Introducción a la pan-genómica microbiana – GET_HOMOLOGUES

Pablo Vinuesa

Centro de Ciencias Genómicas – UNAM, vinuesa@ccg.unam.mx

http://www.ccg.unam.mx/~vinuesa/

Introducción a la filoinformática – pan-genómica y filogenómica. IBBM-CONICET, UNLP Argentina. Julio 2018


Outline

I. Basic concepts in microbial comparative- and pan-genomics
   • Methods for orthology inference
   • Genome mosaicism, genomic islands and HGT
   • The prokaryotic gene and genome space: cloud, shell, core
   • Microbial core- and pan-genomes

II. get_homologues, a powerful and highly configurable software package for microbial pan-genomics
   • Overview of the package’s capabilities (get_homologues pipelines)
   • Using GET_HOLOLOGUES to explore factors affecting the calling of orthologs

III. A pangenomic analysis of pIncA/C plasmids using GET_HOMOLOGUES
   • Defining a robust core- and pan-genome of pIncA/C plasmids
   • Exploring the gene space of pIncA/C plasmids: core, shell, cloud
   • Pan-genome trees vs. core genome trees (supermatrices)
   • Identifying lineage-specific genes in NDM-1 producing plasmids

Computational methods to identify orthologs

Orthology, co-orthology and paralogy relationships in the evolution of four genes that arose from a single common ancestor.

Gene trees are not to be equated with species trees!!!


Computational methods to identify orthologs

• Most direct approach but known to be subjected to increasing artifacts as evol. dist. increases due to:
  - alignment problems
  - long-branch attraction
  - heterotachy
  - tree searches stuck in local sub-optimal maxima

• More adequate for bacteria, as they don’t follow a clear tree-like evolutionary path (HGT).
• The only practical strategy for large datasets, therefore most common method in comparative genomics

Methods to identify orthologs

• Orthostrapper
• RIO
• COCO-CL
• TReP
• PHOG
• MetaPhoR

• Best-based heuristic
• Blast-based heuristic

• BDBHs
• COGs
• OrthoMCL
Computational methods to identify orthologs: BBHs

Grouping of genes in different species that are each others’ BLAST BBHs into sets of orthologs and co-orthologs.

Linking pairs of BBHs from multiple genomes has a property of self-verification, as their consistency would be very unlikely due to chance, especially between phylogenetically distant lineages.

Methods for clustering of pairwise BBHs vary, but the most widely used approach involves a single-linkage clustering procedure, where any two clusters sharing a common BBH are merged until convergence.

Problems with: differential gene loss, domain recombination/gain/loss


Computational methods to identify orthologs: BDBHs

The bidirectional best-hit approach (BDBH) as implemented in GET_HOMOLOGUES (Contreras-Moreira & Vinuesa 2013)

1. Find inparalogues in all genomes
2. Find reciprocal best blast hits between reference and remaining genomes

1. Cluster sequences based on filtering criteria such as:
   - % alignment overlap
   - min. % sequence identity
   - E-score cut-off
   - Pfam domain composition
   - synteny


Computational methods to identify orthologs: OrthoMCL

OrthoMCL forms groups of orthologs and co-orthologs using a Markov clustering process involving iterative simulations of stochastic (randomized) flow on the edges of a BBH graph, with clusters of desired tightness identified depending on a given ‘inflation’ parameter determined by trial and error.

Li L et al. Genome Res. 2003;13:2179-2189

Basic concepts in bacterial comparative genomics – Mosaic genome structure and the pan-genome

**Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli***

- Patchy distribution among strains
- Ecological specialization
- Plasmids, phages, GIs, ICE, integrons ...
- Highly transferable (HGT)

*accessory/feible genome (shell + cloud genes)*

- Core genome
- Orthologous gene families
- Species/population trees
- Population genetics
- Genome annotation

**Distribution of COGs by the number of organisms included in each cluster:**

- Highly conserved core genes
- Narrowly distributed cloud genes
- Moderately conserved shell genes

Fit of an exponential decay function with three exponents:

The core (~70 clusters), shell (~5700 clusters) the cloud (~24,000 clusters)

Koonin & Wolf, NAR 2008. Vol 36, No. 21

Basic concepts in bacterial comparative genomics – the pan-genome (Tettelin et al. 2005)


Size estimations of the core and pan-genomes of the animal-associated Rhizobiaceae (5 Brucella spp., 4 Bartonella spp.)

- Brucella suis 1330
- Brucella melitensis 16M
- Brucella abortus bv. 1 strain
- Brucella canis ATCC 23365
- Brucella ovis ATCC 25840
- Brucella henselae str. Houston-1
- Brucella canis ATCC 23365
- Bartonella bacilliformis K583
- Bartonella quintana str. Toulouse
- Bartonella bongori str. 2569
- Bartonella henselae str. Houston-1
- Bartonella quintana str. Toulouse
- Ochrobactrum anthropi ATCC 49188
- Mesorhizobium loti MAFF303099

Size estimations of the core and pan-genomes of the plant-associated Rhizobiaceae (2 Rhizobium, 2 Sinorhizobium, 3 Agrobacterium)

- Rhizobium leguminosarum bv. viciae 3841
- Rhizobium etli CNF 42
- Agrobacterium radiobacter K84
- Agrobacterium tumefaciens str. C58
- Sinorhizobium meliloti 1021
- Mesorhizobium nonagri nsp. 1
- Agrobacterium vitis S4
- Sinorhizobium medicae WSM419
- Brucella suis 1330
- Brucella melitensis 16M
- Brucella abortus bv. 1 strain
- Brucella canis ATCC 23365
- Bartonella bacilliformis K583
- Bartonella quintana str. Toulouse
- Bartonella henselae str. Houston-1

GET_HOMOLOGUES + GET_PHYLOMARKERS: open-source code for phylogenomics and microbial pan-genomics

- Open source package written in Perl and R, freely available through GitHub
- Computes clusters of homologous clusters (paralogues and orthologues)
- Computes pan-genome matrices, ANI matrices, fits mixture-binomial models to estimate pan-genome sizes, identifies clade-specific genes and clade-specific gene family expansions, computes diverse genome trees...
Overview of the GET_HOMOLOGUES package

1. Clustering of prot. and nt secs. (CDSs) in homologous groups (paralogs, orthologs).
2. Identification of orthologous intergenic regions.
3. Definition, statistical analysis and plotting of pan- and core-genomes, LSEs...
4. Comparative genome analyses, including analysis of the pan-genome structure (core, soft-core, shell and cloud) and identification of lineage-specific clusters.
5. Computation and graphical representation of pan-genome matrices, trees and ANIb

Overview of the GET_HOMOLOGUES package

Block 1: Extracting features from GBK files

Block 2: Blasting, sorting & indexing

Block 3: Clustering

Block 4: Parsing, statistical analysis and graphical display of pan-genomes and core-genomes with auxiliary scripts

get_homologues: a software package for microbial comparative genomics.

get_homologues installation

- Written in Perl, tested in 32 and 64-bit Linux, Mac OS X systems
- Installed along with dependencies using install.pl
- Bundled with precompiled binary files (COGtriangles, MCL and BLAST)
- Required dependencies (Perl modules)
  - Bio::Seq, Bio::SeqIO,
  - Optional software and database dependencies:
    - hmmscan (from the HAMMER3 package) for Pfam domain scanning
    - Pfam HMM library and hmmpress for db formatting
    - BerkeleyDB (perl module)
  - R

Running get_homologues.pl

Input data: whole genome GenBank or FASTA files

Typing $ ./get_homologues.pl -h, on the terminal will show the basic options:

-usage: ../get_homologues.pl [options]
  -h this message
  -v print version, credits and checks installation
  -d directory with input amino acid FASTA files (.faa) or (overrides -i) GenBank files (.gbk), 1 per taxon; allows for new files to be added there later, creates output folder named 'directory_homologues'
  -i input amino acid FASTA file with [taxon names] in headers, (required unless -d is set, creates output folder named 'file_homologues'
  -o only run BLAST/Pfam searches and exit (useful to pre-compute searches)
  -c report genome composition analysis (follows order in -I file if enforced, ignores -r,-t,-e)
  -s save memory by using BerkeleyDB; default parsing stores sequence hits in RAM
  -r runmode [local|cluster] (default local)
  -I file with .faa/.gbk files in -d to be included (takes all by default, requires -d)

Optional parameters:
- only run BLAST/Pfam searches and exit (useful to pre-compute searches)
- report genome composition analysis (follows order in -I file if enforced, ignores -r,-t,-e)
- save memory by using BerkeleyDB; default parsing stores sequence hits in RAM
- runmode [local|cluster] (default local)
- file with .faa/.gbk files in -d to be included (takes all by default, requires -d)
- algorithms instead of default bidirectional best-hits (BBBHs):
  -G use Cogtrriangle algorithm (COGS, PubMed=20439257) (requires 3+ genomes|taxa)
  -M use orthoMCL algorithm (OMCL, PubMed=12952885)
  - output truncated
get_homologues: a software package for microbial comparative genomics.

**Running get_homologues.pl**

- **Usage:** `../get_homologues.pl` [options]
  
  - `--usage` shows usage information
  - `--cont` continues script execution
  - `--options` control clustering:
    - `-D` require equal Pfam domain composition
      when defining similarity-based orthology
    - `-S` min % sequence identity in BLAST query/subj pairs
      (range [1-100], default=1 [BDBH|OMCL])
    - `-N` min BLAST neighborhood correlation
      (PubMed=18475320, range [0,1], default=0 [BDBH|OMCL])
    - `-b` compile core-genome with minimum BLAST searches
    - `-t` report sequence clusters including at least t taxa
    - `-a` report clusters of sequence features in GenBank files
    - `-g` report clusters of intergenic sequences flanked by ORFs
    - `-f` filter by % length difference within clusters
    - `-r` reference proteome .faa/.gbk file
    - `-e` exclude clusters with inparalogues
    - `-x` allow sequences in multiple COG clusters
  - OPTIONS that control clustering:
    - `-t` report sequence clusters including at least t taxa
    - `-a` report clusters of sequence features in GenBank files
      instead of default "CDS" GenBank features
      - requires -d and .gbk files
      - example -a 'tRNA, rRNA'
    - `-g` report clusters of intergenic sequences flanked by ORFs
      in addition to default "CDS" clusters
      - requires -d and .gbk files
    - `-f` filter by % length difference within clusters
    - `-r` reference proteome .faa/.gbk file
    - `-e` exclude clusters with inparalogues
    - `-x` allow sequences in multiple COG clusters
    - `-F` orthoMCL inflation value

**Running get_homologues.pl – choosing a proper clustering algorithm**

**Dataset name** (number of genomes) **number of sequences** **smallest-largest proteome (KB)** **taxonomy** **reference**

<table>
<thead>
<tr>
<th>Dataset name</th>
<th>number of genomes</th>
<th>number of sequences</th>
<th>smallest-largest proteome (KB)</th>
<th>taxonomy</th>
<th>reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enterobacteriaceae42</td>
<td>196654</td>
<td>3725-5768</td>
<td>family</td>
<td>this paper</td>
<td></td>
</tr>
<tr>
<td>Escherichia coli101</td>
<td>497131</td>
<td>4968-5608</td>
<td>species</td>
<td>this paper</td>
<td></td>
</tr>
<tr>
<td>Escherichia coli57</td>
<td>269710</td>
<td>4688-5972</td>
<td>species</td>
<td>this paper</td>
<td></td>
</tr>
<tr>
<td>Escherichia coli20</td>
<td>93612</td>
<td>4688-5608</td>
<td>species</td>
<td>Touchon et al. (2)</td>
<td></td>
</tr>
<tr>
<td>Vibrionales38</td>
<td>165536</td>
<td>3425-6064</td>
<td>species</td>
<td>Thompson et al. (5)</td>
<td></td>
</tr>
<tr>
<td>Gammaproteobacteria13</td>
<td>42011</td>
<td>564-5571</td>
<td>phylum</td>
<td>Leral et al. (3)</td>
<td></td>
</tr>
<tr>
<td>yeast6</td>
<td>32079</td>
<td>4714-5861</td>
<td>class</td>
<td>Salichos &amp; Rokas (4)</td>
<td></td>
</tr>
</tbody>
</table>
Alignment coverage is the most influential parameter for the calculation of core-genomes.

Increasing coverage thresholds improves the comparison with Pfam protein domains.

Estimating core genomes:
- E. coli 20 Willembrock-fit
- E. coli 20 Tettelin-fit
- E. coli 57 Willembrock-fit
Introducción a la pan-genómica microbiana – GET_HOMOLOGUES

Estimating pan-genome sizes for *E. coli* 20 (upper row) and *E. coli* 57 (lower)

- **BDBH**: 124 new genes /genome
- **COG**: 105 new genes /genome
- **OMCL**: 107 new genes /genome

- **66 new genes /genome**
- **63 new genes /genome**
- **62 new genes /genome**

Running `get_homologues.pl` – what are the sizes of the *E. coli* core- and pan-genomes

- Draft genomes
- Complete genomes
- Core genome for 89 *E. coli* genomes
- Pan-genome for 89 *E. coli* genomes

- **BDBH + COG + OMCL intersection**
- **COG + OMCL intersection**

On average, every new *E. coli* genome adds some 55 new genes to its known pangenome

get_homologues: a software package for microbial comparative genomics.

**Different algorithms produce notably different pan-genomic cluster sizes**

- Gammaproteobacteria13
- Escherichia coli20

Venn analysis as a way to define consensus pan-genomes

- A
- B
- C
- D

Introducción a la pan-genómica microbiana – GET_HOMOLOGUES

E. coli pangenome phylogeny estimated using the consensus panmatrix of presence/absence data for 46 fully sequenced genomes of Escherichia spp. / Shigella spp. and Citrobacter spp.

standard parsimony analysis: heretic search+TBR swapping + 1000 bootstrap replicates

K12 & B strains

126 EHEC-specific genes when compared with K12 groups

EHEC

Shigella spp.

E. albertii, E. fergusonii, Citrobacter spp.

The GET_HOMOLOGUES tutorial:

1) cp the instructions script to your home
/export/space2/tib/filo/martes/protocols/code4_GET_HOMOLOGUES_TIB17.txt

2) Open the script with nedit and keep a terminal open to issue the commands

Analyses to be performed:

A pangenomic analysis of plnCA/C plasmids using GET_HOMOLOGUES

• Defining a robust core- and pan-genome of plnCA/C plasmids
• Exploring the gene space of plnCA/C plasmids: core, shell, cloud
• Pan-genome trees vs. core genome trees (supermatrices)
• Identifying lineage-specific genes in NDM-1 producing plasmids