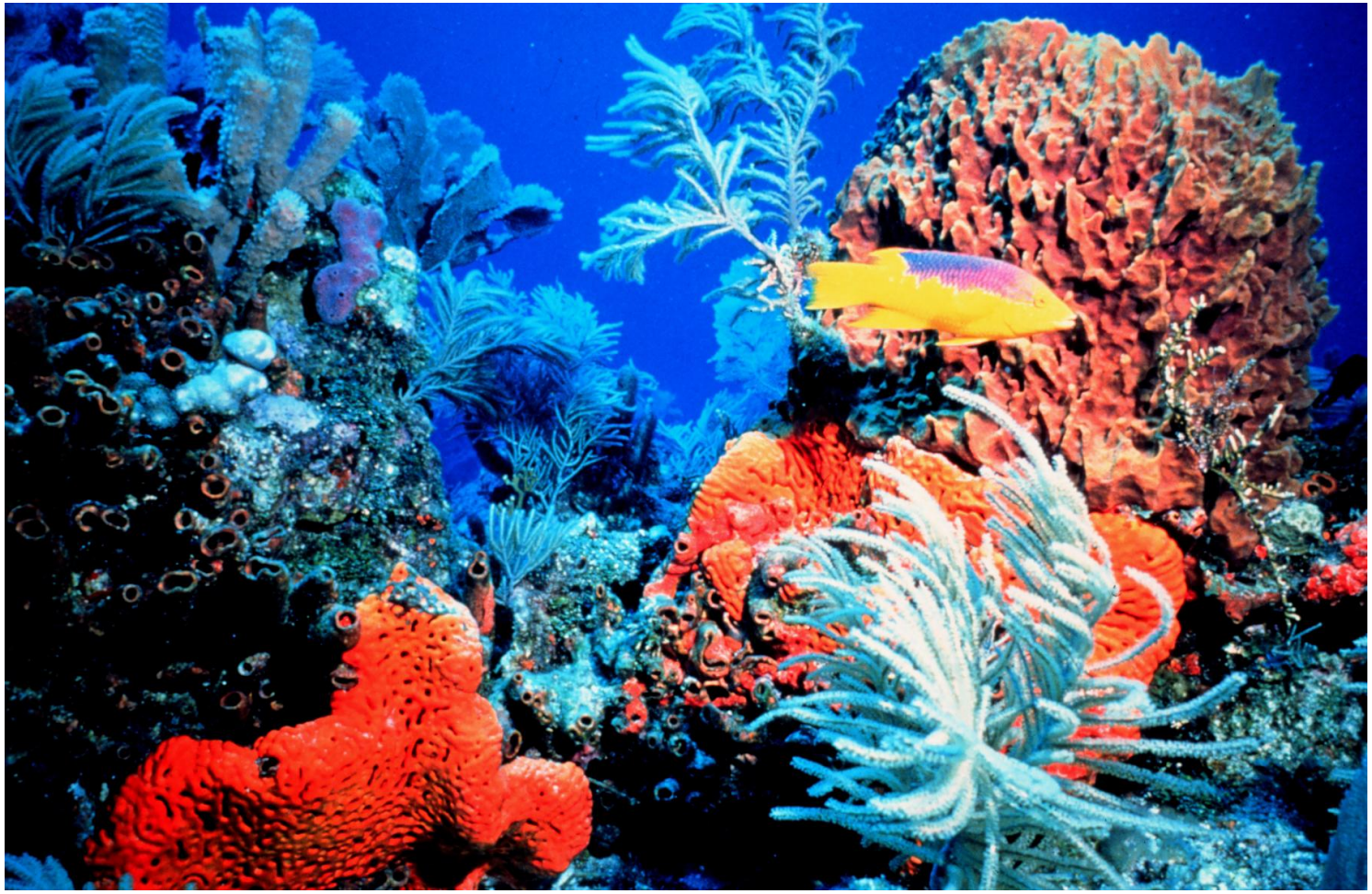
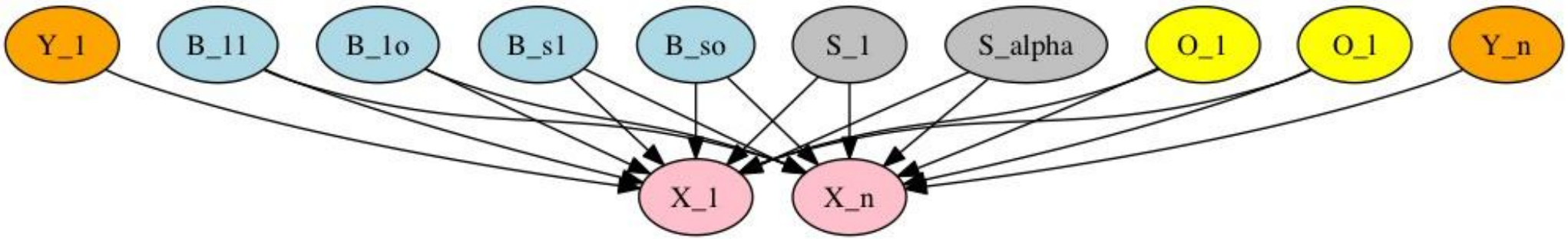
Final Project - Bayesian Probabilistic Assembler

Farhan Damani, Dan Adler



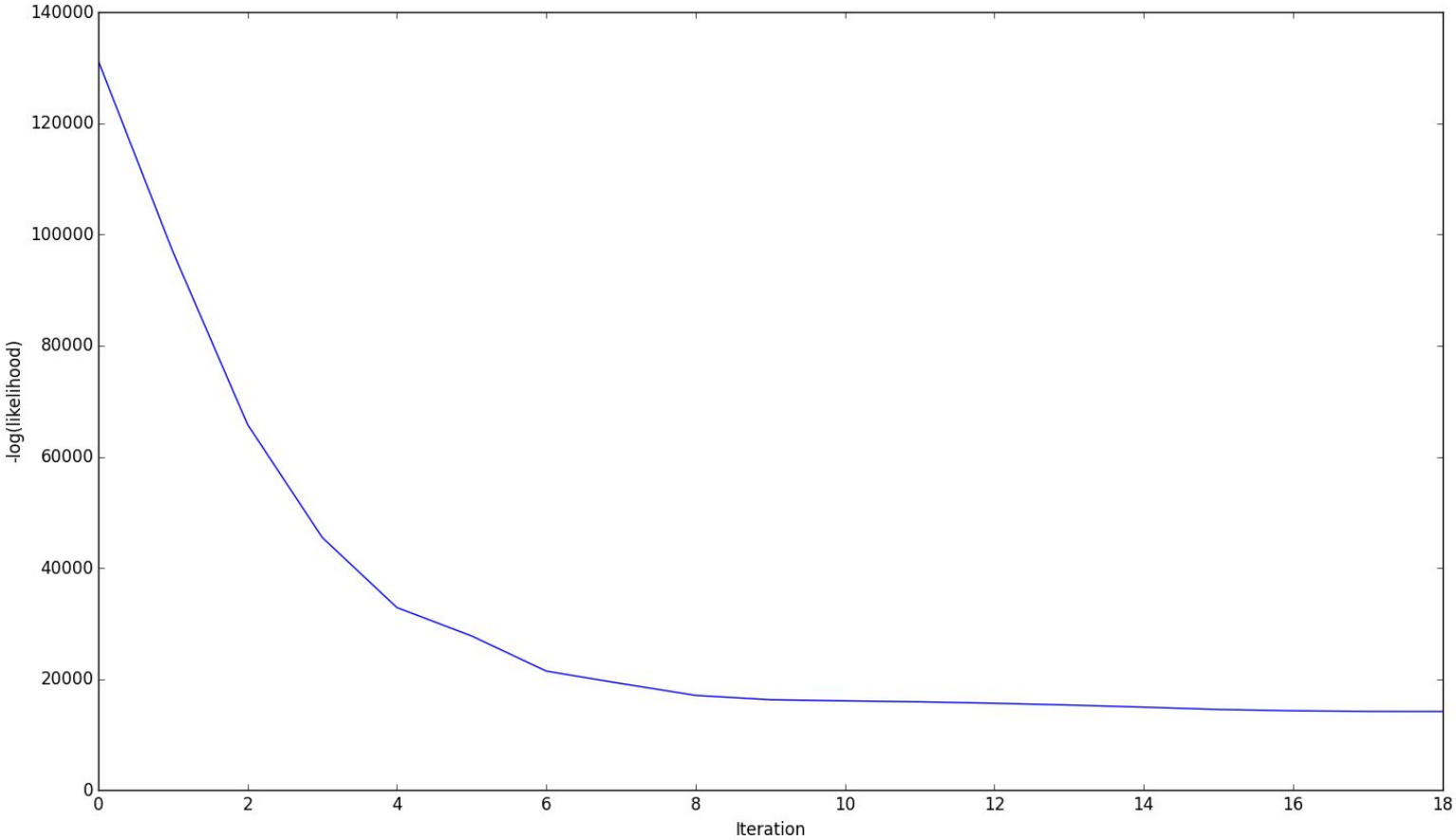

Y_1
B_11
B_1o
B_s1
B_so
S_1
S_alpha
O_1
O_1
Y_n
X_1
X_n

$$p(s) = (1-p)^{s-1}p,\ p = \frac{1}{Expected\ Contigs}$$

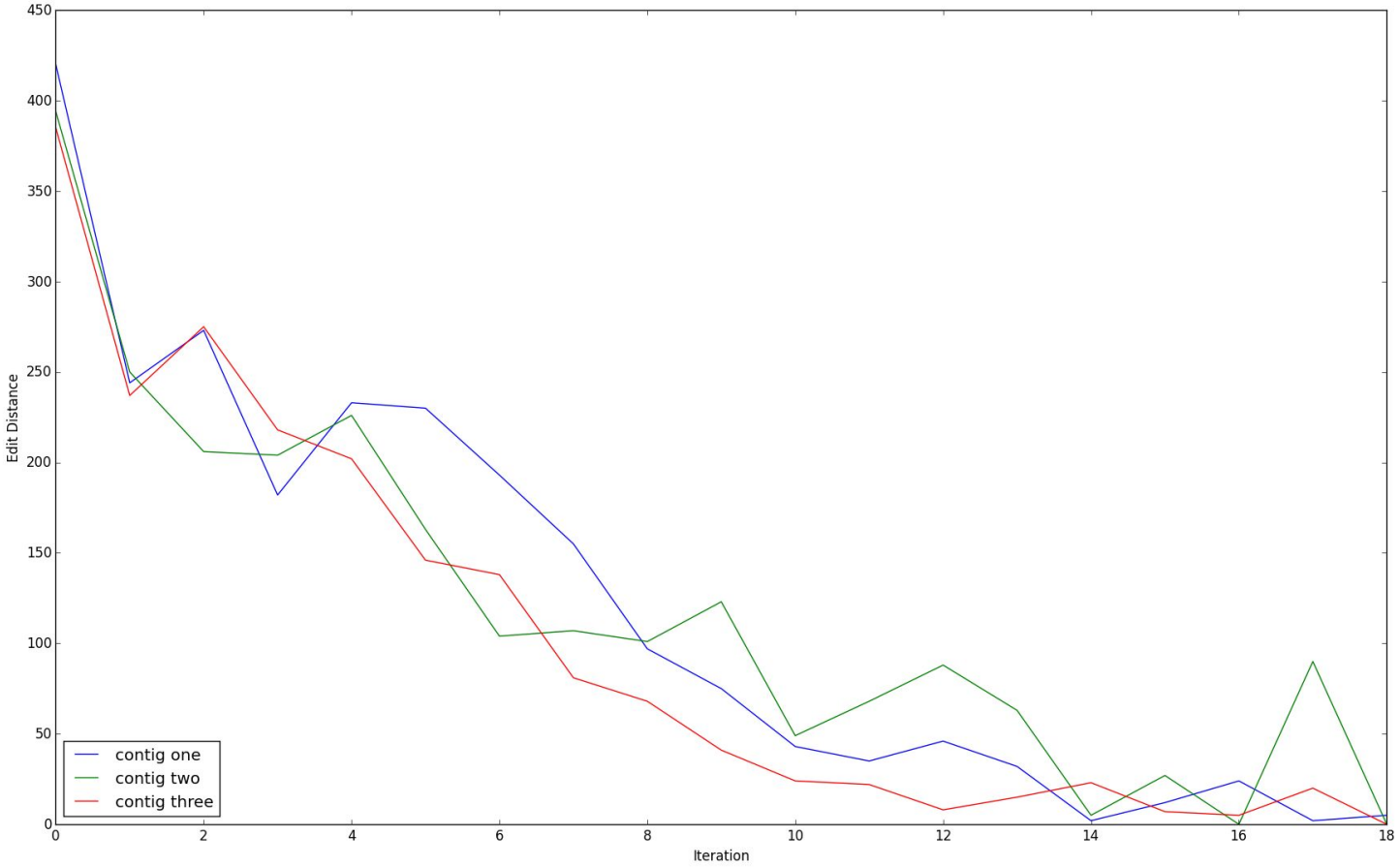
$$p(o) = \frac{1}{Contig\ Length}$$

$$p(x,y) = (1-p_{miss})^{n_{hit}} p_{miss}^{k-n_{hit}}$$

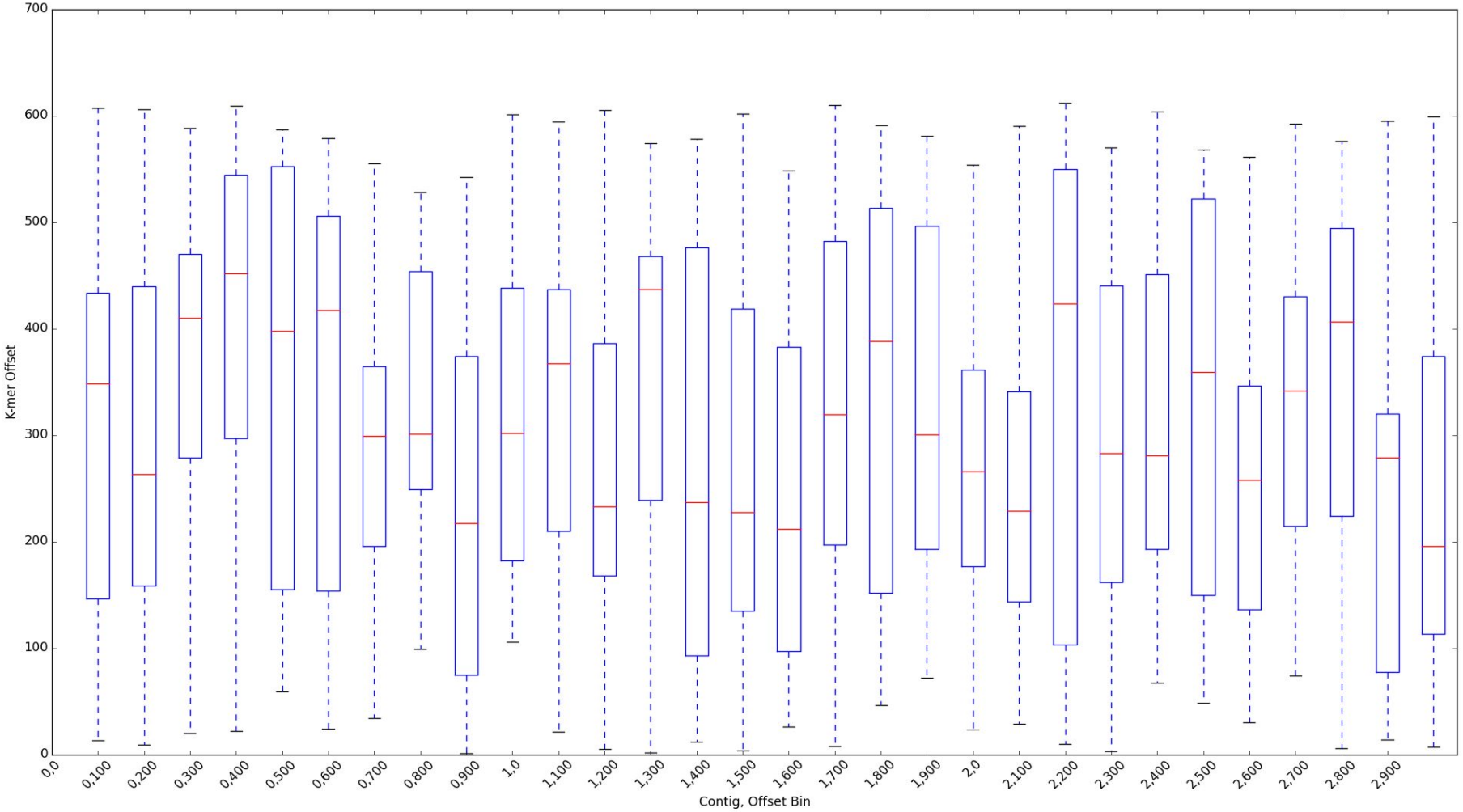
Likelihood and Convergence Guarantees



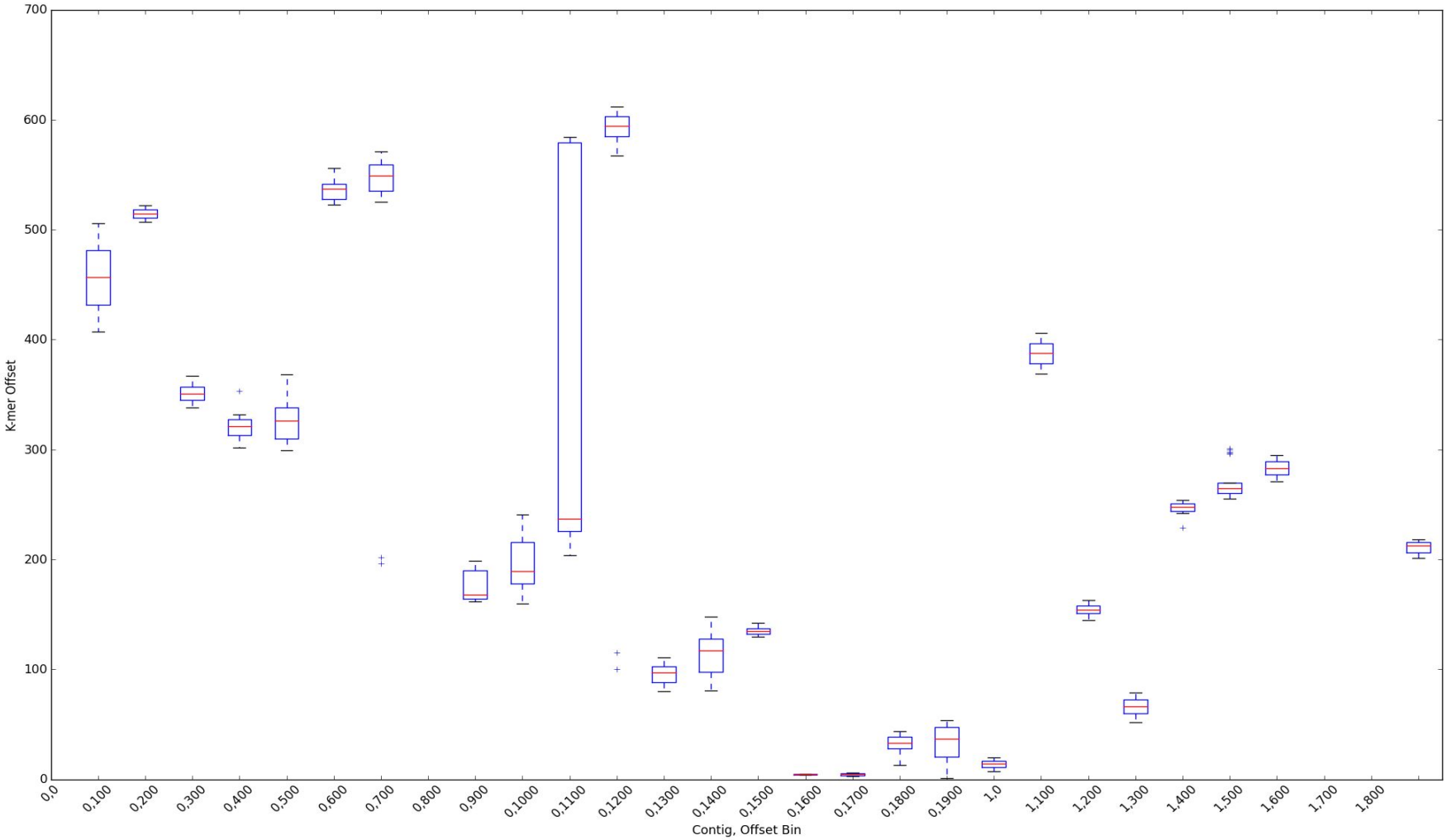
How do our contigs evolve over iterations?



Read Positions in Contigs at Initialization



Read Positions in Contigs at Convergence



Merges happened!

- merging contigs criteria: exact match of 15 nucleotides from the tail ends of any two nucleotides.
- P(2 contigs merging randomly) = 4.963083675E-24
  - (1/(4^30)) * 2 * (n choose 2), where n = number of contigs. In this case, n = 3.

Future work

- Use model to fully reconstruct original contig sequence
  - incorporate better prior on P(O) -- when contigs merge, reads need to be pushed to lower offsets, i.e. spaces of higher probability
- Accommodate for insertions/deletions
- Incorporate genetically diverse species into model

Literature cited and Support

1. Laserson, Jonathan et. al. “Genovo: De Novo Assembly for Metagenomes” Journal of Computational Biology, Vol. 18, No. 3. 2011.
2. Ray, Priyadip et. al. “Bayesian joint analysis of heterogeneous genomics data” Bioinformatics, Vol 30, No. 10. 2014.
3. Li-Thiao-Te Sebastian et. al. “Bayesian model averaging for estimating the number of classes: applications to the total number of species in metagenomics” Journal of Applied Statistics, Vo. 39, No. 7. 2012.
4. Alexis Battle
5. Ben Langmead