README file for GRASTv1.0.pl

Genome Reduction Analysing Software Tool (GRAST).

Produced by Christina Toft and Mario A. Fares
Date 03/04/06

Reference and more information: Toft, C and Fares, MA (2006). GRAST: a new way of Genome Reduction Analysis using comparative genomics. Bioinformatics *(accepted)*

GRAST allows the user to analyse genome reduction by whole genome comparison between a reduced genome and a reference genome (a close relative to the reduced genome). A number of the options in GRAST can also be used for a whole genome comparison of non-reduced genomes.

Requirements:

- installed bastsall
- perl modules:
    - Math::BigFloat
    - GD::SVG or GD::Graph depending on the version of GRAST the user runs (outputs svg or gif files) plus the modules they depend on.
- input files should be in genbank format (.gbk or .cgi).
- .ptt (COG) files should by in the same directory as the genbank files (only required for options that involve knowledge of gene function)
- the genbank and the COG files should have the same name except for the extensions, .gbk/.cgi and .ptt respectively

GRAST can been run automatically or with command line arguments.

Running GRAST using command line arguments:

GRASTv1.0.pl <reduced genome's genbank file> <reference genome's genebank file> <options and output options>

Output options:

-g shows the placement of common/none common genes in the two genomes
-c shows the placement of common genes in the two genomes
-n shows the placement of none common genes in the two genomes

-r gene pair plot of the two genomes
-s CSGC: the start position in the two genomes, the genes names and the number of genes that has been lost between genes
-m shows the placement of the CGSC on the two genomes
-y overall statistics on CSGC's
-x statistics on CSGC
-l distribution of lost genes in COG - text file
-b shows the distribution of lost genes within/in COG categories and the overall loss of genes
-p orthologous gene pairs - gene names + e-values
-k list of genes that don't have a ortholog in the other genome
-t observed and expected (before and after genome reduction) on seeing gathering of genes with similar function
-j shows the distribution of junk DNA
-h distribution of junk DNA - text file
-all runs all output options

Options:

-a <number of simulations>
default=100
The number of simulations to be performed to estimate the expected probability of seeing the seen translocation events by random chance (before and after genome reduction)

-e <e-value>
default=1e-6
The cut-off e-value for the BLAST searches between the two genomes

-o <e-value>
default=1e-6
The cut-off e-value for the intra BLAST searches.

-d <directory>
default=current directory
The directory in which the output files are created (including the blast files created BLAST searches between the two genomes).

-f
GRAST will ask the user for the name of output file for each output.

(only available when GRAST is run with user interaction)

To run GRAST with user interaction:

GRASTv1.0.pl <reduced genome's genbank file> <reference genome's genebank file> <options>

if the `reduced genome's genbank file` and/or the `reference genome's genebank` file does not exist GRAST will ask the user to enter the file names.

Screen shot:

```
OPTIONS:
     1    shows the placement of common/non-common genes in the
            two genomes (g)
     2    shows the placement of common genes in the two genomes
            (c)
     3    shows the placement of non-common genes in the two
            genomes (n)
     4    gene pair plot of the two genomes (r)
     5    CSGC: the start position in the two genomes, the genes
            names and the number of genes that has been lost
            between genes (s)
     6    shows the placement of the CGSC on the two genomes (m)
     7    overall statistics on CSGC's (y)
     8    statistics on CSGC (x)
     9    distribution of lost genes in COG - text file (l)
     10   shows the distribution of lost genes within/in COG
            categories and the overall loss of genes (b)
     11   orthologous gene pairs - gene names + e-values (p)
     12   list of genes that don't have a homolog in the other
            genome (k)
     13   observed and expected (before and after genome
            reduction) on seeing gathering of genes with similar
            function (t)
     14   shows the distribution of junk DNA (j)
     15   distribution of junk DNA - text file (h)
     q    quit.
Pick one of the options:
```

Options:

See above.

Definitions:

Common genes = gene pairs that has been found each other by mutual top hit BLASTP search (orthologous gene pair).

Clusters of Orthologous Groups of proteins (COGs) = Each COG consists of individual orthologous genes or orthologous groups of paralogs from three or more phylogenetic

lineages (Tatosow et al., 1997). The COG also classifies the genes according to their function. This is predicted by knowing the function of one or more of the members of the COG and assuming the rest of the genes in the COG have the same function.

Translocated = when two genes are together in the reference genome but not in the reduced genome (figure 1 A).

Gathering caused by gene loss = two genes have been gathered by gene loss in the reduced genome when all the genes that are between them in the reference genome has been lost - no rearrangement have occurred (figure 1 B).

Gathering caused by movement = two gene have been gathered by movement in the reduced genome when the genes have had to move to come together in the reduced genome (figure 1 C).

Conserved Gene Succession Cluster (CGSC) = Genes that have retained there order over evolution. For two genes to be in a CGSC they have to be adjacent in the reduced genome and the genes if any between them in the reference genomes have to have been lost (figure 1 E).

Note: GRAST performs a intra BLAST search of the two genomes. For non-common genes that have a hit in the common genes that satisfy a set cut-off values are said to have a paralog and are removed from the analysis.

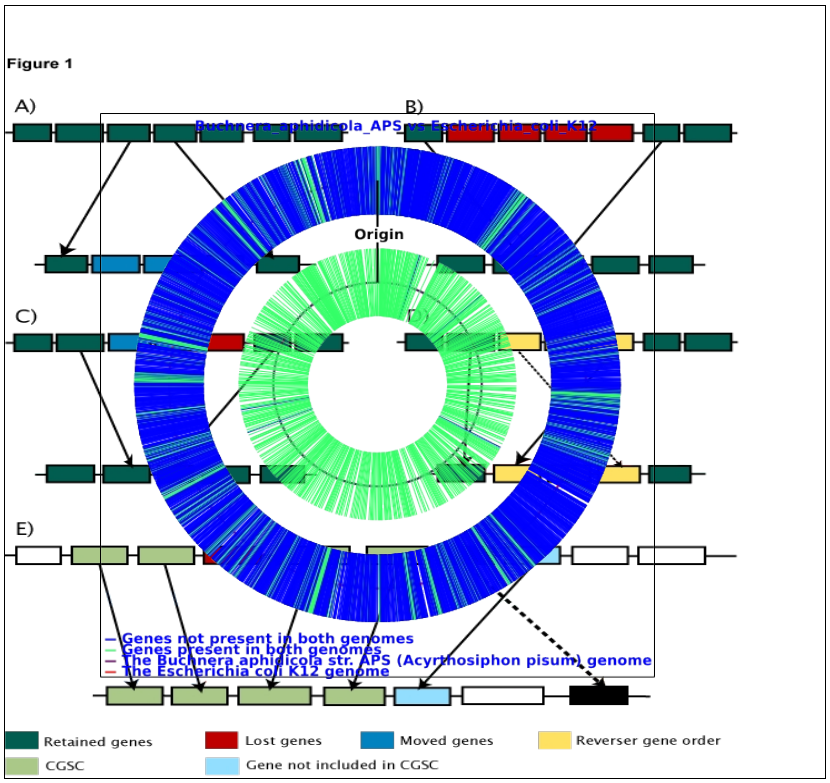


Figure 1. Gene rearrangements in the endosymbiont (reduced) genome identified by GRAST. **A)** neighbor genes in the ancestral genome translocated to different positions in the reduced genome; **B)** genes gathering as a result of disintegration of non-functional genes included between gathered genes; **C)** Genes gathering by translocation; **D)** Gene translocation and genome segment inversion; **E)** definition of Conserved Gene Succession Clusters (CGSCs). The light blue gene is not in the CGSC because of the black gene.

Description of the output options:

1) shows the placement of common/non-common genes in the two genomes (-g) default file name=[reduced genome's locus name]_[reference genome's locus name].genes.(gif/svg)

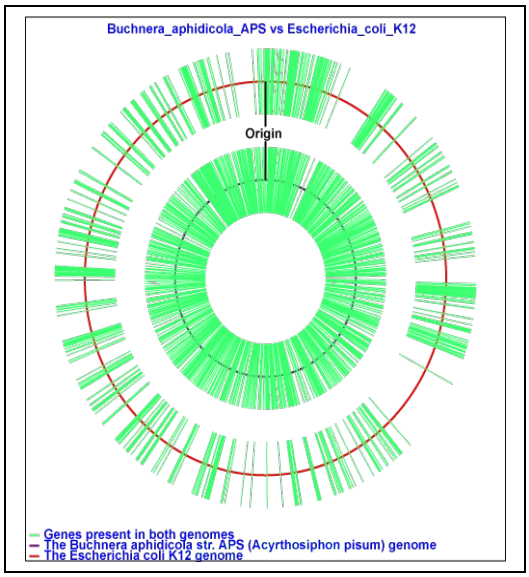
The genes the two genomes have in common are marked in green onto the two genomes at the start position of the genes. The genes that did not find a corresponding gene in the other genome are marked in blue onto the two genomes. The innermost genome is the reduced genome where the outermost genome in the reference genome. The origin marked on the two genomes is the origin of replication for the two genomes.
Note that the two genomes are NOT scaled.

output file:

2) shows the placement of common genes in the two genomes (-c)
default file name=[reduced genome's locus name]_[reference genome's locus name].common_genes.(gif/svg)

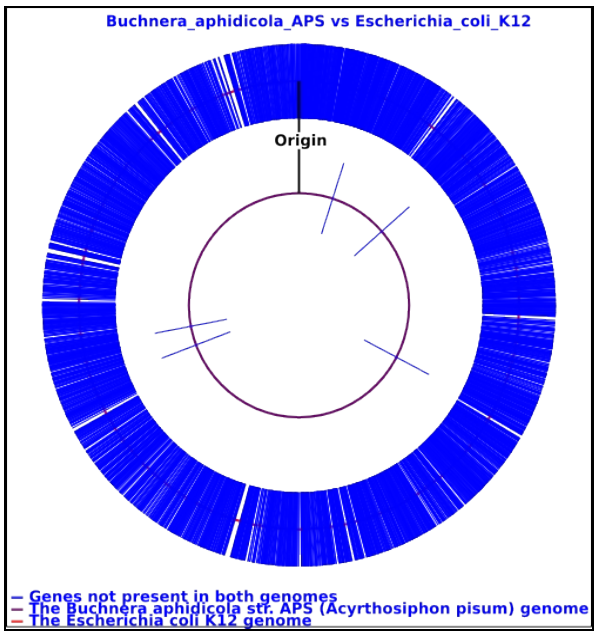
The genes the two genomes have in common are marked in green onto the two genomes at the start position of the genes. The innermost genome is the reduced genome where the outermost genome in the reference genome. The origin marked on the two genomes is the origin of replication for the two genomes.
Note that the two genomes are NOT scaled.

Output file:



3) shows the placement of none common genes in the two genomes (-n)
default file name=[reduced genome's locus name]_[reference genome's locus name].non-common_genes.(gif/svg)

The genes in the two genomes that did not find a corresponding gene in the other genome are marked in blue onto the two genomes. The innermost genome is the reduced genome where the outermost genome in the reference genome. The origin marked on the two genomes is the origin of replication for the two genomes.
Note that the two genomes are NOT scaled.

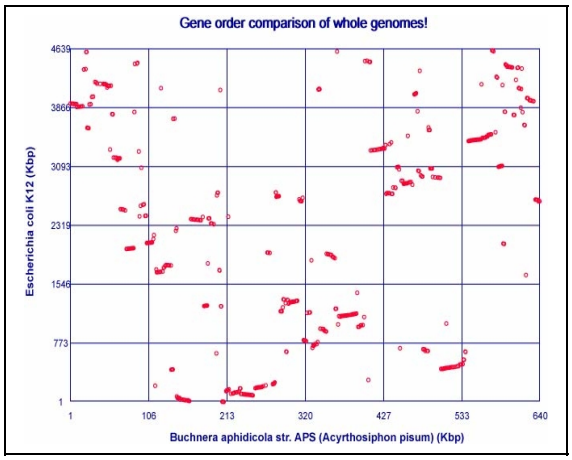


Output file:

4) gene pair plot of the two genomes (-r)
default file name=[reduced genome's locus name]_[reference genome's locus name].rearrange.(gif/svg)

The plot shows the placement of the orthologous gene pairs on the two genomes. The reduced genome is on the x-axis and the reference genome is on the y-axis.
If genes have been lost by random and no rearrangements have been take place, it would be expected that the gene pairs would be on the diagonal.

Output file:



5) CSGC: the start position in the two genomes, the genes names and the number of genes that has been lost (-s)
default file name=[reduced genome's locus name]_[reference genome's locus name].succession.txt

The first line in the output file: reference genome/reduced genome.
The rest of the file contains the CGSC found in the two genomes. The start positions in the two genome for each of the CGSC, the gene locus tag and the number of genes that been lost between adjacent genes in the CGSC. The CGSC's that have a reverse gene order in the reduced genome compared to the reference genome are at the end of the output file (after the line: REVERSE).

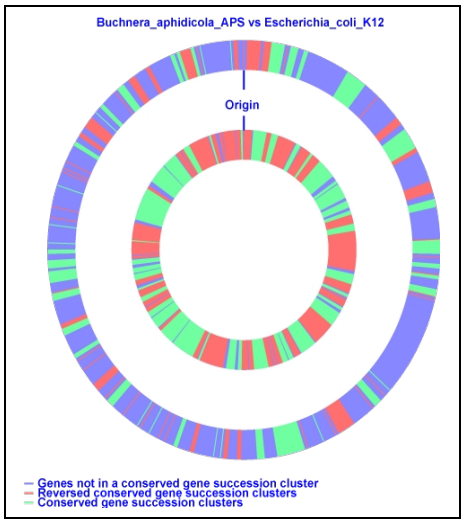
Output file:

```
Escherichia_coli_K12/Buchnera_aphidicola_APS
3875728/8911 : b3699/BU010 (1) b3701/BU011 (0) b3702/BU012 (0)
b3703/BU013 (0) b3704/BU014 (0) b3705/BU015 (0) b3706/BU016

4368711/18376 : b4142/BU018 (0) b4143/BU019 (3) b4147/BU020

4598261/21614 : b4361/BU021 (0) b4362/BU022
...

REVERSE

2654770/638074 : b2525/BU606 (0) b2526/BU605 (0) b2527/BU604 (1)
b2529/BU603 (0) b2530/BU602
...
```

6) shows the placement of the CGSC on the two genomes (-m)
default file name=[reduced genome's locus name]_[reference genome's locus name].succession.(gif/svg)

The CGSC are marked in green on the two genomes and red if the gene order has been reversed. Genes that are not in a CGSC are marked in blue. The innermost genome is the reduced genome where the outermost genome in the reference genome. The two genomes are not scaled. The origin marked on the two genomes is the origin of replication for the two genomes.

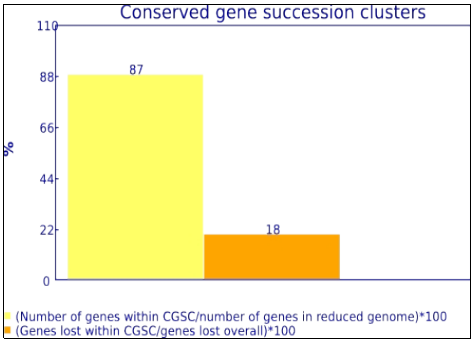
Output file:



7) overall statistics on CSGC's (-y)
default file name=[reduced genome's locus name]_[reference genome's locus name].cgsc_stats2.(gif/svg)

Bar graph showing the percentage of genes in the reduced genome that are contained in a CGSC and the percentage of gene lost within CGSC.

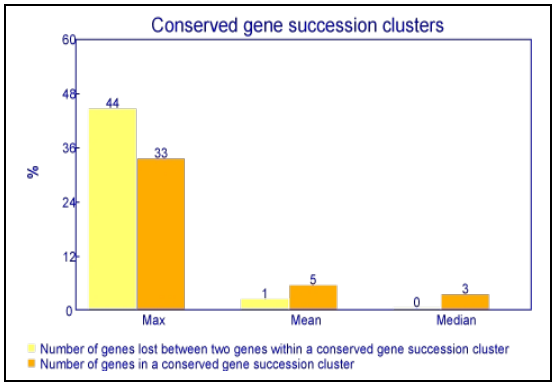
Output file:



8) statistics on CSGC (-x)
default file name=[reduced genome's locus name]_[reference genome's locus name].cgsc_stats1.(gif/svg)

Bar graph showing the maximum, mean and median number of genes lost between genes within a CGSC and number of gene contained within a CGSC.

Output file:



9) distribution of lost genes in COG - text file (-l)
default file name=[reduced genome's locus name]_[reference genome's locus name].dist.txt

The first line contains the number of protein coding genes (minis paralogous) in the reduced genome, the number of protein coding genes (minis paralogous) in the reference genome and the percentage of gene lost from the reference genome to the reduced genome.
The rest of the file contains statistical information for each of the functional categories defined by COG. The number of lost genes (from reference to reduced genome) over the number of genes present in the reference genome, the percentage of genes lost within a functional category and the percentage of lost genes in a functional category out of the total number of lost genes.

The information will indicated if the functional categories are more conserved than others. This can help in determining, which of the functional categories that is most and least imported for the survival of the organism.

Output file:

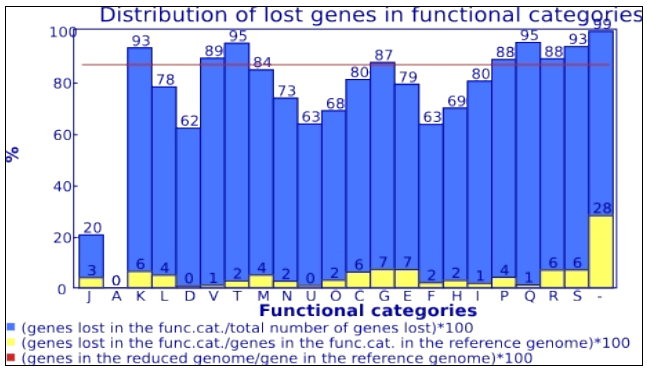
```
562     3886   85.538
```

COG	genes lost of	% g.l in cat	% g.l overall
J	30/146	20.548	0.903
A	0/1	0.000	0.000
K	219/235	93.191	6.588
L	142/182	78.022	4.272
D	18/29	62.069	0.542
V	33/37	89.189	0.993
T	95/100	95.000	2.858
M	155/183	84.699	4.663
N	70/95	73.684	2.106
U	21/33	63.636	0.632
O	75/109	68.807	2.256
C	183/226	80.973	5.505
G	232/265	87.547	6.980
E	208/263	79.087	6.258
F	47/74	63.514	1.414
H	74/106	69.811	2.226
I	53/66	80.303	1.594
P	134/151	88.742	4.031
Q	41/43	95.349	1.233
R	226/254	88.976	6.799
S	240/256	93.750	7.220
-	1028/1032	99.612	30.927

10) shows the distribution of lost genes within/in COG categories and the overall loss of genes (-b)
default file name=[reduced genome's locus name]_[reference genome's locus name].dist.(gif/svg)

Graphical representation of output option 8/-l.
A bar-graph where each of the bars represents a functional category is produced in this option. The blue bars are the percentage of genes lost within a functional category, the yellow bars are the percentage of lost genes in a functional category out of the total number of lost genes and the red line is the expected lost of genes (the percentage of lost genes from the reference genome to become the reduced genome).

Output file:



11) orthologous gene pairs -gene names + e-values (-p)
default file name=[reduced genome's locus name]_[reference genome's locus name].common_genes.txt

The first column contains the locus name and the gene names of the identified orthologous gene pairs in the reduced genome. The fourth and last column contains the locus name and the gene names of the identified orthologous genes pairs in the reference genome. Column two and three contains the e-values from the BLAST searches. The second column is the e-values computed from the BLAST search of the reference genome against the reduced genome and the third column is the e-values computed from the BLAST search of the reduced genome against the reference genome. Each row except the first contains an orthologous gene pair with the corresponding e-values.

Output file:

```
NC_002528          e-value   e-value   NC_000913
BU001   0.0        0.0       b3741
BU002   1e-101     1e-101    b3738
BU003   6e-25      4e-21     b3737
BU004   5e-31      4e-30     b3736
BU005   3e-29      5e-30     b3735
BU006   0.0        0.0       b3734
BU007   4e-99      1e-95     b3733
BU008   0.0        0.0       b3732
BU009   2e-34      2e-30     b3731
BU010   0.0        0.0       b3699
BU011   4e-81      7e-86     b3701
                     ...
```

12) list of genes that dont have a homolog in the other genome (-k)
default file name=[reduced genome's locus name]_[reference genome's locus name].non-common_genes.txt

The first line contains the name and locus name of the reduced genome. The following lines contain the locus tag of the genes in the reduced genomes that are not found in the reference genome by mutual top hit BLAST search. These lines are followed by a blank line and a line contains the name and locus name of the reference genome. The remaining lines in the file then contain the locus tag of the genes from the reference that are not found in the reduced genome. All the genes that have been determined to have a paralog in the common genes have 'paralog to locus tag of the gene' following it in the output file.

Output file:

```
Buchnera aphidicola str. APS (Acyrthosiphon pisum) (NC_002528)
BU029
BU078
BU181
BU252   paralog to BU184
BU380
...

Escherichia coli K12 (NC_000913)
b0001
b0005
b0006
b0007
b0008   paralog to b2464
b0009
...
```

13) observed and expected (before and after genome reduction) on seeing gathering of genes with similar function (-t)
default file name=[reduced genome's locus name]_[reference genome's locus name].translocations.txt

It is assumed that each of the translocation events is independent and that they have not occurred in any specific order. The probability of seeing genes being gathered caused by movement to genes with similar function can be calculated by using multinomial density function.
The probability of gathering (caused by movement) genes with similar function is calculated for the observed and the expected. The expected values are found by simulating the same number of translocation events in the reference genome and in a reduced reference genome (gene order of the reference genome but contains only common genes found between the two genomes) as observed and determining the probability of seeing these events. By default there are done 100 simulations for the reference genome and for the reduced reference genome. The expected values are the average of the simulations.

Output file:

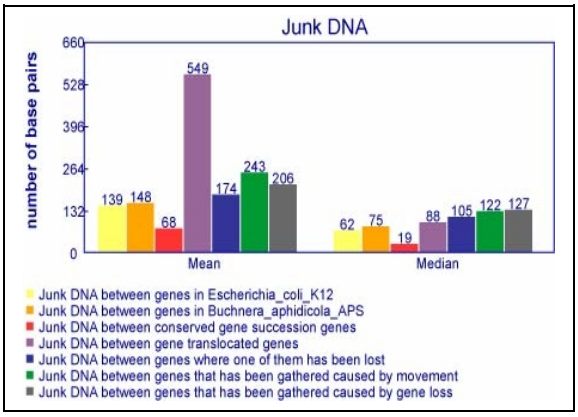
```
Gathering caused by movement probability:
        - observed                      = 1.5486e-12
        - random before genome reduction  = 1.9185e-03
        - random after genome reduction   = 7.6295e-07
```

```
Note that 97.34% of the reduced genome has a corresponding gene in
the reference genome
```

14) shows the distribution of junk DNA (-j)
default file name=[reduced genome's locus name]_[reference genome's locus name].junkDNA.(gif/svg)

Bar graph showing the mean and median length of junk DNA between genes in the reference genome, the reduced genome, between conserved gene succession genes, between translocated genes, between genes where one of them has been lost, between genes that have been gathered caused by movement and between genes that has been gathered caused by gene loss.

Output file:



15) distribution of junk DNA - text file (-h)
default file name=[reduced genome's locus name]_[reference genome's locus name].junkDNA.txt

File containing the number of genes, the maximum, the mean and median length of junk DNA between genes in the reference genome, the reduced genome, between conserved gene succession genes, between translocated genes, between genes where one of them has been lost, between genes that has been gathered caused by movement and between genes that has been gathered caused by gene loss. Last line tells how many genes that is in the reduced genome that is not found in the reference genome.

Output file:

```
Number-       Max-      Mean- Median- Group
   563-      3813-    148.34-     75- Buchnera_aphidicola_APS
  4236-      9159-    139.41-   62.5- Escherichia_coli_K12
```

```
   238-       606-     68.63-   19.5- Length of junk DNA between
conserved gene succession genes
    13-      5771-    549.54-     88- Length of junk DNA between
gene translocated genes
   596-      5701-    174.11-    105- Length of junk DNA between
genes where one of them has been lost
   147-      3813-    243.24-    122- Length of junk DNA between
genes that has been gathered caused by movement
   150-      1365-    206.52-    127- Length of junk DNA between
genes that has been gathered caused by gene loss
    15 Number of genes present in the reduced genome but not found
in the reference genome
```