Fast Identification of Gene Clusters in Prokaryotic Genomes

Karine St-Onge[1,2], Anne Bergeron[1,3], Cedric Chauve[1,2,3]
(1) Comparative genomics laboratory, UQAM
(2) LaCIM, UQAM
(3) Computer Science Department, UQAM

Poster at *3rd RECOMB Workshop on Comparative Genomics*, 2005.

Fast Identification of Gene Clusters in Prokaryotic Genomes

Karine St-Onge[1,2], Anne Bergeron[1,3] and Cedric Chauve[1,2,3]

(1) CGL, (2) LaCIM and (3) Computer Science Department, UQÀM, Montréal

email: {anne.bergeron,cedric.chauve}@uqam.ca

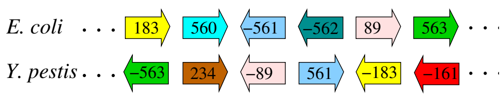
Introduction

The detection of **gene clusters** that are **conserved** in several genomes, in terms of gene proximity and **gene content**, have proved to be an invaluable tool in the comparative analysis of prokaryotic genomes. It has applications, for example, in predicting **operons**, **functional association** between groups of genes or putative **genome rearrangements**. We propose an **efficient algorithm** for computing gene clusters, that can handle large scale comparisons.

The algorithm

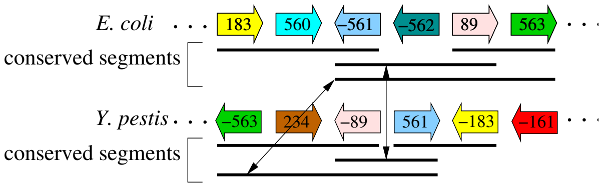
Principle. Our algorithm is based on the hypothesis (Sankoff 2003) of the **prevalence of short inversions** in the evolution of prokaryotic genomes. Its principle is to first detect short gene segments conserved among several genomes, then to "assemble" these segments into gene clusters.

Input. Sequences of signed integers (gene families):

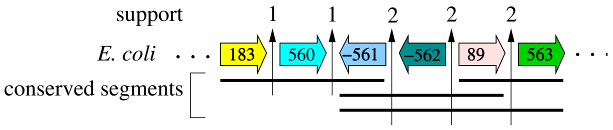


Gene family 183: threonine synthase
Gene family 560: hypothetical protein
Gene family 562: inner membrane transport protein
Gene family 89: transaldolase B . . .

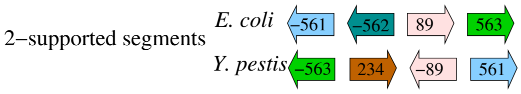
Phase 1. Identification of δ-conserved segments. (δ: short positive integer) A segment of at most δ genes, $g_1 \ldots g_k$ is δ-conserved if there is another segment, either in the same genome, or in another genome, of at most δ genes that is bounded by genes belonging to the families g_1 and g_k. For example, with $\delta = 4$,



Phase 2. Segmentation of genomes into ω-supported gene segments. (ω: positive integer) An intergenic region is ω-supported, if it belongs to at least ω conserved segments.

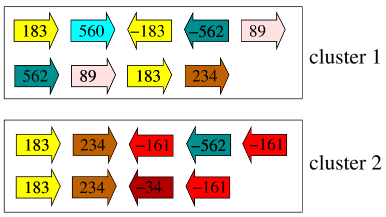


A segment of genes is ω-supported if all its intergenic regions are ω-supported.



The result of this phase is a set of gene segments.

Phase 3. ρ-clustering of the set of ω-supported gene segments. (ρ: positive integer) Single-linkage clustering: if the segments $g_1 \ldots g_\ell$ and $h_1 \ldots h_k$ share respectively $\rho/100\,\%$ of common families, they belong to the same cluster. For example, if $\rho = 80$, the four gene segments below form two clusters,



while they would form a single cluster if $\rho = 50$.

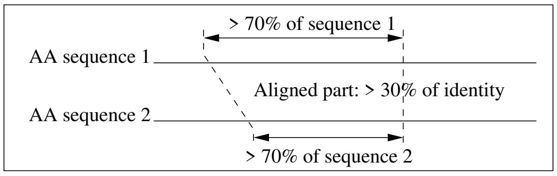
Time and space complexities. $O(\delta n^2)$ worst-case time complexity and $O(n^2)$ worst-case space complexity.

Remark. Approaches similar to our, based on the detection of short conserved segments, have been used (Fujibuchi *et al.* 2000, Rogozin *et al.* 2002), but focusing only on *conserved pairs of adjacent genes*. The other common approach, based on a formal definition of gene cluster – gene teams (Nicolas *et al.* 2003), homology teams (He *et al.* 2005), domain teams (Pasek *et al.* 2005), conserved intervals of sequences (Schmidt *et al.* 2005) –, faces the problem of finding a good tradeoff between the realism of such a definition, with respect to the biological data, and the computational complexity of the related algorithms.

Application: γ-Proteobacteriae

Genomes. Comparison of 12 genomes of γ-Proetobacteria: *Escherichia coli K12*, *Salmonella typhimurium*, *Yersinia pestis CO92*, *Yersinia pestis KIM*, *Buchnera aphidicola*, *Wigglesworthia glossinidia*, *Haemophilus influenzae*, *Pasteurella multocida*, *Pseudomonas aeruginosa*, *Xylella fastidiosa*, *Xanthomonas axonopodis*, *Xanthomonas campestris*.

Gene families. They are based on annotations for tRNAs and rRNAs, and on a single linkage clustering based on sequence similarity analysis with BLASTp for coding genes: two genes are linked if their BLASTp alignment overlaps at least 70% of the sequence of both genes, with at least 30% of identity of the aligned amino acids.



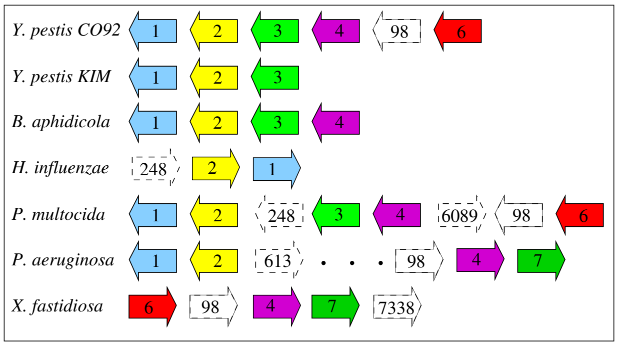
The total number of genes in these genomes is 39050, and we identified 27726 gene families. Of these, 26154 families contains only one gene, and thus are present in only one genome, and one contains 432 genes, composed exclusively of genes coding for ATP-binding proteins. A total of 26681 families are present in only one genome, while 93 are present in at least 10 genomes, and 25 families are present in all 12 genomes.

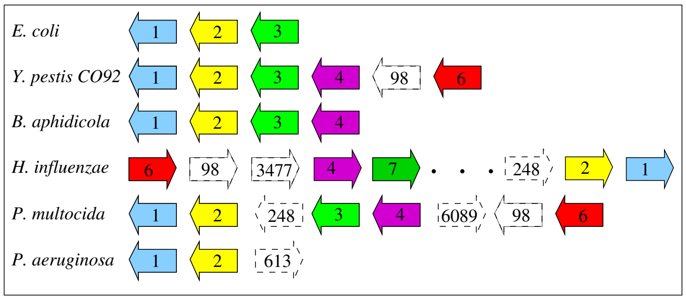
Parameters and performances. We computed gene clusters with all combinations of the three parameters δ, ω and ρ, with $\delta = 2, 3$, $\omega = 1, \ldots, 2\delta - 1$ and $\rho = 50, 60, 70, 80, 90, 100$, which yields 48 different combinations. This resulted in a 10 min. computation, with a linear space practical complexity.

The number of clusters for the data set of 12 genomes goes from 199 clusters ($\delta = 3$, $\omega = 2$ and $\rho = 100$) to 930 clusters ($\delta = 2$, $\omega = 2$ and $\rho = 60$). The set of all clusters is available at `adn.bioinfo.uqam.ca/~genoc/CLUSTERS`.

Examples of clusters: the tryptophan operon. We present below two clusters, highlighting how our algorithm was able to detect biologically relevant clusters related to the tryptophan operon.

Gene families of the tryptophan operon:
1 (trpA) – 2(trpB) – 3,7(trpC) – 4,5(trpD) – 6(trpE)





Conclusion.

The algorithm we propose is both time and space efficient, which allows to handle large datasets. It relies strongly on the hypothesis of short inversions and on the notion of support of an intergenic region, that is new, as far as we know. Moreover, the parameters δ, ω and ρ allow the user to modify the quality of the conservation defining clusters.

The simplicity of our algorithm leaves the door open for many improvments: we are currently testing more sophisticated clustering approaches, and a generalized notion of support that would handle the phylogenetic relationship between species.

Address for correspondence: Départment d'Informatique, UQÀM, CP 8888, succ. centre-ville, Montréal (QC), H3C 3P8