Supplementary material

1. DomainTeam : the case of inserted domains

In those few cases where a domain is inserted within another domain (Bateman et al., 2004), the two domains are considered as adjacent. The Figure below illustrates such a case.

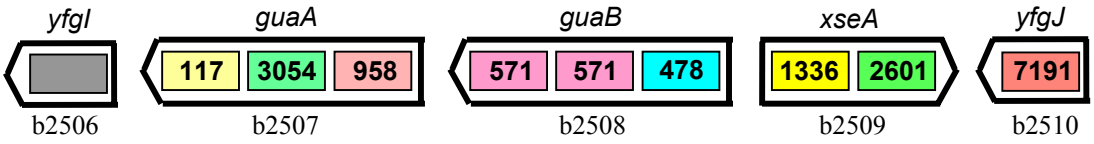

Figure Legend

Schematic representation of five successive genes on the chromosome of *E. coli*. Genes are represented by thick arrows and labeled both by their name above the arrows and by their "ordered locus name" under the arrows. Within the arrows are given the PfamA domain numbers found in the corresponding encoded proteins, where 571 stands for PF00571. The protein encoded by *yfgI* does not contain a Pfam domain. In GuaB, the two domains PF00571 are inserted within the domain PF00478. They are considered by DomainTeam as being outside this last domain. The two genes *xseA* and *yfgJ* are partially "end-on" (Fukuda et al., 1999) overlapping genes (4 bases).

2. Pathogenic island : type III secretion system

This exemple depicts a team found in a set of 10 pathogenic bacteria, comprising genes coding for proteins belonging to the Type III secretion system and for flagellar switch proteins (Hueck, 1998; Pallen et al., 2003). Here we can see that DomainTeam detects conserved inter-chromosomal segments as well as intra-chromosomal duplications. It may be noted that one gene in *P. aeruginosa* (between *pscT* and *pscR*) has no name in EMBL/Uniprot. Indeed, all the Blast2 comparisons with all the other proteins in the team sharing the same domain (PF01313) resulted in high E-values ($> 2.10^{-3}$) except with YscS from the *Y. pestis* plasmid. The conserved position of this gene in a highly conserved team suggests that it could be safely named *pscS*.

Set of 10 pathogenic bacteria (chromosomes and plasmids if any): *Chlamydia muridarum, Escherichia coli O157-H7, Pseudomonas aeruginosa, Pseudomonas syringae, Ralstonia solanacearum, Rhizobium meliloti, Salmonella typhimurium, Shigella flexneri, Xantomonas campestris, Yersinia pestis* plus *Bacillus subtilis*.

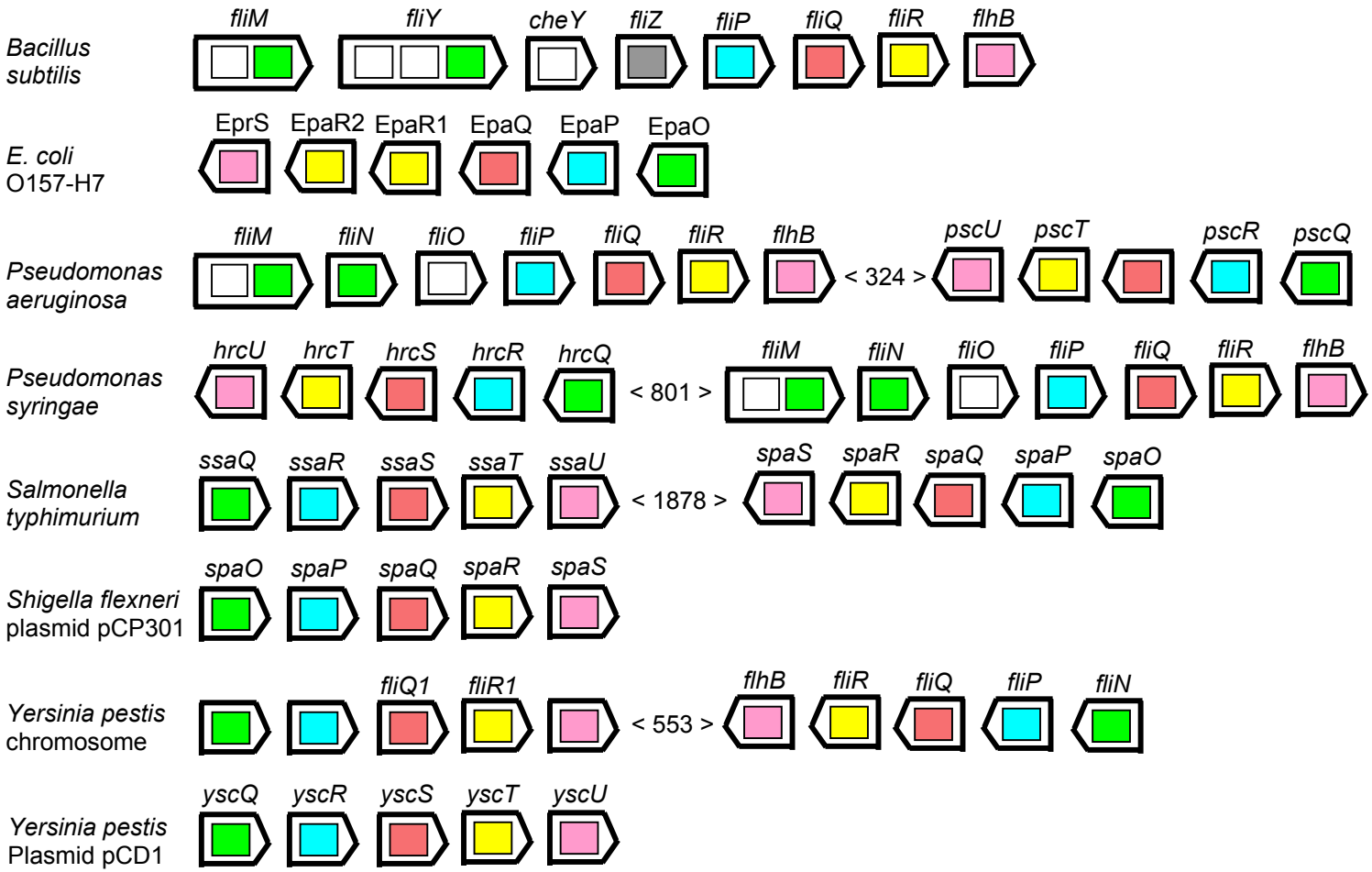
Bacillus subtilis
fliM
fliY
cheY
fliZ
fliP
fliQ
fliR
flhB
E. coli O157-H7
EprS
EpaR2
EpaR1
EpaQ
EpaP
EpaO
Pseudomonas aeruginosa
fliM
fliN
fliO
fliP
fliQ
fliR
flhB
< 324 >
pscU
pscT
pscR
pscQ
Pseudomonas syringae
hrcU
hrcT
hrcS
hrcR
hrcQ
< 801 >
fliM
fliN
fliO
fliP
fliQ
fliR
flhB
Salmonella typhimurium
ssaQ
ssaR
ssaS
ssaT
ssaU
< 1878 >
spaS
spaR
spaQ
spaP
spaO
Shigella flexneri plasmid pCP301
spaO
spaP
spaQ
spaR
spaS
Yersinia pestis chromosome
fliQ1
fliR1
< 553 >
flhB
fliR
fliQ
fliP
fliN
Yersinia pestis Plasmid pCD1
yscQ
yscR
yscS
yscT
yscU

Figure Legend

An example of a team ($\delta = 4$) found in 10 pathogenic bacteria (plus *B. subtilis*). Each color corresponds to one particular Pfam domain (their labels have been omitted due to space constraints). White domains are not members of the team. Protein FliZ in *B. subtilis* has no Pfam domain and is considered as an insertion. The gene names (or protein names in the case of *E. coli*) are given above their schematic representation. A figure between two occurrences of the team on the same chromosome represents the number of domains between the two occurrences.

3. List of the fully recovered operons and their phylogenetic distribution

* Operon dnaKJ *

Size : 2
B0014
B0015
2/2 gene(s) found in domain teams number : 862 1615 3887 3893 11917 14910
BEST domain team 3893 : anaba borbu ecoli haein pseae rhilo salty vibch1 xylfa yerpe
PHYLO : cyanobacteria spirochaetes gammaproteobacteria alphaproteobacteria

- Operon ribF-ileS-lspA-slpA-lytB *

Size : 5
B0025
B0026
B0027
B0028
B0029
5/5 gene(s) found in domain teams number : 1615 7988 10998 15980
BEST domain team 7988 : ecoli pseae salty vibch1 yerpe
PHYLO : gammaproteobacteria

* Operon carAB *

Size : 2
B0032
B0033
2/2 gene(s) found in domain teams number : 1615 4926 10652 10998 12955
BEST domain team 12955 : bactn ecoli pseae salty thema vibch1 xylfa yerpe
PHYLO : bacteroidetes gammaproteobacteria thermotogae

* Operon caiTABCDE *
Size : 6
B0035
B0036
B0037
B0038
B0039
B0040
6/6 gene(s) found in domain teams number : 1615 10652
BEST domain team 10652 : ecoli salty
PHYLO : gammaproteobacteria

Etc. For the full list see : http://lgi.infobiogen.fr/DomainTeams

4. Average number of proteins per occurrence of domain teams and score

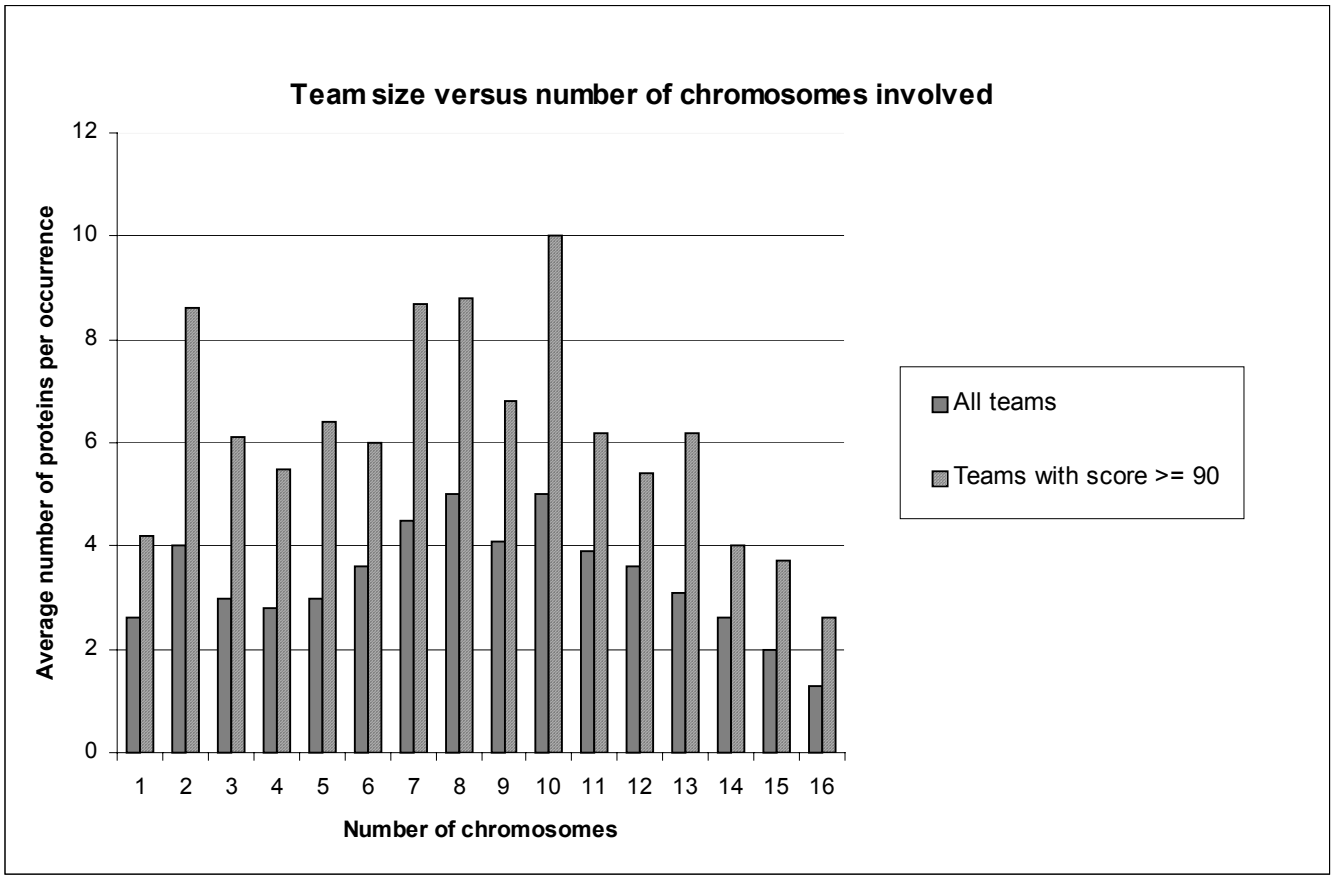

Figure Legend

Distribution of the average number of proteins per occurrence of the teams ($\delta = 3$) found in 16 archaebacterial chromosomes, as a function of the number of chromosomes they belong to.

Those teams with a score ≥ 90 contain a higher number of proteins.

5. Practical computing considerations

The computation time required to compare a set of chromosomes is a function of the number of chromosomes, the number of proteins in the set, the value of δ and the degree of conservation between the organisms under study. We tested the efficiency of DomainTeams on a 1 Ghz Sun ultrasparc III+ processor in the following way : a) we downloaded a superset of 95 eubacterial and archaebacterial chromosome tables from EBI; b) for *n* ranging from 4 to 20 (see Table a) we extracted for each *n* 25 subsets of *n* chromosomes chosen at random from the superset; c) for each *n* we run DomainTeams on each of the 25 subsets of *n* chromosomes with $\delta = 3$. The results (mean values for the 25 runs) are given in Table a. As stated in the DomainTeam section, the number of teams can become exponential when the chromosomes under study are very similar, which explains the high standard deviations of the CPU times (Table a) : they were significantly higher when the subsets contained closely related species such as *E. coli* and *Y. pestis*, even more with two strains of the same species such as *E. coli* K12 and O6. Not surprisingly, the CPU time increases with both the number of chromosomes and δ. Thus the value of δ must be kept small (*e.g.* 3) to compare simultaneously tens of chromosomes at the same time. Note, however, that the comparison of the two closely related yeasts *Saccharomyces cerevisiae* (6 222 proteins) and *Ashbya gossypii* (4 712 proteins) (Dietrich et al., 2004) was performed in 3 seconds with $\delta = 5$. The comparison with $\delta = 3$ took 5 minutes for the set of 16 Archaebacteria, 320 minutes for the set of 15 Gram- bacteria and 29 minutes for the set of 13 Gram+ bacteria. Thus, whatever the closeness of the species and the value of δ, DomainTeam is extremely efficient when comparing a small number of chromosomes.

proteome set	4 bacteria	8 bacteria	16 bacteria	18 bacteria	20 bacteria
nb of proteins	10 638	22 402	48 257	52 401	58 323
nb of domains	13 211	27 392	57 609	63 740	71 381
nb of teams	2 099	4 228	8 821	11 169	12 220
nb of fusions	69	170	338	447	502
time user CPU (s) (standard deviation)	294 (983)	1 199 (2 769)	29 809 (50 439)	59 146 (53 810)	49 763 (52 782)

Table a : Average parameters and computation times for different sets of complete proteomes with $\delta = 3$. In each column the values are the means of 25 different runs.

value of δ	3	4	5	6	7	8	9	10
CPU time (s)	159	230	381	757	1 522	1 615	3 617	9 010

Table b : Computation times for processing one set of four bacterial proteomes with different values of δ. Note that a value of $\delta = 3$ corresponds to a maximum gap length of 2 domains.