Baculovirus phylogeny based on genome rearrangements

Daniel Goodman, Noah Ollikainen, Chris Sholley, Pavel Pevzner

UNIVERSITY OF CALIFORNIA - SAN DIEGO

1 Introduction

Why Gene Rearrangement?

As the availability of whole genome sequences has increased over the past decade, the data has provided unprecedented insight into evolutionary relationships. Recent sequencing of complete genomes has led to an explosion of phylogenomics studies addressing evolutionary problems by considering entire genomes as compared individual genes. Analyzing rare genetic changes in whole genomes has the potential to overcome many of the problems commonly associated with gene sequenced-based phylogenetic analysis.

Why Viruses?

Viral evolution remains poorly addressed by this technique. In the last ten years the number of fully sequenced virus genomes has increased by an order of magnitude.

We chose baculoviridae as a test family because their evolutionary relationships are well studied and significant genome rearrangement has occurred between members of the family.

2 GenomeOrder package

To automate our analysis we developed a general purpose genome phylogeny software package called **GenomeOrder**. At its core is a clustering and refinement algorithm tied to NCBI's **BLASTCLUST**, which identifies orthologous sets of single genes shared between all queried genomes. We then use the order of those shared genes to find whole genome rearrangements and the associated phylogeny through the **MGR** package.

Some clades within the tree will have the same order among these conserved genes. These more closely related genomes are run again through GenomeOrder in an iterative process, until no new clusters can be found.

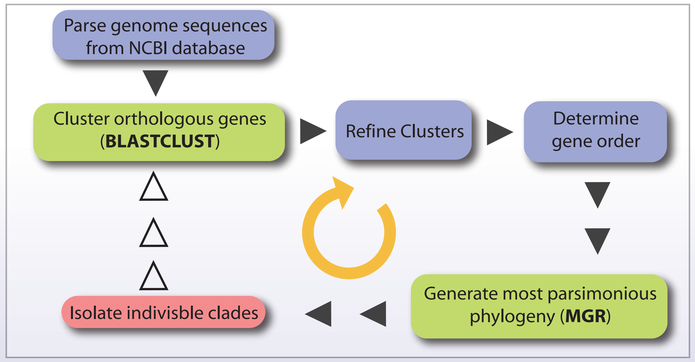

Figure 1. Overview of GenomeOrder software. Steps carried out by GenomeOrder are in blue, external programs are in green. After parsing, clustering is done using all-against-all BLAST. Clusters containing one gene per genome are extracted, larger clusters are refined. Gene order data is sent to MGR, which calculates relatedness. If MGR finds no differences for some subset of the genomes, that subset is run again to find more homologs.

3 Cluster Refinement

With this method, all genomes have to be represented in the alphabet of the same genes such that each gene appears once and only once in each genome. Many clusters contain multiple paralogous genes per genome, and cannot be used to determine relative gene ordering.

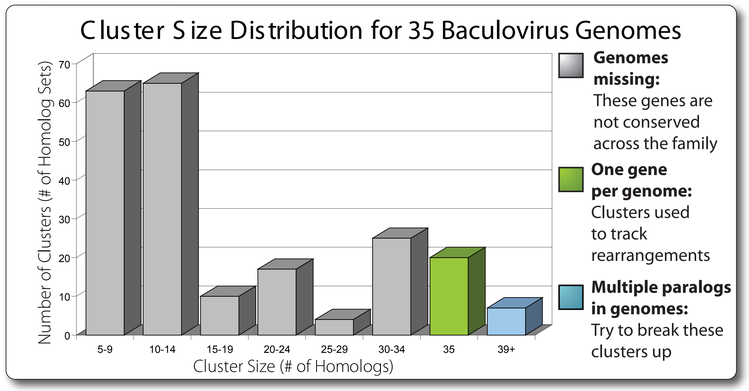

Figure 2. A distribution of cluster size after the initial clustering step for 35 genomes. Most clusters do not contain a member from every genome (grey), some contain one member from each of the 35 (green), and some contain many more. In this case, one had 320 genes.

To recover larger clusters, we employ a two-part heuristic algorithm that treats genes as vertices and their pairwise BLAST scores as edges described in Figure 3.

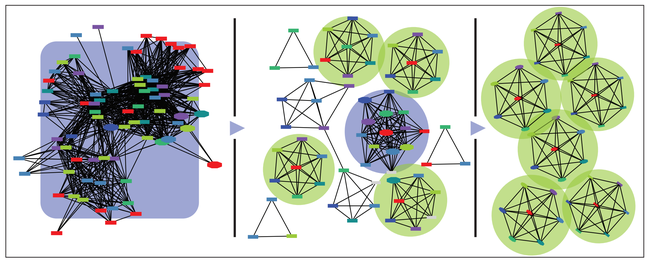

Figure 3. A schematic showing the cluster refinement algorithm. Genes in the cluster are vertices, colored by species. Blast hits are edges. First, weak edges are removed, and the remaining edges are searched for best subclusters (in green) in a manner similar to Prim's algorithm. The algorithm is run iteratively on large subclusters that remain (blue).

4 Constructing Phylogeny

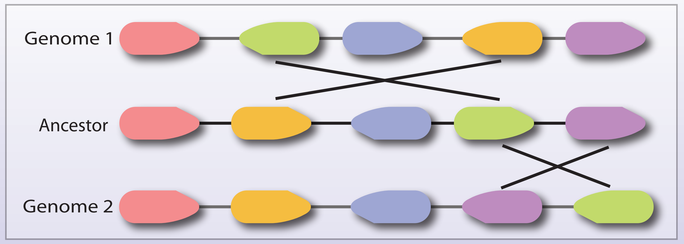

Figure 4. Schematic diagram showing gene orders and two rearrangements between hypothetical genomes. Rearrangements indicate distinct evolutionary events, allowing us to use gene order to derive evolutionary relatedness. MGR (Multiple Genome Rearrangements) was used to construct phylogeny based on most plausible rearrangement scenario for a set of gene orders.

5 Phylogeny of Baculovirus

We used GenomeOrder to derive baculovirus phylogeny. In Figure 5, the unrooted tree for 9 baculovirus genomes is shown, along with a comparison with a gene sequence-based phylogenetic tree from *Herniou 2001*. These two trees largely agree. Figure 6 is a more complete tree of the baculovirus family, including several new genomes recently sequenced.

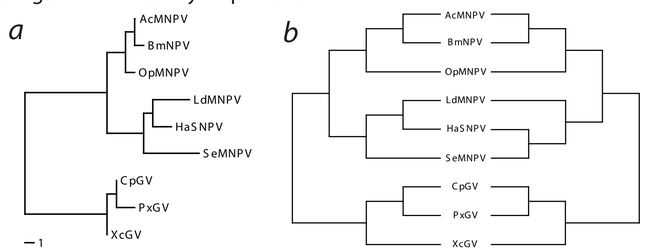

Figure 5. (a) A phylogenetic tree of 9 baculovirus genomes based on gene order. Granuloviruses, MNPVs, and NPVs separate as expected. (b) A comparison with the phylogeny described in Herniou 2001. Our tree is on the left.

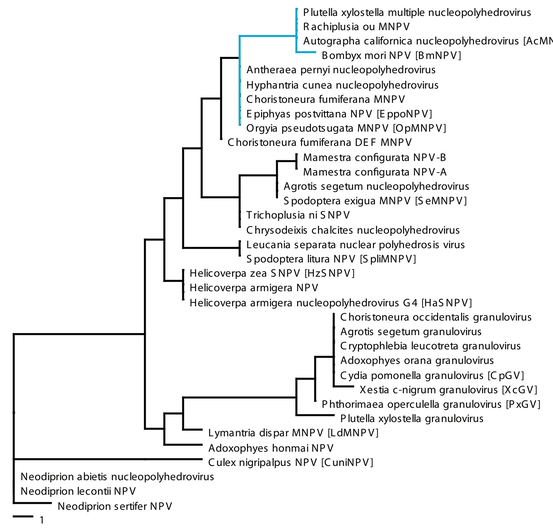

Figure 6. A phylogenetic tree of all available baculovirus genomes. The second iteration of GenomeOrder is shown in blue.

6 Conclusion

Genome rearrangements were used to construct baculovirus phylogeny. Our tree largely agrees with previous baculovirus studies that derived phyogeny based on gene sequence. Our software has demonstrated itself to be accurate in constructing phylogenetic relationships, and could also be useful in constructing phylogeny of bacterial and mitochondrial genomes.

8 References

Hannenhalli, S., Chappey, C., Koonin, E., Pevzner, P. (1995). Genome Sequence Comparison and Scenarios for Gene Rearrangements: A Test Case. Genomics. 30:299-311.

Herniou, E., Luque, T., Chen, X., Vlak, J., Winstanley, D., Cory, D., O'Reilly, D. (2001). Use of Whole Genome Sequence Data To Infer Baculovirus Phylogeny. Journal of Virology. 75(17):8117-8126.

Bourque, G. and Pevzner, P.A. (2002) Genome-Scale Evolution: Reconstructing Gene Orders in the Ancestral Species. Genome Res., 12(1), 26-36.