions.

A Phylogenetic Gibbs Sampler

- The Gibbs Sampler is based on the Gibbs Recursive Sampler (Thompson et al. 2003).
- Sequences from closely related species are globally pre-aligned and their relationship described by a phylogenetic tree. See Fig. 1.
- Optimal sequence weights are generated from the phylogenetic tree (Newberg 2005).
- The joint probability of the aligned sequences at a given position is calculated using Felsenstein's (1981) tree-likelihood algorithm. Motif sites of width w are modeled as product multinomials with Dirichlet priors.
- For the model-update step, a Metropolis-Hastings algorithm is used. Starting with an existing model, the algorithm first draws a proposed model using sequence-weighted counts from the posterior Dirichlet distribution. The proposed model is accepted with a probability based on a Metropolis-Hastings ratio:

$$\min\left\{1,\prod\frac{\left(\dfrac{Fels(\Theta_p)}{\prod_{b=a,t,c,g}\Theta_p^{c_b}}\right)}{\left(\dfrac{Fels(\Theta)}{\prod_{b=a,t,c,g}\Theta^{c_b}}\right)}\right\}$$

where:
Θ is an existing model and Θ_p is a proposed next model
$Fels(\Theta_p)$ is the result of the Felsenstein calculation for the motif or background sites
c_b is the sequence-weighted counts for each base

Centroid alignment solution

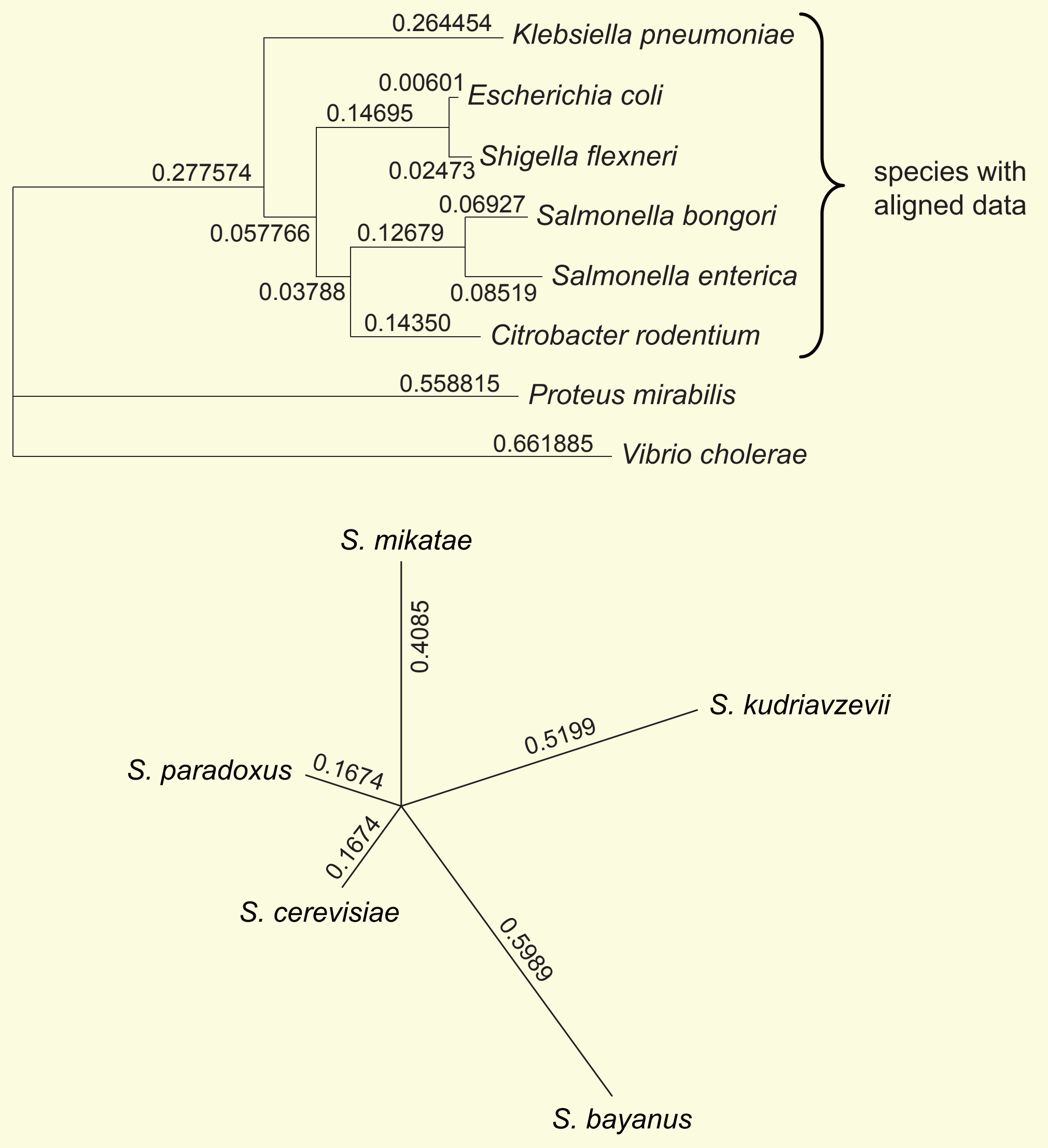
The Gibbs sampling procedure is initialized with a random alignment, the algorithm proceeds as follows:

1. a sequence is selected, and the probability of each possible number of sites, up to the maximum specified by the user, is calculated based on the current model;
2. the number of sites is sampled;
3. the predicted positions and types of the sites are sampled based on their probabilities, calculated as described by Thompson et al. (2004).
4. the motif models are updated from the sampled sites in all sequences.

- The recursive sampler previously reported the MAP (maximum *a posteriori* probability) alignment, the alignment with the maximimum log of the alignment probability minus the log of a background alignment.
- The above steps are executed during a "burn-in" period (typically 2,000-3,000 iterations of steps 1-4), followed by a fixed number of sampling iterations (typically 8,000-10,000) during which each sampled site is tracked.
- The centroid is calculated from these samples by identifying the alignment that has the minimum total distance to the other alignments in the ensemble of sampled alignments.
- Sites in two different samples have a distance of 1 if they do not overlap by half the site width.

Grant Support
2P20 RR01 5578-06 (COBRE NIH/NCRR)
DE-FG01-03ER0305 to CEL/LAM (DOE)
5K25HG003291 to LAN (NIH)

Figure 1. Phylogenetic trees



Simulated sequence data

Data sets:

- 100 sets of simulated sequences were generated from a phylogenetic tree (γ-proteobacterial or yeast).
- 0 to 4 simulated sites were planted in the simulated sequences: Crp sites for γ-proteobacterial and STB5p sites for yeast.
- Four different algorithms were tested:
  – Gibbs Recursive Sampler (provides MAP solution)
  – Gibbs Recursive Sampler with phylogeny incorporated (MAP solution)
  – Gibbs Centroid Sampler with phylogeny incorporated
  – PhyloGibbs, Siddharthan *et al*. (2005).

Results:

Specificity - The number of false positives (FP) predicted in a sequence with no planted motif sites.

Positive Predictive Value (PPV) - of the predictions made, what proportion are true positives (TP):

$$PPV = \frac{TP}{TP + FP}$$

Sensitivity - of all the positive sites, what proportion are detected:

$$Sensitivity = \frac{TP}{TP + FN}$$

Table 1. Specificity of the algorithms reported as the number of false positive predictions in 100 simulated proteobacterial or yeast sequences containing no simulated transcription factor binding sites averaged over 3 runs.

Algorithm	Proteobacterial	Yeast
MAP	218.3	n.d.
MAP+Phylogeny	2.3	93.3
Centroid+Phylogeny	1.0	26.0
PhyloGibbs	49.3	100.3

Figure 2. (A) PPV and (B) Sensitivity as a function of the number of planted simulated Crp sites in each simulated γ-proteobacterial sequence; (C) PPV and (D) Sensitivity as a function of the number of planted simulated STB5p sites in each simulated yeast sequence.

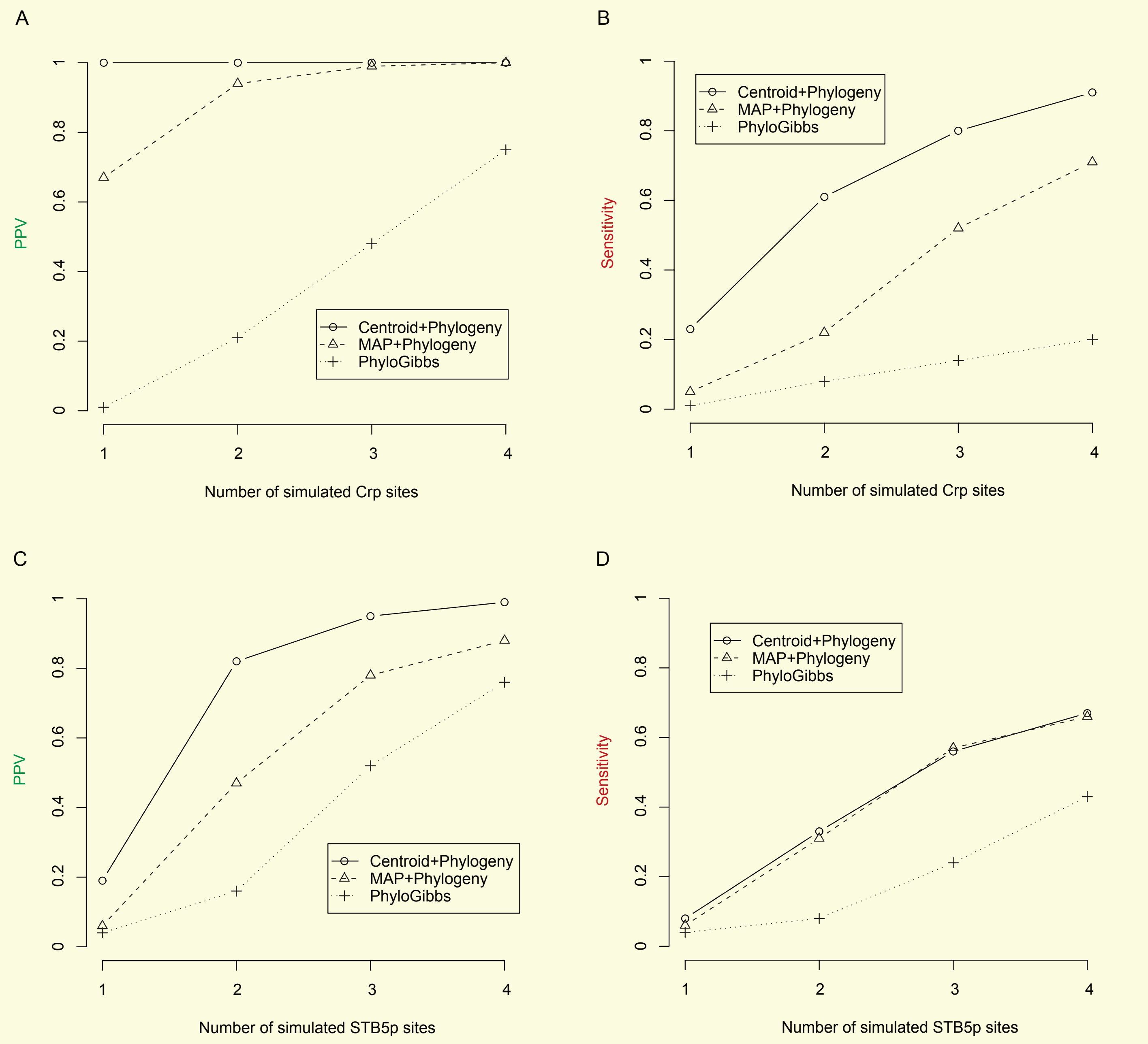


Table 2. The phylogenetic Gibbs Centroid Sampler on proteobacterial promoter sequences
72 sets of orthologous promoters across eight γ-proteobacterial species with at least one experimentally validated transcription factor binding site is present in each of the *E. coli* sequences

Motif Models	Total Predictions	True Positives	False Positives	False Negatives	PPV	Sensitivity
1	57.7	47.3	10.3	55.7	0.82	0.46
2	74.3	57.3	17.0	53.7	0.77	0.45
3	79.3	61.3	18.0	70.7	0.77	0.46

Table 3. The phylogenetic Gibbs Centroid Sampler on sets of co-regulated promoters
Orthologous promoters across eight γ-proteobacterial species grouped by regulon

Regulon	Genes	TF sites	Total Predictions	TP	FP	FN	PPV	Sensitivity
Crp	25	29	19.7	13.7	6.0	9.3	0.69	0.47
LexA	8	11	9.0	9.0	0.0	2.0	1.00	0.82
TyrR	5	8	5.0	5.0	0.0	3.0	1.00	0.62

Conclusions

The combination of a centroid predictor and a full phylogeny model yeilds enhanced specificity, sensitivity and PPV to Gibbs sampling based motif detection algorithms:

- Specificity - in the absence of transcription factor bindings sites, the phylogenetic model enables the Gibbs sampler (MAP and centroid versions) to avoid false positive predictions.
- PPV - the phylogenetic Gibbs sampler centroid alignment is more resistant to false predictions and thus achieves higher PPV than the optimization-based approaches.
- Sensitivity - the ensemble centroid alignment has improved sensitivity over optimization-based algorithms, suggesting that optimization approaches focus on a subset of the true sites, and in so doing derive an overly-focused model.

References

[1]Ding, Y.E., C.Y. Chan, and C.E. Lawrence. 2005. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA* **11**: 1157-1166.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**: 368-376.

Miyazawa, S. 1995. A reliable sequence alignment method based on probabilities of residue correspondences. *Protein Eng*. **8**: 999-1009.

Newberg, L., W.A. Thompson, S.P. Conlan, T.M. Smith, L.A. McCue, and C.E. Lawrence. 2007. A phylogenetic Gibbs sampler that yields centroid solutions for cis regulatory site prediction. *Bioinformatics* Submitted.

Newberg, L.A., L.A. McCue, and C.E. Lawrence. 2005. The Relative Inefficiency of Sequence Weights Approaches in Determining a Nucleotide Position Weight Matrix. *Statistical Applications in Genetics and Molecular Biology* **4:** 1-18.

Siddharthan, R., E.D. Siggia, and E. van Nimwegen. 2005. PhyloGibbs: A Gibbs Sampling Motif Finder That Incorporates Phylogeny. *PLoS Computational Biology* **1**: e67.

Pacific Northwest National Laboratory
Operated by Battelle for the U.S. Department of Energy

Center for Computational Molecular Biology

Wadsworth Center · NYS Department of Health

Rensselaer