A Novel Approach for Comparative Genomics & Annotation Transfer

Alban Mancheron, Raluca Uricaru, Eric Rivals

A Novel Approach for Comparative Genomics & Annotation Transfer

Alban MANCHERON <alban.mancheron@lirmm.fr>

Raluca URICARU <raluca.uricaru@lirmm.fr>

Eric RIVALS <eric.rivals@lirmm.fr>

UNIVERSITÉ MONTPELLIER II — SCIENCES ET TECHNIQUES DU LANGUEDOC

LIRMM

Context "What is genome comparison good for?"

Genome comparison

- is crucial for genome annotation, regulatory motifs identification, and vaccine design
- aims at finding genomic regions either specific to or in one-to-one correspondence between individuals/strains/species
- proves useful to transfer annotations from a known genome to a new one

However, current methods do not suit the whole spectrum of applications and genome sizes.

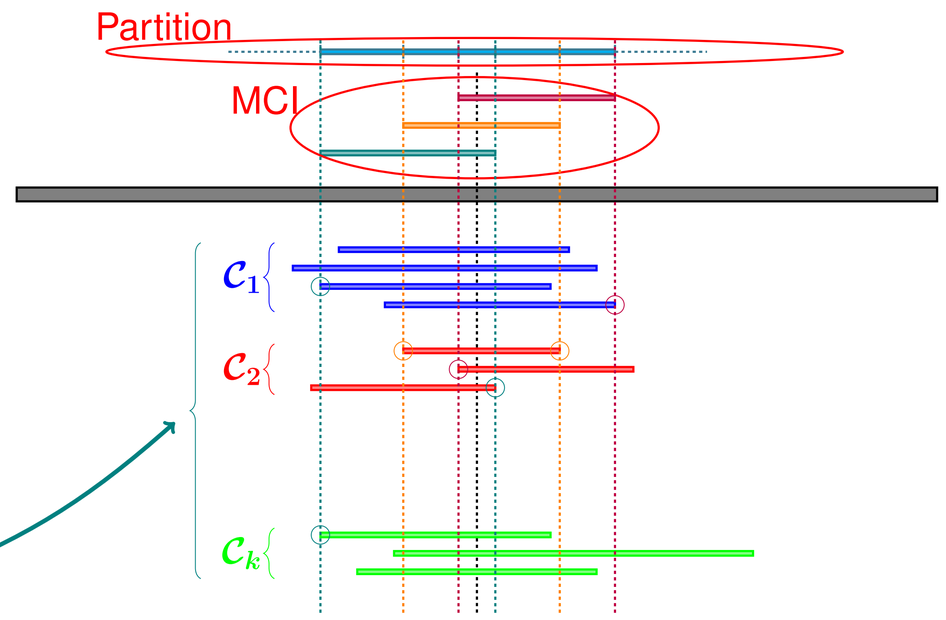
Innovative approaches are still needed

QOD GUI: a novel approach to genome comparison

We propose, QOD GUI, an approach to whole genome comparison that differs from multiple alignment:

- it compares one target genome to k reference genomes
- uses as input pairwise local alignments (BLAST-like) between the target and each of the references
- it computes on the target genome all Maximum Common Intervals (MCI, see Definition)
- MCI: maximal region from the target that can be aligned pairwise to regions in each of the references
- QOD GUI partitions the target genome into regions that are
  1. not alignable with each reference genome
  2. shared, but having a single alignment possibility in each reference genome
  3. shared with several potential alignments
- Class 2 regions have potential orthologs in the other genomes, Class 1 regions help finding genome specific regions.

The MCIs, the partitions, and annotations features are visualized and browsable on the GUI.

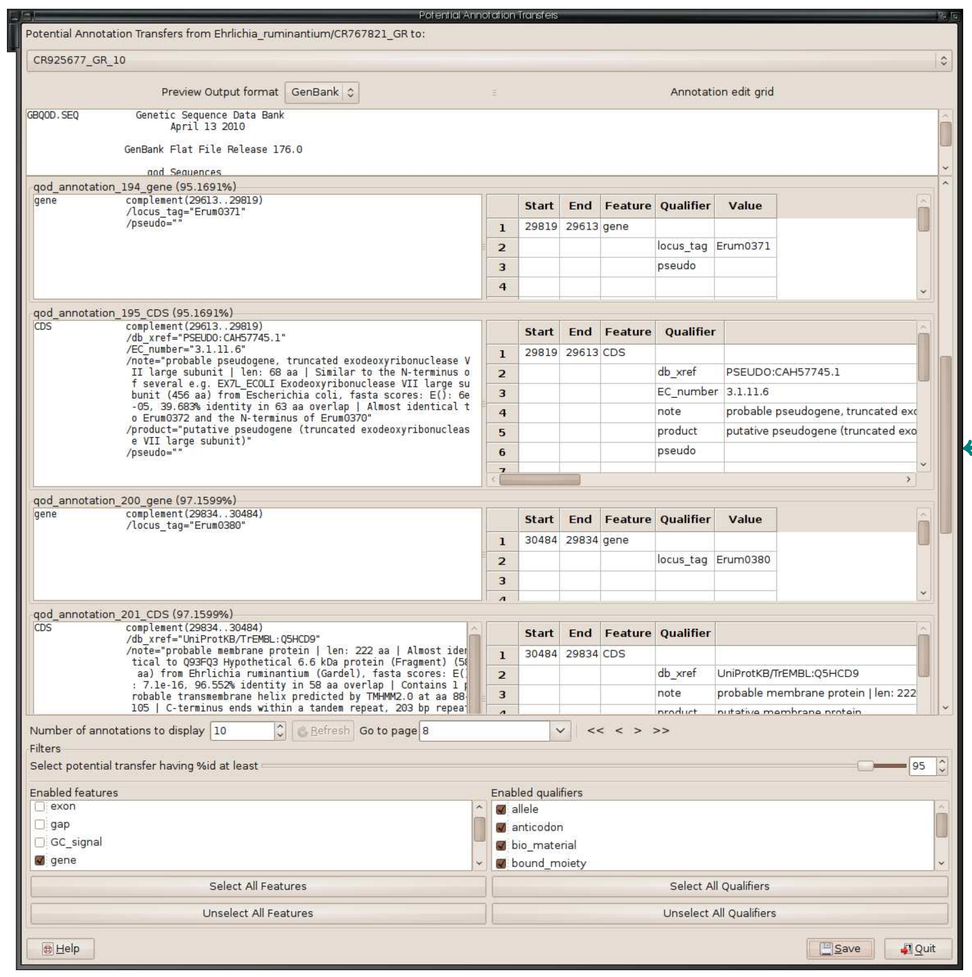


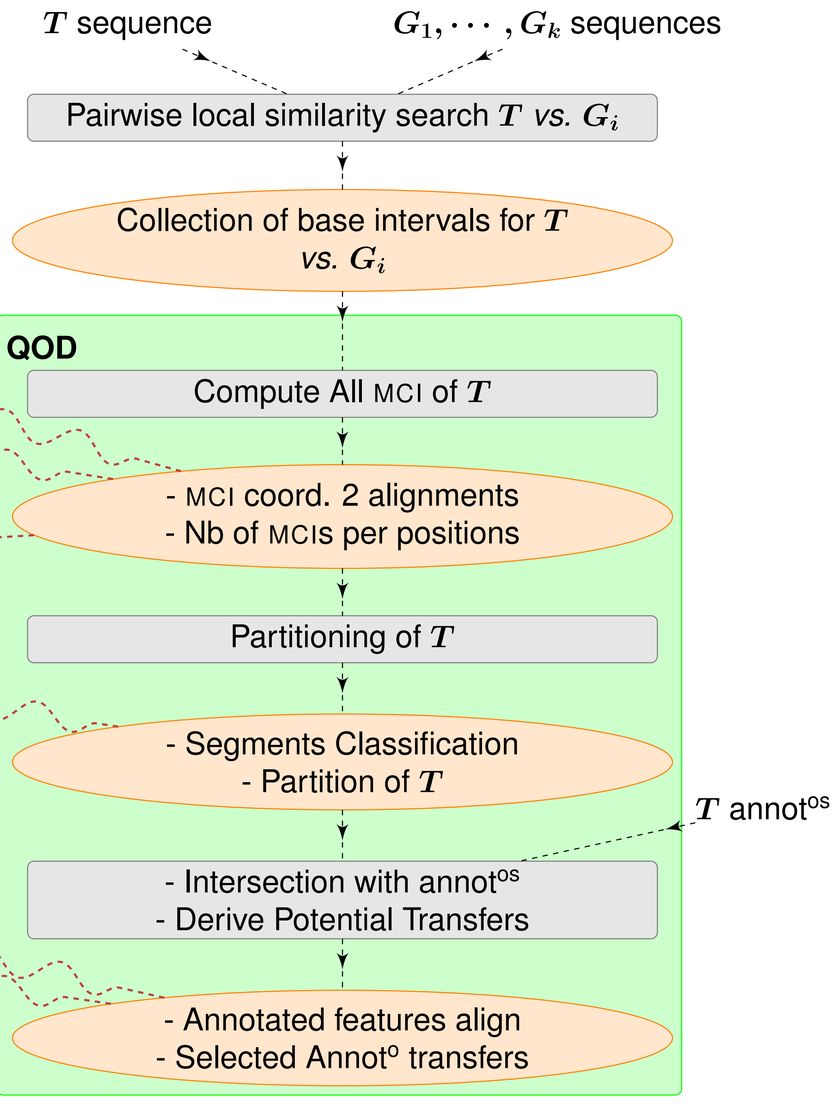
Maximal Common Intervals (MCI)

For $1 \leq i \leq k$, let $\mathcal{C}_i$ be the set of local alignments between target genome T and reference genome G_i.

An interval J is *common* to all $\mathcal{C}_{1 \leq j \leq k}$ if and only if for any collection $\mathcal{C}_j$ there exists an interval say I_i^j from $\mathcal{C}_j$ such that $J \subseteq I_i^j$.

Assume $J = [p, q]$ with $p < q$ is a common interval. J is said to be a *maximal* if neither $[p-1, q]$ nor $[p, q+1]$ are common intervals.





Annotation Transfer

QOD GUI

- intersects the target genome's annotations with regions from the partition, and
- derives for annotation in shared region their pairwise alignments

QOD GUI declares as **potentially transferable**

- all features falling entirely in a region offering only one alignment possibility
- the user can interactively select among those features to transfer
- according to various criteria: alignment's percent of identity, feature class, etc.

QOD GUI features

- is fast, flexible, interactive
- works on multiple OS
- exploits multi-core computers
- accepts gzip compressed inputs
- can process unassembled/unfinished genomes
- distributed under CÉCILL license (GPL compliant).

Conclusion

- QOD GUI is a different and novel approach to compare multiple whole genomes
- Case study on bacterial strains of E. ruminantium exhibits 92 genes and many orthology relationships that have been missed by previous genome sequencing and comparative projets.
- More informations are available at `http://www.atgc-montpellier.fr/qod/`

ANR

Methods and Algorithms for Bioinformatics LIRMM

A Novel Approach for Comparative Genomics & Annotation Transfer

Alban MANCHERON <alban.mancheron@lirmm.fr>
Raluca URICARU <raluca.uricaru@lirmm.fr>
Eric RIVALS <eric.rivals@lirmm.fr>

LIRMM, CNRS Université Montpellier 2 - CC 477
161, rue ADA 34095 Montpellier CEDEX 5

Abstract

With the rapid development of sequencing techniques, the situation where a newly sequenced genome needs to be annotated using available genomes from close species should become more prevalent in the future. However, because of the cost of genome finishing we may have to handle incomplete or not fully assembled genomes. Undoubtedly, the need for comparative annotation will increase, but the genomic community still lack computational solutions that are both efficient and sensitive under various conditions. Present approaches are mainly based on the sequence similarity detected at the gene or protein levels, which are mostly further analysed independently one of each other, despite the dependency implied by the genome.

Hence, we propose a novel approach to genome comparison and use it to develop a system that transfers annotations between the compared genomes. Besides features' sequence similarity, it accounts for the synteny it detects across multiple genomes. This approach is simple for it avoids to solve complex questions that makes other approaches computationally hard.

The underlying idea is to partition a focus genome according to its pairwise similarities with the other compared genomes. The question is formulated as searching for the intervals that are shared across all genomes under consideration, and maximal in length (*i.e.*, not extendible). If a genomic region is covered by at least one interval it is conserved across all genomes, and the number of such intervals tells how many possibilities exist for aligning it with different regions of the other genomes. Hence, our algorithm partitions the genome into regions following two criteria: 1/ being shared or unshared across all genomes, 2/ offering a unique or several alignment possibilities. The annotation transfer procedure crosses the focus genome's annotations with these regions and automatically derives the possible alignments for each feature. All features falling entirely in a region offering only one alignment possibility are declared as *potentially transferable*, and the user may interactively select among those according to various criteria: alignment's percent of identity, feature class, etc.

We implemented these procedures in an efficient and flexible tool, named QOD, equipped with a user-friendly graphical interface. Graphical and textual results representations allow both to grasp the overall genome similarity at a glance and to browse the conserved and unshared features in various ways. This enables the investigation of genome specific genes or of rearrangements, and copy number variations, for instance. For it does not require the genome sequence to be completely assembled, our approach allows to compare and pre-annotate unfinished genomes, as well as assemblies of Next Generation Sequencing data.

Keywords: *bioinformatics, comparative genomics, computational tools, evolution, genomic structure*

Web Page: `http://www.atgc-montpellier.fr/qod/`