Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion

Julien Guglielmini¹,²*, Bertrand Néron³, Sophie S. Abby¹,², María Pilar Garcillán-Barcia⁴, Fernando de la Cruz⁴ and Eduardo P. C. Rocha¹,²

¹Microbial Evolutionary Genomics, Institut Pasteur, Paris 75015, France, ²UMR3525, CNRS, Paris 75015, France, ³Centre d’Informatique pour les Biologistes, Institut Pasteur, Paris 75015, France and ⁴Institute of Biomedicine and Biotechnology of Cantabria (CSIC-Sodercan-University of Cantabria), Santander 39011, Spain

Received December 5, 2013; Revised February 06, 2014; Accepted February 23, 2014

ABSTRACT

Conjugation of DNA through a type IV secretion system (T4SS) drives horizontal gene transfer. Yet little is known on the diversity of these nanomachines. We previously found that T4SS can be divided in eight classes based on the phylogeny of the only ubiquitous protein of T4SS (VirB4). Here, we use an ab initio approach to identify protein families systematically and specifically associated with VirB4 in each class. We built profiles for these proteins and used them to scan 2262 genomes for the presence of T4SS. Our analysis led to the identification of thousands of occurrences of 116 protein families for a total of 1623 T4SS. Importantly, we could identify almost always in our profiles the essential genes of well-studied T4SS. This allowed us to build a database with the largest number of T4SS described to date. Using profile–profile alignments, we reveal many new cases of homology between components of distant classes of T4SS. We mapped these similarities on the T4SS phylogenetic tree and thus obtained the patterns of acquisition and loss of these protein families in the history of T4SS. The identification of the key VirB4-associated proteins paves the way toward experimental analysis of poorly characterized T4SS classes.

INTRODUCTION

Prokaryotes have the ability to adapt quickly by acquiring genes from other prokaryotes (1–3). Conjugation, which is one of the major mechanisms of gene transfer, requires cell-to-cell contact and is able to deliver the whole genome of one cell into another. Conjugation-specific proteins are found in all major taxa of prokaryotes, even though experimental evidence is still mostly restricted to Proteobacteria and Firmicutes (4–8). The most frequent mechanism of DNA conjugation involves the passage of single-stranded DNA (ssDNA) from the donor cell to the recipient, upon which replication re-establishes double-stranded (dsDNA) copies in each cell (7). This mechanism relies on three major components: a relaxosome, a coupling protein (T4CP) and a type IV secretion system (T4SS). The relaxosome includes a protein essential for conjugation—the relaxase (MOB)—that nicks the dsDNA and binds the resulting ssDNA at the origin of transfer (see (7, 9) for reviews). The relaxase bound to the ssDNA molecule is coupled to a T4SS by a T4CP and translocated through the donor membrane(s) to the cytoplasm of the recipient. Two different coupling proteins have been identified: VirD4 and TcpA. TcpA is found within certain systems of Firmicutes, and is more closely related to FtsK, a protein involved in chromosome segregation, than to VirD4 (10, 11). VirD4 is associated with the vast majority of T4SS and probably originated from an ssDNA translocase (12). Some mobile genetic elements encode a relaxase and occasionally a T4CP but no T4SS. These elements are very abundant in bacterial genomes and are called ‘mobilizable’ because they use a T4SS encoded in trans. Most T4SS are thought to be involved in conjugation (nucleo-protein secretion), but some are specialized in protein secretion, allowing the delivery of effector proteins to the cytosol of eukaryotic organisms. These T4SS typically lack a relaxase (MOBless T4SS), but require a T4CP (see (5) for exceptions). Several systems are able to deliver both the DNA-bound relaxase and other protein effectors (13–15). There are also examples of T4SS involved in DNA import (the ComB system of Helicobacter pylori (16)) or in DNA secretion (the GGI system of Neisseria gonorrhoeae (17)) associated with natural transformation. T4SS are thus remarkably diverse...
The plasticity of T4SS results in a diversity of systems, most of which are yet poorly characterized. One complication in the study of T4SS is the lack of a standard nomenclature; genes of similar names are not necessarily homologs and homologs do not necessarily have the same names. Here, we follow the convention that mating-pair formation (MPF) genes are labeled with the name of the associated mobile element, and protein profiles are labeled with the name of the MPF class. For example, TraB is the protein TraB encoded by plasmid F and TraB_{MPFF} is the protein profile for the protein family including TraB of MPF_F. The only exception concerns the VirB proteins, which by default concern the T-DNA transfer VirB system of Agrobacterium tumefaciens, and whose names are in general used without ambiguity (i.e. similar names correspond to homologs). We use the VirB system, composed of 11 genes from VirB1 to VirB11, as a model because it is by far the best characterized T4SS (9,18) (Figure 1). The core secretion channel complex of the VirB T4SS (including VirB7, VirB9 and VirB10) that spans the periplasm and both cell membranes is thought to be the first to assemble. VirB10 lines the inner surface of the core complex chambers, and spans the whole length of the core complex, forming the bottom ring of the inner membrane layer. The protein VirB9 forms the outer sheath of the core complex that interacts and is stabilized by VirB7, a small lipoprotein (9,19). The three proteins VirB3, VirB6 and VirB8 are thought to join the core complex to produce the inner-membrane pore (20). VirB6 is a polytopic membrane protein with a number of transmembrane domains (TMDs) and a large central periplasmic loop (21). The role of VirB3 is not yet clear, even if the protein is essential for pilus assembly and substrate translocation. In some systems, VirB3 is fused to VirB4, which suggests a strong functional link between the two (22). Three AAA+ ATPases (VirB4, VirB11 and the T4CP) join the assembled complex, one of which (VirB4) is implicated in energizing pilus biogenesis (23). The pilus is composed of a major and a minor pilin (respectively VirB2 and VirB5). Finally, VirB1 is a non-essential transglycosylase that degrades peptidoglycan and thus facilitates T4SS assembly across the cell wall (24,25).

In a recent study, we found that the phylogeny of VirB4, the only ubiquitous protein with recognizable homologs in all known T4SS, is divided in eight large robust clades that correspond to eight MPF classes (11). Based on these results we proposed an evolution-aware classification of T4SS that shows strong associations with prokaryote systems and, to some extent, to the structure of the cell envelope (11). The association between T4SS composition and cell envelope structure has been recently reviewed (20). It should be stressed that within a given MPF class there is also co-evolution between the composition of the membrane and the T4SS (26). Four MPF classes encompass the conjugation systems of Proteobacteria and closely related taxa: VirB-like systems are the most numerous (MPF_F); F-like systems are particularly abundant in plasmids of γ-Proteobacteria (MPF_G); R64-like systems are much rarer (MPF_H); and ICEHin1056-like systems are almost exclusively found as integrative elements (ICE) in Proteobacteria (MPF_G) (6,8,27). Two MPF classes are much less well known and include systems present only in Cyanobacteria (MPF_C) or in Bacteroidetes (MPF_J). Finally, we delimited two different MPF classes in monoderms (organisms devoid of an outer membrane). One class is found in Firmicutes and Actinobacteria (MPF_K) whereas the other is also found in Tenericutes and Archaea (MPF_FATA).

T4SS are very diverse: 10 of 11 VirB genes are essential (24), whereas the conjugative system of R64 is encoded by 49 genes of which 23 are essential for solid mating plus 12 for liquid mating (28). The exact number of genes and their essentiality for T4SS assembly or function is unknown in many classes of T4SS. Some systems of Firmicutes are thought to lack pili altogether (4), e.g. the T4SS of the pCF10 plasmid of Enterococcus faecalis, the pG01 plasmid or the ICE TnGBS of Streptococcus encode adhesins that stabilize the mating process. In taxa such as Archaea, Actinobacteria or Cyanobacteria, there are very few reports on the mechanisms of ssDNA conjugation, and none, to the best of our knowledge, on the T4SS structure and composition. Yet these taxa encode homologs of VirB4 and VirD4 in their mobile genetic elements suggesting the presence of T4SS-mediated conjugation (8).

Here, we took advantage of our previous dataset of VirB4 homologs to characterize all MPF classes. The repertoires, diversity and evolution of relaxases have been recently reported (29). Therefore, we focused our attention on T4SS. To establish the repertoire of protein families typical of each class of T4SS, we detected the families of genes systematically associated and co-localized with VirB4 in the eight MPF classes. With these families we built protein profiles that we used to detect and classify T4SS. This resulted in a much more complete database of T4SS than those currently available. Using profile–profile alignments we identified distant homologies between protein families of different MPF classes thereby providing the first large-scale systematic analysis of protein homology between all different MPF classes. Finally, we built a web resource, called CONJdb, that presents all our data in a searchable and comprehensive manner.

MATERIALS AND METHODS

Data

Data on complete prokaryotic chromosomes and plasmids were taken from GenBank Refseq (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/, last accessed February 2013). The dataset of 2262 complete prokaryotic genomes comprised 2393 chromosomes and 1813 plasmids. We used the annotations of the GenBank files, having removed all pseudogenes and proteins with inner stop codons. Some protein profiles were taken and improved from our previous work (8).

Construction of protein profiles

Figure 1 describes the procedure used to construct the protein profiles. We searched for genes encoding TraU/VirB4 in the genomes using HMMER (see below). We then gathered the 20 genes on each side of each traU/virB4 gene. This resulted in seven sets of proteins, one for each T4SS class (except for the previously analyzed MPF_F). We made all-against-all BLASTP searches in each set (default settings)
Figure 1. Procedure used to create mating-pair formation (MPF) protein profiles. We used all the genes found within a frame of $-20$ to $+20$ genes around the VirB4 proteins of each clade (named MPF$_1$, MPF$_C$, MPF$_G$, MPF$_T$, MPF$_B$, MPF$_F$, MPF$_FATA$, and MPF$_FA$—see text for details) of its phylogeny (11). Numbers on the phylogeny correspond to bootstrap values, from (11). With these seven datasets of proteins (one for each class, except for MPF$_T$ for which we already had the protein profiles) we performed all-versus-all BLASTP and used the scores to build protein families. We made protein alignments of these families, and kept the ones that give hits within the class. Thumbnail: scheme of the virB system.

and used the output to build protein clusters using SiLiX (30) (identity $\geq 30\%$ and overlap $>50\%$). We carried out a multiple alignment of the proteins in each cluster with more than five proteins using MUSCLE (31) (default parameters). We used the multiple alignments to build phylogenetic trees using PHYML (32). With these two pieces of evidence we removed the very few cases of extreme divergence, the proteins that were too short and the proteins that were too long (typically false positives, fusions or fissions of proteins motivated by sequencing errors or pseudogenization). Then, we re-built multiple alignments of the selected proteins with MUSCLE, checked manually the alignments and trimmed them to remove poorly aligned regions at the edges, if relevant. Finally, we used HMMER 3.0 (33) to build protein profiles from the manually curated multiple alignments.

It should be pointed out that homologous proteins with very little sequence similarity might escape the BLASTP detection used in the clustering procedure if a protein does not find a single hit in a family. However, the use of a method allowing detection of these homologs would probably not help at this stage, since the protein profiles must be built from proteins providing reliable multiple alignments. For a few very divergent proteins this may lead to their exclusion from the analysis. For families aligning poorly with other sub-families, our method will provide several independent profiles. In this case, the profile–profile alignment procedure will identify the homology between families.

Identification and analysis of MPF protein profiles

To identify the profiles corresponding to MPF proteins within all the protein profiles, we performed hidden Markov model (HMM) searches on the genome data. We used HMMER to identify components of the T4SS. We kept the hits showing an i-e-value $< 0.01$ and a coverage of the protein profile higher than $50\%$. Some T4SS genes may be more than 20 genes apart from the VirB4 homolog (even if we could not find a single such occurrence in the model systems). The proteins encoded by these genes will not be used to build the protein profiles. Yet these proteins will be identified by these profiles when we scan the genomes. We will keep them for further analysis when more than $50\%$ of the hits of the protein family are located in a neighborhood of $-20$ to $+20$ genes around a truU/virB4 gene, i.e. as long as the elements of the protein family are not systematically distant from VirB4. All genes regarded as essential components of the T4SS were in this situation. Then, we applied the following procedure on the remaining HMM profiles. We compared all pairs of profiles with each other and with the ones of the Pfam 26 database (13 672 protein families) using HH-search (34) ($p < 0.001$ threshold) (35). As common usage
in comparative genomics, the pairs of significant hits were regarded as homologs. We cannot totally exclude the possibility that the very low similarity between some families could be due to convergence. We removed from further analysis the profiles whose Pfam annotation was clearly related to functions other than conjugation. We inspected the colocalization patterns of the hits of the remaining profiles. We removed protein profiles that gave hits systematically distant (more than five ORFs) from the others. The resulting set of profiles was used to scan the genomes and build CON- Jdb.

**RESULTS**

**Characterization of T4SS protein families**

We built a procedure to identify T4SS based on two well-established features (5). (i) Genes encoding the T4SS are generally grouped together in one or a few operons. (ii) VirB4 is the only protein family identified in all functional T4SS. We identified the occurrences of virB4 in 2269 complete prokaryotic genomes using a previously defined protein profile (see the Materials and Methods section). These proteins were found in all major taxa. We then fetched the genes in the genomic neighborhood of each of the 1623 virB4 genes. Pairwise similarity searches followed by clustering and curation resulted in 652 families. The multiple alignments of each family were used to build protein profiles (HMMs), which were applied to scan the genomic data. Some T4SS components may be more than 20 genes apart from virB4. This is the case of T4SS encoded in several loci scattered in the genomes of *Rickettsiales* (36). These proteins were not used to build the protein profiles. Nevertheless, they were subsequently identified in the step of genome scanning with the protein profiles. They are therefore included in the analysis of T4SS. Our method has no phylogenetic bias, i.e. we do not *a priori* restrict protein families associated with a VirB4 class to a given taxonomic group. However, most studied T4SS are from Proteobacteria and Firmicutes and this may lead to two inevitable implicit biases in the analysis. Firstly, analogous, not homologous, proteins may fill the same function in different taxa and might be missed if there are few representatives of the taxa or if these genes are not encoded systematically close to virB4. For example, we have reported that we probably miss relaxases from Archaea and Actinobacteria (11). Secondly, proteins evolving too rapidly produce smaller protein families that might miss representatives from the T4SS more evolutionarily distant from the model systems. Candidates for such functions can be fetched using other protein features like peptide signals or TMDs (see below). In spite of this, we found T4SS in nearly all taxa for which there is a significant number of genomes. Profile–profile comparisons also showed homology between many of these profiles between MPF classes and between taxa (see below). Therefore, we believe to have uncovered the majority of protein families systematically associated with T4SS.

Some of the new profiles are not specific for T4SS because they match genes lacking a neighboring virB4 more than 50% of the times. Most of these non-specific profiles match proteins typically encoded by mobile genetic elements, like primases or zinc-finger proteins, which are not directly associated with conjugation or T4SS (Supplementary Table S1). These profiles may help the characterization of the genetic context of the T4SS but they are of little use to the computational study of these systems. In our genome scans we ignored them, except when experimental data showed their implication in conjugation. This was the case of relaxases and T4CP. Some profiles match proteins encoded by genes systematically neighboring virB4. These profiles are specific to T4SS, i.e. they contribute to their accurate identification. In the vast majority of cases, they also match proteins from one single T4SS class (Figure 2) and can therefore be used to class T4SS. Out of the 1623 VirB4 hits, 93% co-localized with other T4SS-associated profiles in Proteobacteria and 80% in other taxa. Hence, in the vast majority of cases, VirB4 is indeed associated with T4SS. Exceptions may be due to ongoing genetic degradation of conjugation systems, to the existence of unknown classes of T4SS and/or to cooperation of VirB4 for other functions. A T4SS is assigned to a given class if the direct neighborhood of VirB4 harbors the hits of at least three protein profiles from this class. Some systems cannot be classified this way because they contain less than three hits, or no hits at all, of MPF protein profiles. We call them MPFO (‘O’ for ‘others’). They are mostly (75%) T4SS loci lacking a neighboring relaxase, thus probably devoted to protein secretion or ongoing genetic degradation.

The observation that profiles typically match T4SS of one single class means that homologous proteins are more similar within than between T4SS classes. This can be explained either by lack of homology between components of different T4SS classes or by tight co-evolution between these proteins and VirB4. The latter hypothesis applies to at least certain components that have known homologs between T4SS classes (e.g. VirB6 or VirB1) (20,37). Hence, homologous components exist but are not usually exchanged between classes. We used profile–profile comparisons to systematically detect significant sequence similarity between protein families of different T4SS classes (34). These protein families are similar in sequence and thus they are likely to correspond to bona fide homologs. Nevertheless, since divergence between the proteins precludes the use of phyloge-
metic methods to test homology, we cannot exclude the possibility of convergent evolution. Our analysis pinpointed 57 relations of similarity between protein families associated with different classes of T4SS (Figure 3), in addition to the relations between VirB4, T4CP and MOB homologs. In the following sections, we describe the protein families identified in the different T4SS (Figure 4). We then compare these among themselves and with our previously defined set of profiles for the VirB system (11).

**MPF**

We used the F plasmid T4SS as the model of MPF. This system is composed of 18 proteins, some homologs for VirB components (reviewed in (38)). We built profiles for the three components of the core complex of the T4SS—TraB, TraV and TraK—that display structural similarities to the VirB10, VirB9 and VirB7 complex (40). While these proteins do seem to have analogous roles, only the profile–profile alignments of VirB9 and TraKMPFF and TraVMPFF were significant (P-values of 7.6 × 10^{-6} and 8.3 × 10^{-5} respectively). Profile–profile comparisons of VirB10 and TraBMPFF are not significant, whereas they both exhibit the same PFAM domain (PF03743, with e-values of 4.8 × 10^{-6} and 1.3 × 10^{-15} respectively). This is because our long profiles (respectively 375 and 431 positions) only align at the region common to the much smaller profile PF03743 (187 positions). The inner-membrane pore is thought to be composed of TraGF in interaction with TraLF and TraEF (homolog to VirB3 and VirB8 respectively). The N-terminal region of TraGF is homolog to VirB6 while the C-terminal part is involved in mating-pair stabilization (41). TraNF interacts with OmpA and LPS moieties during conjugation, resulting in mating-pair stabilization (41,42). Its profile did not match any other profile. TraA is the only pilin of MPF (43,44), but its rapid evolution precluded the definition of a protein profile for the family. We built four different profiles for the periplasmic proteins TraW, TraU, TraF and TrbcF (45), all homologous according to our profile–profile alignments (Figure 3). The inactivation of these proteins leads to shortened pili (46,47). We also built a profile for TraH that is thought to participate in pilus extension (44,45). Overall, we have obtained 12 protein profiles for MPF matching all known essential genes, except TraQ (pilin chaperone) and the fast-evolving TraX (50). Interestingly, these three latter genes interact physically (51,52).

**MPF**

We used the IncI plasmid R64 conjugative system as the model for MPF. We built 16 protein profiles for MPF, few of which have homologs in other classes. The profile TraOMFF is homologous to VirB10 and thus probably part of the core complex (53). TraSMFF is homologous to TraEMFF/VirB8, and thus is probably part of the inner scaffold. TraJMPF is homologous to VirB11 (BlastP e-value <10^{-14}), as reported (54). TraOMFF and TraRMPF are homologs and profile–profile comparisons show an indirect relation of homology with the pilin VirB2 (Figure 3). We also built profiles for proteins required for conjugation both in liquid and in surfaces that are encoded outside of the main T4SS operon: TrbAMFF and TrbBMFF (28) and for TraEMPFF whose function is still unknown (55). TrbMMFF is homolog to TraFMPFF as previously suggested (44). We could not obtain profiles for TraRMPF, TraHRMPF, TraSRMPF and TraXRMPF that are not essential for conjugation (28,55). We also could not build a specific protein profile for the SogLMPF and SogSRMPF proteins; they have been reported to be essential for conjugation, but not directly involved in DNA transfer (55). Except these ones, the set of profiles for MPF includes all proteins for which the corresponding gene disruption completely abolishes transfer (55).

**MPF**

MPF were originally described from ICEIn1056 from *Haemophilus influenzae* (27,56). Here, we used the very closely related ICEIn10810 element as a model because it is included in a complete genome sequence. The T4SS is encoded in the 24-genes operon tfe. We could find no studies on the structure or assembly of this T4SS, but comparative analyses identified 13 genes present in many homologous elements and a number of essential genes (27,56,57). We built profiles for 18 different proteins of MPF (Figure 4). We were unable to retrieve profiles for Tfe1ICEIn10810, Tfe2ICEIn10810, Tfe20ICEIn10810 and Tfe21ICEIn10810; the first two seem non-essential for conjugation (27). The core complex of MPF might resemble the one of MPF since Tfc13MPFG, Tfc14MPFG and Tfc15MPFG (as well as Tfc2MPFG) are respectively homologs to TraKMPFF, TraBMPFF and TraVMPPF. Profile–profile alignments revealed similarities between MPF and the inner scaffold of MPF and MPF with Tfc1MPFG, Tfc18MPFG and Tfc2MPFG being homologous to VirB3/TraLMPF, VirB6/TraVMPPF and VirB8/TraEMPPF proteins. MPF, like MPF, exhibits two VirB-like pilins, Tfe9MPFG and Tfe10MPFG. Interestingly, MPF contains TraWMPFG, TraUMPPF, TraCMPPF and TraFMPFF homologs, although they were thought to be specific to MPF (38). Deletion of Tfc24, the homolog of TraVMPPF, results in decreased transcription of many of the T4SS genes (27). We could also find the only detected homolog to TraHMPF among all T4SS classes: Tfc22MPFG. Overall, our set includes 18 protein profiles including nearly all T4SS essential proteins, except Tfe20ICEIn10810 and Tfe21ICEIn10810. Profile–profile alignments show striking homologies between MPF and MPF in spite of their evolutionary distance and the use of very different pilin by the two systems.

**MPF**

*Bacteroides thetaiotaomicron* CTnDOT ICE encodes the MPF model system in a 17-genes operon: *traA-traQCTnDOT* (38). Only one of these proteins (TraGCTnDOT) was previously found to be homologous to components of the T4SS of Proteobacteria (VirB4 protein) (59). The genes *traA-traDCTnDOT* seem more variable than the rest of the system when compared with the closely related element CTnERL (60); they might thus be non-essential or evolve so fast that sequence similarity between distant elements is lost. The proteins encoded by *traG-NCTnDOT* are reportedly essential
for conjugation whereas traO-QCTnDOT might have a regulatory role (58,60). We obtained profiles for all proteins TraEF, TraQCTnDOT. The vast majority (99%) of T4SS that we detected with these profiles is encoded in genomes of the Bacteroidetes phylum. Profile–profile comparisons show that MPFB has at least two VirB9 homologs, TraQMPFB and TraNMPFB. TraMMPFB matches the VirB10 PFAM domain, but not our VirB10 protein profile. As for the homology between VirB10 and TraB, this is because the PFAM domain is much shorter and includes more distant homologs. The comparisons also revealed that the most-conserved proteins of the inner scaffold (VirB3, VirB6 and VirB8) have homologs in MPFB (respectively TraFMPFB, TraJMPFB and TraKMPFB). The TraECTnDOT and TraIcTDnDOT proteins are thought to be pilins since we find them to be homologous to VirB2 and VirB5 in our analysis. TraOMPFB is homologous to TraMPFB, suggesting that MPFB might have the peculiarity of encoding three different pilins. We were thus able to define 12 protein profiles for MPFB, including most known essential genes of CtnDOT, of which eight have homologs in other systems. This analysis suggests that this class of T4SS strongly resembles some proteobacterial T4SSs. Hopefully, this will facilitate further studies on these systems.

MPFC

There is some circumstantial evidence of conjugation in Cyanobacteria (61), and we have previously identified VirB4 homologs in genomes of this taxa (11). Yet we were unable to find experimental or computational studies on cyanobacterial conjugative systems. We built eight protein profiles that were highly MPFC-specific (>95% hitting genomes of Cyanobacteria). These hits are systematically associated with VirB4 and form compact ge-
Figure 4. Representation of the different mating-pair formation (MPF) classes. The length of the arrows is proportional to the mean length of the corresponding genes. Bold arrows represent genes for which the corresponding protein profile was already available (6,8,11). White arrows represent genes for which we did not obtain a profile. Gray arrows represent genes for which we built a protein profile lacking homologs in other classes. A given color (except white and gray) corresponds to a single family of homologs.

Genetic loci. We used the alpha plasmid of Nostoc sp. PCC 7120 as a model for MPF<sub>C</sub>. Profile-profile comparisons show that MPF<sub>C</sub> have two proteins homologous to the core complex in other systems (a VirB<sub>9</sub>/TraV<sub>MPFF</sub> and a VirB<sub>10</sub>/TraB<sub>MPFF</sub> homolog). The MPF<sub>C</sub> inner scaffold might resemble those of MPF<sub>T</sub> and MPF<sub>E</sub> since we found homologs to VirB3/TraL<sub>MPFF</sub>, VirB6/TraK<sub>MPFF</sub> and VirB8/TraE<sub>MPFF</sub> (Figure 3). Overall, among the eight new MPF<sub>C</sub> protein profiles, five show homologs with components of MPF from Proteobacteria.
that some MPFFA systems use VirD4 as coupling protein, whereas others correspond to those in Figure 4. The bicolor VirD4 protein losses. The colors of the arrows (as shown on the right) in (11). Green shapes represent inferred protein gains, whereas red shapes of VirD4 and encode a peculiar relaxase (75). We have a hydrolase-like activity on C. perfringens (69), which is consistent with a VirB1-like role in MPF. We could not build a profile for Orf15t916 matching our specificity criteria (less than 50% of the hits neighbored virB4). This gene encodes an anti-restriction protein that is probably not part of the typical MPF system.

The 12-gene model conjugative locus of Tn916 is the best-described MPF system. This class is mostly found in Firmicutes and Actinobacteria (62,63). Tn916, ICE61 of Bacillus subtilis and the plasmid pCW3 of Clostridium perfringens, all use TcpA as the coupling protein instead of VirD4 and encode a peculiar relaxase (orf20 MPs, a MOB) related to rolling-circle replication initiators from plasmids and phages (29,64). The presence of a relaxase within the MPF region is unique among the systems we considered in this work. These systems do not encode cell surface adhesins (5). We built seven specific profiles for this system: the TcpA distant homolog of VirD4, and six putative components of the T4SS (Figure 4). Orf22 MPs and Orf23 MPs are the only pair of protein families from a single system for which we could only build one single profile that systematically matches both proteins. As expected, given the absence of an outer membrane, our analysis suggests that MPF lacks homologs to the components of the core complex found in proteobacterial systems. Orf15 MPs is the only profile homologous to components of the inner scaffold (VirB6/TraMPF) that we could identify in this T4SS class. The structure of TcpC pCW3, a close homolog to Orf13t916, has remarkable similarities with the one of VirB8 even though sequence similarity is not significant (65). Hence, MPF might have two components homologous to the inner scaffold of MPF—Orf15 MPs and Orf13 MPs—and these proteins have been shown to interact (66). MPF also has a component homologous to VirB1: Orf14 MPs. Its homolog TcpG pCW3 has a hydrolase-like activity on C. perfringens peptidoglycan (67), which is consistent with a VirB1-like role in MPF. We could not build a profile for Orf18 t916 matching our specificity criteria (less than 50% of the hits neighbored virB4). This gene encodes an anti-restriction protein that is probably not part of the typical MPF system (68). The last gene of the operon, orf24 t916, is often found in other MPF systems isolated from the rest of the operon (69), or even not mentioned as part of the conjugative machinery (70). It is thus probably not an essential component of MPF.

MPF

This class includes systems from an extremely diverse group of Prokaryotes (Firmicutes, Actinobacteria, Tenericutes and Archaea). These monoderms have very diverse cell envelopes—no cell wall in Tenericutes, different lipids in Archaea, thick cell walls in Firmicutes and Actinobacteria—and this may have accelerated their diversification. This might explain why we could not create a single set of protein profiles for all the systems in the class. The lack of experimental studies in most MPF systems further complicated our task. Nevertheless, we identified proteins associated with four sub-classes of T4SS for which conjugative systems have been studied experimentally. These sub-classes correspond to monophyletic sub-groups in the MPF systems VirB4 phylogeny. Some of the components have highly similar homologs between sub-classes.

The 13-genes operon encoding the conjugative system of the pGO1 plasmid from Staphylococcus aureus (71,72) was used as a model for a sub-class only found among plasmids of Firmicutes. We built six profiles specific to genes encoded in the operon. The Trs pGO1 and Trs pGO1 profiles were not specific, with respectively only 7 and 2% of the hits associated to a VirB4. The three other sub-classes were modeled from the prg pCF system of plasmid pCF10 from Enterococcus faecalis (73,74), from the ICE CTn2 from Clostridium difficile (75) and from an ICE of the Streplococcus agalactiae NEM316 genome (ICESaNEM316) of the ICESa2603 family (76,77). We did not use ICESa2603 itself as a model because it lacks a T4CP and therefore it does not strictly fit our definition of an ICE. These three MPF sub-classes share a core of three proteins (in addition to the VirB4 and VirD4 homologs): Prf F pCF10 is homologous to CD414 CTn2 and GBS1363 ICESaNEM316; Prg F pCF10 is homologous to CD417 CTn2 and GBS1361 ICESaNEM316; Prg F pCF10 is homologous to CD415 CTn2 and GBS1362 ICESaNEM316. Thus, these nine proteins are recovered by only three different profiles. ICESaNEM316-like systems are only found within some Streptococcus, and never on plasmids. This element also shares an additional profile with the Prg/Pcf system, namely Prg L pCF10, which corresponds to GBS1349 ICESaNEM316.

Profile–profile comparisons showed no protein profile in MPF with homologies to the T4SS core complex. The exception is the homology between Prg L pCF10 and VirB9 at the N-terminus of the latter, where VirB9 interacts with the inner membrane. The VirB9 region interacting with the outer membrane has no discernible homologs in MPF. On the other hand, we found some homologs to components of the inner-membrane pore complex: VirB3 homologs—Trs C pGO1, Prg L pCF10/CD417 CTn2/GBS1361 ICESaNEM316—and VirB6 homologs (Prgh pCF10/CD415 CTn2/GBS1362 ICESaNEM316). Besides, Trs M pGO1, homolog of Trs M pGO1 for which we could not build a profile, is structurally similar to VirB8 (78). All the MPF systems that we describe except ICESaNEM316 encode a VirB1 homolog (Trs G pGO1).
VirB3 has homologs in every class except MPF I and MPF B. VirB6 and VirB8 are found in almost all MPF classes, except its absence from monoderms. Part of the VirB10 protein forms the outer membrane pore (40) and this may explain its absence from monoderms.

The three proteins of the inner-membrane pore (VirB3, VirB6 and VirB8) are found in almost all MPF classes. VirB3 has homologs in every class except MPF I and MPF A. The VirB3 partner ATPase (VirB4) has a domain typical of the conjugation system or sequencing errors. Importantly, very few large (> 5) clusters of profiles lack VirB4 (9%). Absence of VirB4 might be due to pseudogenization of the conjugation system or sequencing errors. Accordingly, some of these large clusters matching a large number of profiles for a given class include a pseudogene of VirB4.

Two of the ATPases of the VirB system (VirB4 and VirD4) are widespread. VirB4 is thought to be ubiquitous in T4SS and our analysis relied on this hypothesis. Importantly, very few large (> 4) clusters of profiles lack VirB4 (9%). Absence of VirB4 might be due to pseudogenization of the conjugation system or sequencing errors. Accordingly, some of these large clusters matching a large number of profiles for a given class include a pseudogene of VirB4.

Among the 104 protein profiles used in this study, 34 have no homologs. These ‘orphan’ profiles are not equally distributed (gray arrows in Figure 4). Only one is found in MPF A, whereas they represent most of the profiles of MPF I. This matches the high complexity of MPF I and its early divergence of MPF A and MPF G from the remaining classes (11). Some of these ‘orphan’ profiles might be distant homologs of proteins present in other systems that passed unnoticed in our sequence-based analysis. For example, we found no homologs among pilins, which evolve very fast in sequence, of different classes. Structural data, when it becomes available for the different classes, will help highlighting these cases and complete the networks of homology of Figure 3.

Evolution of the T4SS gene repertoires

The analysis of the total number of relationships of homology between profiles shows that MPF T is the class with the largest number of homologs in other classes (23 links) (Figure 5). MPF I and MPF G are highly connected among themselves suggesting some homology between these classes at least at the level of the core complex and inner-membrane pore, in spite of not being neighbors.
in the VirB4 phylogeny (11) (Figure 1). MPF_C have a number of components homologous with other MPF classes organized in loci resembling the VirB operon, in spite of the large evolutionary distance between MPF_T and MPF_C in the VirB4 phylogeny (11). MPF_B has more homologs with MPF_T (two VirB9) and MPF_F (homology to the pilins VirB2 and VirB5). Interestingly, MPF_F is the least connected class. This is consistent with its position at the base of the VirB4 phylogeny and its use of a distant homolog of VirB4, TraU (11). MPF_T also has the peculiarity of encoding two types of pili, one of which being necessary for liquid mating and homologous to the type 4 pili (55,91). The systems of monodermers, and particularly MPF_FA, also have few homologs reinforcing the claims that they have fewer components than the other T4SS (20,92), possibly as a result of adaptation to monoderm.

We mapped the patterns of the presence/absence of VirB homologs in the phylogenetic trees and drew the most parsimonious scenario for the recruitment of the different components to the T4SS (Figure 6). The scenarios for fast-evolving proteins must be taken with care because lost sequence similarity may lead to an under-estimation of the proteins already present in the last common ancestor of all T4SS. Nevertheless, our analysis suggests that a large number of components of the VirB systems were present early in the history of T4SS.

Web servers

We had previously made available a web site (now called CONJscan, http://mobyle.pasteur.fr/cgi-bin/portal.py#forms:CONJscan-T4SSscan or the short URL http://bit.ly/CONJscan) that allows searching for T4SS and conjugation-related protein profiles. In the present work, we increased the number of profiles available for search from 45 to 116. Our previously identified T4SS and relaxases were distributed as a simple text file. This made its analysis difficult with no intuitive or user-friendly way to search or filter the results according to different criteria. We have now created a web site, CONJdb, which allows searching and browsing our extended dataset thanks to a graphical interface. Our web site also implements the classification scheme. Finally, the graphical interface allows to performs searches by bacterial species, by T4SS class, or according to the presence or absence of T4SS components, namely the T4SS, the coupling protein and the relaxase. This should facilitate the discrimination of T4SS dedicated to conjugation from T4SS dedicated to protein secretion. We are working on the development of a stand-alone application that will be distributed and should be of use to analyze large metagenomic datasets with local computational resources. In the future, CONJdb will be regularly updated and linked with other genomic databases. The web site is accessible at http://conjdb.web.pasteur.fr.

The present release of CONJdb has the results of the analysis of 2393 chromosomes and 1813 plasmids for a total of 2262 complete prokaryotic genomes. Plasmids sequenced without the corresponding chromosome were not included. We detected 947 conjugative systems (up from 515), 1181 mobilizable elements (up from 595) and 646 T4SS lacking nearby relaxases (up from 243), for a total of 1623 T4SS. The most recent and comprehensive T4SS database to date, SecReT4 (93), contains 811 T4SS present on 638 different replicons. SecReT4 does not contain Cyanobacteria or Archaea T4SS. Another tool, AtlasT4SS (94), contains information from 70 genomes (58 Bacteria, 1 Archaea and 11 plasmids) and contains 134 clusters of orthologs. Importantly, CONJdb includes information on the T4SS class and on the putative T4SS function. Finally, because proteins are annotated using HMM profiles, instead of blast-like searches in a large databank of sequences, CONJscan and CONJdb provide standard sequence annotations. We hope this study and these web resources will be useful for scientists studying T4SSs and particularly for those engaging in the study of poorly known ones.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Spanish Ministry of Economy [BFU2011-26608]; European Seventh Framework Program [282004/FP7-HEALTH.2011, 612146/FP7-ICT-2013]; European Research Council Grant [EVOMOBILOME no. 281605]. Source of open access funding: European Research Council grant to the PI.

REFERENCES


