Statistical Evaluation of Genome Rearrangement

David Sankoff

Department of Mathematics and Statistics, University of Ottawa,
585 King Edward Avenue, Ottawa, ON, Canada, K1N 6N5
sankoff@uottawa.ca

Abstract. Genomic distances based on the number of rearrangement steps – inversions, transpositions, reciprocal translocations – necessary to convert the gene or segment order of one genome to that of another are potentially meaningful measures of evolutionary divergence. The significance of a comparison between two genomes, however, depends on how it differs from the case where the order of the n segments constituting one genome is randomized with respect to the other. In this presentation, we discuss the comparison of randomized segment orders from a probabilistic and statistical viewpoint as a basis for evaluating the relationships among real genomes. The combinatorial structure containing all the information necessary to calculate genomic distance d is the bicoloured "breakpoint graph", essentially the union of two bipartite matchings within the set of $2n$ segment ends, a red matching induced by segment endpoint adjacencies in one genome and black matching similarly determined by the other genome. The number c of alternating-colour cycles in the breakpoint graph is the key component in formulae for d. Indeed, $d \geq n - c$, where equality holds for the most inclusive repertory of rearrangement types postulated to account for evolutionary divergence.

Over a decade ago, it was observed in simulations of random genomes with hundreds of genes that the distance d seldom differed from n by more than a few rearrangements, even though it is easy to construct examples where d is as low as $\frac{n}{2}$. Our main result is that in expectation $c = C + \frac{1}{2}\log n$ for a small constant C, so that $n - d = O(\log n)$, thus explaining the early observations. We derive this for a relaxed model where chromosomes need not be totally ordered – they may include circular "plasmids" – since the combinatorics of this case are very simple. We then present simulations and partial analytical results to show that the case where all chromosomes are totally linearly ordered (no plasmids) behaves virtually identically to the relaxed model for large n.

Consider the "reuse" statistic $r = \frac{2d}{n}$. Although r can be as low as 1, in which case the breakpoint graph contains d cycles of the smallest size possible, r can also be as high as 2, in which case the cycles become larger and less numerous. Our results show that the latter is the case for random gene orders as well. Inference about evolution based on r, then, is compromised by the fact that a pattern of larger and fewer cycles occurs both when comparing genomes that have actually diverged through via high "reuse" rates and in genomes that are purely randomly ordered with respect to each other.

A. Apostolico et al. (Eds.): RECOMB 2006, LNBI 3909, p. 84, 2006.