

# Inferring orthologous regions via a pseudo-Gibbs sampler: Finding the pieces of the rearrangement puzzle

Bob Mau<sup>1,2</sup>, Aaron Darling<sup>1,3</sup>,  
Frederick R. Blattner<sup>4,5</sup>, Nicole T. Perna<sup>1,5</sup>

Departments of Animal Health and Biomedical Sciences<sup>1</sup>, Oncology<sup>2</sup>, Computer Science<sup>3</sup>,  
Laboratory of Genetics<sup>4</sup>, and Genome Center of Wisconsin<sup>5</sup>  
University of Wisconsin – Madison, WI 53706.

## ABSTRACT

Chromosomes evolve by the modification, loss, gain, duplication, and rearrangement of the DNA sequences that comprise their genomes. Comparisons of complete genome sequences should illuminate the evolutionary dynamics that gave rise to a group of organisms, but except for closely related species or bacterial strains, the true evolutionary path is complex, ambiguous, and by its very nature, irreproducible.

A necessary first step is the identification of orthologous regions: DNA sequences that descend vertically from an ancestral genome. Historically, conserved gene order across a wide spectrum of species has been a litmus test for orthology. But breaks in the conserved order among orthologous sequence occur frequently, the product of genomic rearrangement, gene duplication, gene loss, or lateral gene transfer. Of these, only genomic rearrangements break order while preserving orthology.

We present a methodology to construct orthologous regions from a dense set of DNA markers that appear once and only once in each genome of interest. We cast this as a spatial problem in  $K$ -dimensions ( $K$  is the number of genomes) and each marker is a point in  $K$ -space. Dually, each genome is represented as a signed permutation induced by the order and orientation of the complete set of markers. Orthologous regions are  $K$ -dimensional analogs of the diagonals and anti-diagonals present in the global dotplots for pairs of related genomes, such as those generated by MUMmer (Delcher *et.al.*, 1999).

All automated marker schemes interject a certain level of noise, or false orthologs. Noise increases with evolutionary distance among the organisms being compared, as one is forced to lower thresholds in order to extract weak signal. Our approach is able to identify orthologous signals from spatial context, despite the random noise generated by gene duplications, convergent evolution, and random bits of sequence similarity.

Indicator random variables at each marker define a configuration space of zeroes and ones. A ‘one’ at a marker site represents a true ortholog, whereas a ‘zero’ denotes a false ortholog. A pseudo-Gibbs sampler simulates configurations by picking markers at random, then computing a score conditioned on the breakpoints formed by active (=1) markers in the current configuration. This score is transformed into a binomial probability, and the current marker’s status is updated by drawing a uniform random variate. In this manner, the pseudo-Gibbs sampler explores the configuration space, activating (“turning on”) or “turning off” markers one at a time. Some markers are consistently “turned on”, while others are routinely de-activated. The former coalesce into locally collinear blocks that form the backbone of orthologous regions.

By recording the relative frequency with which each marker is updated to a one, a posterior probability of orthology is assigned. Although these probabilities are very sensitive to

parameters in the sampler, their stochastic order is quite robust. Hence, the dimension of the problem is dramatically reduced from exponential in the number of markers to being equal to the number of markers. As we shall demonstrate, the choice of partition often depends on the biological question being asked.

The problem we address bridges the gap between two topics where much progress has been made: a) the automated identification of unique DNA markers for a group of organisms, and b) inferring the phylogeny and history of genomic rearrangements based on the order and orientation of orthologous markers. Phylogenetic inferences premised on genome rearrangements are not at all robust to the false classification of random sequence similarity as orthologous. Protecting against such errors without sacrificing legitimate but weak orthologies is the prime directive of this research.

The pseudo-Gibbs sampler is applied to several groups of bacterial genomes, illustrating the range and some of the limits of this approach. Finally, the method is applied to the eukaryotic troika of human-rat-mouse complete genomes.

We acknowledge the work of those who have pioneered the use of genomic rearrangements for phylogeny reconstruction, many of whom will be in attendance. A partial list includes:

- Ajana, Y., J. F. Lefebvre, et al. (2002). "Exploring the set of all minimal sequences of reversals—An application to test the replication-directed reversal hypothesis." Second International Workshop, Algorithms in Bioinformatics.
- Bafna, V. and P. Pevzner (1996). "Genome Rearrangements and Sorting by Reversals." SIAM Journal on Computing **25**(2): 272-89.
- Blanchette, M., T. Kunisawa, et al. (1999). "Gene order breakpoint evidence in animal mitochondrial phylogeny." J Mol Evol **49**(2): 193-203.
- Bourque, G. and P. A. Pevzner (2002). "Genome-scale evolution: reconstructing gene orders in the ancestral species." Genome Res **12**(1): 26-36.
- Hannenhalli, S. and P. A. Pevzner (1995). "Transforming Cabbage into Turnip (polynomial algorithm for sorting signed permutations by reversals)." Proc. 27th Annual ACM Symposium on the Theory of Computing: 178-89.
- Larget, B., D. Simon, et al. (2002). "Bayesian Phylogenetic Inference from Animal Mitochondrial Genomes." Journal of the Royal Statistical Society - Series B **64**(4): 681-693.
- Moret, B. M., L. S. Wang, et al. (2001). "New approaches for reconstructing phylogenies from gene order data." Bioinformatics **17**(Suppl 1): S165-73.
- Nadeau, J. H. and B. A. Taylor (1984). "Lengths of chromosomal segments conserved since divergence of man and mouse." Proc Natl Acad Sci U S A **81**(3): 814-8.
- Pevzner, P. and G. Tesler (2003). "Genome rearrangements in Mammalian evolution: lessons from human and mouse genomes." Genome Res **13**(1): 37-45.
- Tesler, G. (2002). "GRIMM: genome rearrangements web server." Bioinformatics **18**(3): 492-3.
- Tillier, E. R. and R. A. Collins (2000). "Genome rearrangement by replication-directed translocation." Nat Genet **26**(2): 195-7.

Our work builds and complements much of the above-cited research. In particular, we wish to thank Glenn Tessler and Guillaume Bourque for use of their personal versions of MGR and GRIMM in a subsequent phylogenetic analysis.