Functional Diversification of the Twin-Arginine Translocation Pathway Mediates the Emergence of Novel Ecological Adaptations

Xiaowei Jiang and Mario A. Fares

1Evolutionary Genetics and Bioinformatics Laboratory, Department of Genetics, Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin, Ireland
2Integrative Systems Biology Group, Department of Abiotic Stress, Institute of Molecular and Cellular Biology (Consejo Superior de Investigaciones Científicas (CSIC)-Universidad Politécnica de Valencia (UPV)), Valencia, Spain

Abstract

Microorganisms occupy a myriad of ecological niches that show an astonishing diversity. The molecular mechanisms underlying microbes’ ecological diversity remain a fundamental conundrum in evolutionary biology. Evidence points to that the secretion of a particular set of proteins mediates microbes’ interaction with the environment. Several systems are involved in this secretion, including the Sec secretion system and the Tat pathway. Shifts in the functions of proteins from the secretion systems may condition the set of secreted proteins and can, therefore, mediate adaptations to new ecological niches. In this manuscript, we have investigated processes of functional divergence (FD)—a term used here to refer to the emergence of novel functions by the modification of ancestral ones—of Tat pathway proteins using a large set of microbes with different lifestyles. The application of a novel approach to identify FD allowed us to distinguish molecular changes in the three Tat proteins among different groups of archaea and bacteria. We found these changes as well as the composition of secreted proteins to be correlated with differences in microbe’s lifestyles. We identified major signatures of FD in halophilic and thermophilic archaea as well as in pathogenic bacteria. The location of amino acids affected by FD in functionally important domains of Tat proteins made it possible to find the link between the molecular changes in Tat, the set of secreted proteins and the environmental features of the microbes. We present evidence that links specific molecular changes in secretion mediating proteins of microbes to their ecological adaptations.

Key words: functional divergence, Tat secretion, ecological adaptation.

Introduction

Bacterial, archaeal, and eukaryotic cells use two major secretion systems to transport proteins that are synthesized in the cytosol, namely the general secretion (Sec) pathway and the twin-arginine translocation (Tat) pathway. The Sec pathway transports proteins in an unfolded state, whereas the Sec-independent Tat pathway transports proteins that are partially or fully folded. The N-terminal signal peptide, as part of the substrate protein, plays a significant role in targeting proteins to the cytoplasmic membrane, where they can be integrated or alternatively transported across. This N-terminal peptide presents a tripartite configuration that includes an N-terminal region at the beginning, a hydrophobic region in the middle, and a C-terminal region in the end. A typical Tat signal peptide has an arginine–arginine (twin-arginine) consensus motif in its N-terminal region (Yuan et al. 2010).

Two major differences exist in the mechanism of protein secretion of Sec-dependent and Tat systems. The Sec-dependent translocation pathway uses two modes of protein secretion: the cotranslational translocation and the posttranslational translocation. In the cotranslational translocation pathway, which is conserved in both archaea and bacteria, proteins are translocated across or into the plasma membrane during protein synthesis. In contrast to this, in the posttranslational translocation, proteins are first synthesized in the cytosol and stabilized in an unfolded state through the chaperone activity of SecB and/or SecA for their subsequent translocation across the membrane (McFarland et al. 1993; Eser and Ehrmann 2003). The translocation of proteins occurs through the pore formed by SecY protein and favored by the ATPase activity of SecA. Importantly, SecB and/or SecA mediated translocation pathway is only found in bacteria but not in archaea. The mechanism that archaea use to solve the problem of posttranslational translocation—that is stabilizing proteins in an unfolded or partially folded states for their subsequent translocation across the membrane, remains elusive (Irihimovitch and Eichler 2003).

The Sec-independent Tat pathway is conserved in bacteria, archaea, chloroplasts, and plant mitochondria (Lee et al. 2006), pointing to its essentiality in the viability of these organisms. It consists of three major components, TatA, TatB, and TatC, all of which are present in most gram-negative bacteria. In contrast, most gram-positive bacteria and archaea have only two Tat components, TatA and TatC (Eijlander et al. 2009). Most of our knowledge...
regarding the mechanisms of substrate recognition and translocation by the Tat system is based on studies of the two model organisms, *Escherichia coli* and *Bacillus subtilis* (Panahandeh et al. 2009). These studies show that Tat translocase uses Tat(B)C complex as a receptor site for binding Tat substrate proteins. After substrate binding, Tat(B)C recruits multiple TatA proteins to form a pore-containing protein complex that allows passing substrates through the pore driven by a proton motive force (Lee et al. 2006; Tarry et al. 2009; Yuan et al. 2010). TatC was shown to determine the substrate specificity and thus plays a critical role in Tat-dependent protein secretion (Strauch and Georgiou 2007; Eijlander et al. 2009). What is the mechanism whereby Tat pathway monitors the folding fidelity of proteins for their correct translocation? In one study, authors showed that indeed TatA presents a proofreading function of its substrate FeS proteins NrfC and NapC, so that when such substrates are incorrectly folded, they undergo rapid degradation (Matos et al. 2008).

Although the mechanism is not entirely clear, Tat pathway seems to mediate the emergence of new ecological adaptations in many bacteria and archaea. For example, recently, an association between ecological adaptation and the Tat secretion pathway has been established in halophilic archaea *Halobacterium* sp. *NRC-1* (Rose et al. 2007). Studies showed that this group of archaea uses the Tat translocation pathway extensively to transport most of their secreted proteins (Dilks et al. 2005). Haloarchaea live in an environment with “near-saturation” salt concentrations, a harsh condition for the proper function and secretion of proteins whose functions are important in later stages outside the cytosol. Interestingly, this correlates with the fact that most sequenced haloarchaea have two TatC homologs (TatCo and TatCt). In contrast, therefore, to many bacteria in which Tat transports a small fraction of proteins and are not essential for cell viability (Lee et al. 2006), TatC may be essential to haloarchaea in buffering the environmental conditions.

Not only is the Tat pathway a major route for protein secretion in haloarchaea but also it is essential for cell viability (Dilks et al. 2005; Thomas and Bolhuis 2006). Some other prokaryotes are known to use the Tat pathway extensively, such as *Streptomyces* species (e.g., *S. coelicolor*) (Widdick et al. 2006). This difference in Tat essentiality between microbes seeks an explanation and can be crucial to understand ecological adaptation. What makes Tat essential and what molecular bases contribute to the different essentiality of Tat in different groups of organisms? Could we find molecular changes that are associated with environmental adaptations? Are there any other prokaryotes that present similar patterns in their Tat translocase? These questions remain largely unexplored.

Secretion systems play a significant role in prokaryotes environmental adaptation (Wooldridge 2009), and, as such, understanding the evolution of these systems is a crucial aim in evolutionary biology. Unfortunately, however, contrary to cytosolic proteins, a large amount of secreted proteins are structurally uncharacterized, posing difficulties to perform detailed evolutionary analyses (Elofsson and von Heijne 2007; Popot 2010). Alternative approaches are therefore required to circumvent the limitations to studying the evolution of these systems.

Here, we hypothesize that the Tat pathway has diverged its function from the ancestral ones in microbes with particular ecological traits. As a case in point, we focus in two main types of microbes: halophilic archaea, that live in environments saturated with salt, and pathogenic bacteria, that may require the secretion of proteins to interfere with the immune response of the host. We use prokaryotic genome sequences, together with abundant secondary structure data, a new computational tool to identify divergence in protein functions and a wide range of experimental studies reporting crucial information to test this hypothesis.

**Materials and Methods**

**Sequences and Genomes**

Homologous protein sequences for TatA, TatB, and TatC were retrieved using Application Program Interface (http://www.genome.jp/kegg/soap/) service from Kyoto Encyclopedia of Genes and Genomes pathway database (Kanehisa et al. 2008). Genomes and protein table files were retrieved from NCBI (National Center for Biotechnology Information) Genome database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.ppt.tar.gz). Protein table files were later used for functionally annotating Tat-dependent substrates proteins with Clusters of Orthologous Groups (COG) (Tatusov et al. 2003). We also annotated all pathogenic species with their host (human, animal, and plant) and the diseases they cause (ICD-10, International Statistical Classification of Diseases and Related Health Problems tenth Revision, http://apps.who.int/classifications/apps/icd/icd10online) wherever possible as this allowed us to classify bacteria as being pathogenic when appropriate. Amino acid sequences of all homologs were aligned using MUSCLE 3.7 (Edgar 2004), and alignments were manually checked using Jalview (Waterhouse et al. 2009). Gaps are removed in the functional divergence analysis.

**Phylogeny Reconstruction**

Amino acid substitution model and model parameters were chosen using ProTest (Abascal et al. 2005) for each of the Tat proteins: TatA, TatB, and TatC. These parameters were then used by RaxML program (Stamatakis 2006) to estimate the best Maximum likelihood phylogenetic trees for the three proteins. The phylogenetic trees of TatA, TatB, and TatC with their corresponding protein sequence alignments were used as starting data for the analyses of functional divergence (FD).

**Analysis of FD in Tat Proteins**

FD is a term that refers to a change in function as a result of an amino acid substitution event within a protein. Here, we used a relaxed definition of FD to describe shifts in the ancestral function of a protein as a result of important amino acid substitutions in that protein—these shifts might not involve the emergence of a completely novel function but
the modification of an existing one. FD can therefore refer to complete change in function after gene duplication or slight shifts in the function of a protein after speciation.

Previous studies have used the term FD to exclusively refer to shifts in the function of a protein after gene duplication—that is to say, FD between paralogs. Under this view, they classified FD into two main categories, type I and type II FD (Gu 1999). In type I FD, amino acids are highly conserved at particular sites in one paralog while being highly stochastic in the other, indicating the acquisition of a function at these sites in one paralog (the conserved one) or the loss of the function in the other. In type II FD, amino acids are highly conserved in both paralogs at the same positions of the protein, indicating the divergence of two highly important functions between the paralogs.

Recently, a new method was implemented to indentify type I FD at the genome level between orthologs (Toft et al. 2009) and was proved efficient in identifying conserved amino acid sites with FD in a large phylogenetic tree including orthologs as well as paralogous sequences (Williams et al. 2010). Briefly, the method uses a protein sequence alignment and a phylogenetic tree including paralogs and orthologs. The method then compares a pair of clades in the tree sharing a common ancestral origin to their closest phylogenetic outgroup. The comparison is performed for each amino acid site in the protein and the strength or likelihood of the amino acid state is evaluated using the appropriate BLOSUM matrix (BLOcks of Amino Acid SUBstitution Matrix) (Henikoff S and Henikoff JG 1992). BLOSUM matrices comprise the scores for the transition between the 20 amino acids, with positive scores meaning the transition is more frequent than expected, negative means the transition is less frequent than expected, and 0 means these transitions are as frequent as expected. Using BLOSUM as a score matrix for the transitions between amino acids allows scoring both variable sites and conserved sites: one clade may seem variable yet the amino acid transitions between orthologous sequences within the clade may have taken place between biochemically close amino acids (positive scores). In the comparison between two clades to an outgroup (clade 1 and clade 2), FD in clade 1 would be detected if the BLOSUM scores were positive within clade 1, negative between clade 1 and the outgroup, and positive between clade 2 and the same outgroup. To test the significance of these score variances, we calculated the mean and standard error of the scores for clade 1 (\(\bar{C}_1, SE_1\)) and clade 2 (\(\bar{C}_2\) and \(SE_2\)) (Toft et al. 2009). The significance of the difference in the means was obtained by calculating the Z-score for the comparison of the two clades as:

\[
Z = \frac{\bar{C}_1 - \bar{C}_2}{SE_{C_1,C_2}},
\]

where \(\bar{C}_{1,2}\) are the mean substitution scores for the transition from clades on either side of the bifurcation in the phylogenetic tree relative to the outgroup, and \(SE_{C_1,C_2}\) is the combined standard error of clade 1 and clade 2.

**Protein Secondary Structure Prediction**

Two approaches were used to predict protein secondary structures for TatC, TatA, and TatB, respectively. We first used the SMART database to predict protein domains in TatC (Letunic et al. 2009). We then used the Psipred to predict protein structures of TatA and TatB (Bryson et al. 2005). The latter method is based on artificial neural network approaches and can reach accurate protein secondary structure predictions (http://bioinf.cs.ucl.ac.uk/index.php?id=779). Protein domains of *E. coli* TatC were predicted first and then mapped back to the protein sequence alignment to define protein domains for all the proteins in the alignment. These domains included three cytoplasmic domains (C1–C3), six transmembrane domains (T1–T6), and three periplasmic domains (P1–P3). Similarly, protein sequences of TatA and TatB from *E. coli* were used as reference sequences. N-terminal region, transmembrane domain, hinge region, the amphipathic helix, and C-terminal domain were predicted for *E. coli* TatA and TatB, respectively. We then used the *E. coli* TatA and TatB as reference and mapped these structural regions in their multiple sequence alignments to define the protein domains for TatA and TatB. After FD analysis, we mapped amino acid sites under FD to these protein domains. To understand the relevance of the results of FD analyses, we built a two dimensional matrix, with the rows representing clades under FD and the columns representing the number of sites under FD in each protein domain. Three such data matrices were made, one per protein: TatC, TatA, and TatB, respectively. We then used the heatmap.2 function in R (http://www.r-project.org) to cluster the rows and columns to find common patterns of FD among protein domains and clades in the matrix. The heatmap.2 function scales each element in a row in the data matrix based on a normalized Z-score, which is calculated as follows:

\[
z_{ij} = \frac{x_{ij} - \bar{x}_i}{\sigma_i},
\]

Here, \(z_{ij}\) is the Z-score calculated for element \(x_{ij}\) in the matrix, \(\bar{x}_i\) is the mean of the row \(i\), and \(\sigma_i\) is the standard deviation of all the elements in the row \(i\).

**Prediction of Potential Sec and Tat Substrate Proteins**

We used Tatp to predict potential Tat-dependent secreted proteins (secretome) (Bendtsen et al. 2005) and SignalP to predict Sec-dependent secretome (Emanuelsson et al. 2007). In predicting Sec-dependent substrates, we first used SignalP and then removed those predicted also by TatP to minimize false positives. Predicted substrates from Sec and Tat pathways were used to assess how different the two secretion systems, Tat and Sec, were and how much they explained about ecological adaptations in our organisms. We preferred using TatP instead of Tatfind (Rose et al. 2002) for the prediction of Tat-dependent secretome as the former approach is based on artificial neural networks that largely outperforms other approaches in pattern recognition.
recognition tasks pertaining variant Tat signal peptides. In contrast, Tatfind is a regular expression–based approach, which does not allow recognition of variant Tat signal peptides (Bendtsen et al. 2005).

Predicted Sec and Tat substrates were grouped according to their COG functional categories. To determine if the numbers of substrates for Tat-dependent secretion in each of the categories were significantly higher (enriched category for secreted substrates) or lower (impooverished category for secreted substrates) than expected by chance, a χ² test was performed on each COG functional category with 1 degrees of freedom. The statistical test was performed according to the following equation:

$$\chi^2 = \frac{(O_i - E_i)^2}{E_i},$$

where $O_i$ is the observed number of predicted substrate proteins in COG category $i$, whereas $E_i$ is the expected number of such proteins in that category $i$. The total number of proteins in category $i$ is indicated by $n_i$, $N_{predicted}$ is the total number of predicted substrate proteins, whereas $N_{COG}$ is the total number of proteins in the proteome. Only proteins assigned within single COG functional categories (ambiguous categories R and S were not included) were analyzed.

### Results and Discussion

**Evidence for FD in the Tat Translocation System**

In this study, we analyzed Tat homologs (orthologs and paralogs) retrieved from 944 bacterial genomes, 68 archaean genomes, and 11 plant genomes. We estimated the amount of FD in each of the Tat proteins in the different lineages of the phylogenetic tree (see Materials and Methods for details). FD refers here to the divergence of protein’s function from its ancestral function by amino acid replacements in functionally important amino acid sites. We applied a method to identify FD previously developed (Toft et al. 2009; Williams et al. 2010). Analysis of our trees allowed us to identify 204, 145, and 119 clades with at least one conserved amino acid site with evidence of FD for TatC, TatA, and TatB, respectively (Supplementary Material online). The number of clades under FD in each Tat phylogenetic tree (TatC: 204; TatA: 145; and TatB: 119) seems to be correlated with TatC and TatA being the most important components in determining a minimum Tat translocase (Blaudeck et al. 2005; Panahandeh et al. 2009). In particular, a mutant TatA could complement a strain with a deleted TatB without compromising substrates translocation (Blaudeck et al. 2005).

In total, we identified 167 bacterial pathogens (100% of the pathogenic bacteria in our data) and 8 haloarchaea (100% of the analyzed haloarchaea) to be involved in a cluster with at least one site under FD. To understand the relationship between FD and ecological adaptation, we focused on the five most functionally divergent clades of the tree, that have at least four sites with evidence of having diverged their functions from ancestral clades.

### Strong Functional Divergence in TatC from Halophilic Archaea

Haloarchaea presented substantial FD following the TatC gene duplication event that gave rise to TatCo (table 1, amino acid sites under FD are mapped to Tat alignment, as shown in supplementary fig. S4, Supplementary Material online) and TatCt (table 2, amino acid sites under FD are mapped to Tat alignment, as shown in supplementary fig. S5, Supplementary Material online) (fig. 1A). A close look to the distribution of amino acid sites with evidence of FD shows that the different proteins domains present significant differences in their content of FD (fig. 1A). In general, we could identify three main clusters of domains according to the content in functionally divergent amino acid sites (sites under FD are shown in supplementary tables S4–S6, Supplementary Material online). The first cluster consists of two protein domains (from left to right: P3 and C4, fig. 1A) that were impoverished for FD. The second cluster included five protein domains (from left to right: C3, T6, T5, T1, and T3, fig. 1A) that showed modest number of amino acid sites under FD. Finally, the third cluster comprised six protein domains (from left to right: C1, C2, P1, T2, T4, and P2, fig. 1A) and was highly enriched for FD.

Interestingly, halophilic archaea TatCo and TatCt proteins presented two different patterns of FD: TatCt showed strong FD in the second cluster of FD domains, affecting mostly TatC transmembrane domains while TatCo presented most of its FD in the N-terminal region of the protein (fig. 1A). Nonoverlapping FD between the protein
Fig. 1. Clustering analysis of phylogenetic clades and protein domains according to FD analyses. We examined enrichment for amino acid sites with FD in the proteins of the Tat secretion pathway TatC (A), TatA (B), and TatB (C) under FD. The relative position of the enriched clades for FD is shown in red in the maximum likelihood phylogenetic tree next to each of the heatmaps. The high resolution trees are available as Supplementary Material (TatC: supplementary fig. S10, Supplementary Material online; TatA: supplementary fig. S11, Supplementary Material online; and TatB: supplementary fig. S12, Supplementary Material online).
Table 1. Amino Acid Sites Under Functional Divergence in Haloarchaea TatCo.

<table>
<thead>
<tr>
<th>Cola</th>
<th>Zb</th>
<th>Residues (count) for C1/C2/Outgroupc</th>
</tr>
</thead>
<tbody>
<tr>
<td>226</td>
<td>8.63***</td>
<td>V(4) M(1) L(1) A(2) R(4)/R(1)</td>
</tr>
<tr>
<td>242</td>
<td>3.32***</td>
<td>A(3) T(2) S(3)/I(2) M(2)/V(1)</td>
</tr>
<tr>
<td>245</td>
<td>2.878</td>
<td>V(6) L(1) A(1)/A(2) S(1) I(1)/S(1)</td>
</tr>
<tr>
<td>539</td>
<td>5.549***</td>
<td>H(2) G(2) Q(2) N(2)/S(3) A(1)/S(1)</td>
</tr>
<tr>
<td>543</td>
<td>198.49***</td>
<td>L(7) T(1)/W(3) M(1)/W(1)</td>
</tr>
<tr>
<td>551</td>
<td>34.199***</td>
<td>V(8)/L(3) M(1)/F(1)</td>
</tr>
<tr>
<td>553</td>
<td>5.26***</td>
<td>T(3) G(1) S(3) A(1)/L(3) F(1)/L(1)</td>
</tr>
<tr>
<td>569</td>
<td>10.404***</td>
<td>T(6)/S(2)/A(3) T(1)</td>
</tr>
<tr>
<td>577</td>
<td>197***</td>
<td>M(8)/A(4)/A(1)</td>
</tr>
<tr>
<td>596</td>
<td>73.90***</td>
<td>Y(8)/F(3) L(1)/L(1)</td>
</tr>
<tr>
<td>599</td>
<td>10.58***</td>
<td>S(3) A(5)/I(1) V(3)/I(1)</td>
</tr>
<tr>
<td>604</td>
<td>116.59***</td>
<td>L(7)/V(7)/F(3) L(1)/Y(1)</td>
</tr>
<tr>
<td>607</td>
<td>19.379***</td>
<td>G(2)/S(2) A(4)/F(4)/F(1)</td>
</tr>
<tr>
<td>611</td>
<td>2.798**</td>
<td>V(5)/I(2)/L(1)/A(2) T(1) G(1)/T(1)</td>
</tr>
<tr>
<td>629</td>
<td>43.13***</td>
<td>Y(8)/Y(3) L(1)/I(1)</td>
</tr>
<tr>
<td>709</td>
<td>4.966***</td>
<td>I(3)/V(3) L(1)/I(1) A(3) P(1)/P(1)</td>
</tr>
<tr>
<td>808</td>
<td>34.199***</td>
<td>L(8)/V(4)/A(1)</td>
</tr>
<tr>
<td>811</td>
<td>3.125***</td>
<td>T(3)/A(1) G(2)/R(1) S(1)/Q(3) G(1)/E(1)</td>
</tr>
<tr>
<td>815</td>
<td>3.376**</td>
<td>D(2)/N(6)/S(4)/T(1)</td>
</tr>
<tr>
<td>817</td>
<td>34.199***</td>
<td>I(4)/L(2)/M(2)/V(4)/A(1)</td>
</tr>
<tr>
<td>823</td>
<td>34.199***</td>
<td>L(2)/F(3)/Y(3)/G(4)/S(1)</td>
</tr>
<tr>
<td>839</td>
<td>115.6**</td>
<td>M(8)/K(4)/R(1)</td>
</tr>
<tr>
<td>844</td>
<td>13.071**</td>
<td>T(7)/S(1)/E(1) K(2)/Q(1)/K(1)</td>
</tr>
<tr>
<td>845</td>
<td>131.98***</td>
<td>R(8)/Y(8)/R(1)</td>
</tr>
<tr>
<td>847</td>
<td>135.95***</td>
<td>W(8)/T(4)/T(1)</td>
</tr>
<tr>
<td>859</td>
<td>4.406**</td>
<td>G(7)/A(1)/V(1) A(2) L(1)/L(1)</td>
</tr>
<tr>
<td>860</td>
<td>6.419**</td>
<td>F(3)/S(2)/A(3)/I(3)/L(1)/I(1)</td>
</tr>
<tr>
<td>879</td>
<td>237.7**</td>
<td>P(8)/Q(4)/Q(1)</td>
</tr>
<tr>
<td>883</td>
<td>43.13***</td>
<td>A(7)/T(1)/A(3) S(1)/L(1)</td>
</tr>
<tr>
<td>887</td>
<td>167.22***</td>
<td>T(8)/L(3)/V(1)/L(1)</td>
</tr>
<tr>
<td>893</td>
<td>34.199***</td>
<td>T(8)/S(5)/G(1)</td>
</tr>
<tr>
<td>902</td>
<td>54.715***</td>
<td>W(8)/L(3)/F(1)/L(1)</td>
</tr>
</tbody>
</table>

Note.—Amino acid residues are represented with one letter code. Numbers of the same amino acid residues are also calculated in each clade. Amino acid sites under FD are also mapped to TatC alignment shown in supplementary figure S4, Supplementary Material online.
a Relative position of amino acid sites in TatC multiple sequence alignment.
b Score of FD: *, **, and *** indicate P < 0.05, P < 0.01, and P < 0.001, respectively.
c C1 is the clade under FD, C2 is the clade against which comparative analyses are made. Amino acid residues and their numbers are labelled in underline in C1.

dating proteins with variations in the RR motif of Tat signal peptide (Kreutztebeck et al. 2007; Strauch and Georgiou 2007), a motif that needs to be generally conserved for the translocation to take place. These studies suggest that extensive FD in these domains of haloarchaea TatCo has contributed to the expansion of Tat substrates repertoire in this clade. The same rationale applies to the pathogenic bacteria *Leptospira*, *Bartonella*, and *Rickettsia* (De Buck et al. 2008; Joshi et al. 2010; Reynolds et al. 2011).

TatC transmembrane domains have been shown to interact with other TatC or TatB to form complexes for signal peptide and substrate binding (Alami et al. 2003; Behrendt et al. 2007; Punginelli et al. 2007). However, the exact function of these domains in substrate recognition and binding is less well understood. The observed FD in TatCt transmembrane domains may have led to the formation of a functionally different TatCt complex, which could interact with a specific group of Tat substrates in haloarchaea.

TatC was suggested to play a role in determining the specificity of Tat pathway–dependent secretion (Jongbloed et al. 2000). In *B. subtilis* and *E. coli*, both TatCs were shown to serve as a primary RR motif recognition site (Mendel et al. 2008). Haloarchaea encode two TatC paralogs, and both are among the most functionally divergent clades. Rose et al. (2002) first showed that *Halobacterium* sp. NRC-1 extensively uses the twin-arginine translocation.
pathway, instead of Sec, to transport most of its secreted proteins. Such shift of protein transport from Sec to Tat may have been crucial to solve the protein folding problem faced by microbes in high salt concentration environments: folding proteins first in the cytosol and then exporting them from the cell. This provides an explanation to the essentiality of TatC proteins for cell viability in halophilic archaea (Dilks et al. 2005; Thomas and Bolhuis 2006). Moreover, we analyzed the secretomes of two extreme halophilic bacteria Salinibacter ruber and Halorhodospira halophila, whose genomes are available (see equations 3 and 4). Strikingly, COG category P—that comprises proteins involved in inorganic ion transport and metabolism—is ranked as the largest group with known function in Sec-dependent secretome of S. ruber (chi-square = 23.49, \( P < 0.001 \), supplementary fig. S6C and table S10, Supplementary Material online). Conversely, in Tat-dependent secretome, functional category P is only ranked at sixth position (chi-square = 0.18, not statistically significant, supplementary fig. S6D and table S11, Supplementary Material online). Similarly, P is ranked at the second (chi-square = 9.48, \( P < 0.01 \)) and fourth (chi-square = 1.14, not statistically significant) positions in Sec- and Tat-dependent secretomes in H. halophila, respectively (supplementary fig. S6A and B and tables S10 and S11, Supplementary Material online). Not surprisingly, none of the two halophilic bacteria, which present both secretion systems Sec and Tat, showed any sign of strong FD in their Tat components. In the posttranslational translocation pathway in bacteria, SecB and/or SecA, that present chaperone activity, can bind to the newly synthesized proteines and keep them in an unfolded state, which is required for their Sec-dependent translocation. In halophilic bacteria, SecB and/or SecA can act as chaperones preventing the premature folding of the synthesized proteins, ensuring therefore correct protein translocation despite the high salt concentrations. Contrary to this, halophilic archaea do not have SecB and/or SecA proteins that could mitigate the effects of high salt concentrations on protein translocation (Dilks et al. 2005). In these archaea, proteins need therefore to be translocated in a folded state, which is possible through the Tat pathway that does not require unfolded proteins. This, however, would be possible provided that the substrate specificity of TatC has relaxed (for example, recognition of RR motif in a signal peptide of Tat substrate proteins would be no longer a requirement), possibly through key amino acid substitutions that we detect to be under FD.

In conclusion, the distinct FD patterns observed here between TatCo and TatCt have played an important role in the adaptation of halophilic archaea to environments with high salt concentrations, which is a clear example of how FD of the Tat pathway has contributed to the ecological adaptation of halophilic archaea.

### Functional Divergence of Thermophilic Archaea Sulfolobus TatA

FD analyses of TatA highlighted thermophilic archaea Sulfolobus to be the clade with the strongest profile of FD (fig. 1B, amino acid sites under FD are mapped to TatA alignment shown in supplementary fig. S7, Supplementary Material online). The other two clades containing non-pathogenic Mycobacteria and pathogenic E. coli strains have the same amount of amino acid sites under FD as Sulfolobus (fig. 1B, supplementary table S5, Supplementary Material online). In TatA, we identified a main cluster joining domains APH, NT, and H (fig. 1B). Importantly, the C-terminal domain showed strong signal of FD compared with the other domains (fig. 1B). In contrast to TatC, however, the amount of FD in the different domains of TatA was low, and we decided therefore to look at the phylogeny that included the entire set of clades to gain insight on FD in TatA. Clustering analyses based on FD approach and taking all clades together for TatA allowed identifying a more general pattern with three well-distinguishable clusters (supplementary fig. S2, Supplementary Material online). The first cluster contained the C-terminal domain, being this the one most affected by FD in TatA, the second cluster was almost exclusively represented by the amphipathic helix domain, and the third cluster contained the transmembrane, the hinge, and the N-terminal domains. This clustering pattern is consistent with previous functional studies that reported mutations in the first 42 residues of TatA, particularly in the amphipathic helix region, to affect significantly Tat-dependent secretion (Hicks et al. 2005). The C-terminal domain was suggested to play an important functional role in lineage-specific substrate transport (Warren et al. 2009). The low FD observed in TatA transmembrane domain points to the conserved role of this domain, mainly responsible for the formation and structural support of the pore-containing complexes for protein transport (Leake et al. 2008; Warren et al. 2009).

### Extensive Functional Divergence in TatB of Pathogenic Bacteria

Analysis of TatB showed extensive FD affecting clades mostly containing pathogenic bacteria. For example, the top five clades in terms of the amount of amino acid sites under FD included Corynebacterium, Neisseria, Bartonella, and Salmonella, bacteria well known for their pathogenic lifestyle (supplementary table S6, Supplementary Material online). Moreover, two human pathogens (Neisseria gonorrhoeae and Neisseria meningitidis) from the Neisseria clade were the second most functionally divergent (11 amino acid sites, supplementary table S6, Supplementary Material online, amino acid sites under FD are mapped to TatB alignment, as shown in supplementary fig. S8, Supplementary Material online). FD affecting TatB was heterogeneously distributed amongst clades and protein domains: The amphipathic helix and the C-terminal domains were the most affected. Interestingly, we observed the same three-cluster pattern as in TatA (supplementary fig. S3, Supplementary Material online), suggesting that both proteins may share similar evolutionary histories. In support of this, a fusion protein consisting of the N-terminal region of TatA and the amphipathic helix and C-terminal domain of TatB can maintain a low level of
Tat-dependent secretion (Lee et al. 2002). Moreover, the secretion of a tat substrate in *E. coli* can still be detected after mutating TatA in a TatB knockout strain (Blaudeck et al. 2005). These data indicate that FD of TatB may serve to determine the specificity of some Tat-dependent substrates through the modulation of the interaction of TatB with TatA, which is important to bacterial virulence (De Buck et al. 2008).

**Fig. 2.** Analysis of Tat-dependent secreted proteins (secretome) of halophilic archaea and pathogenic bacteria. Substrates are grouped functionally according to the classification of proteins into the Cluster of Orthologous Groups (COG). These functional categories were ranked and plotted according to the numbers of their substrates. (A) *Halobacterium* sp. NCR-1, (B) *Haloferax volcanii*, (C) *Halomicrobium mukohataei*, (D) *Leptospira interrogans*, (E) *Bartonella quintana*, and (F) *Rickettsia rickettsii*. COG functional categories showing significant enrichment (black stars) or impoverishment (gray stars) of predicted substrate proteins are labeled by *P* < 0.05; **P** < 0.01; and ***P** < 0.001.

Tat-independent secretion (Lee et al. 2002). Moreover, the secretion of a tat substrate in *E. coli* can still be detected after mutating TatA in a TatB knockout strain (Blaudeck et al. 2005). These data indicate that FD of TatB may serve to determine the specificity of some Tat-dependent substrates through the modulation of the interaction of TatB with TatA, which is important to bacterial virulence (De Buck et al. 2008).

**Differential Protein Transport in Functionally Divergent Prokaryotic Lineages**

We identified the Sec- and Tat-dependent secretome of three halophilic archaeal species, *Halobacterium* sp. NCR-1, *Haloferax volcanii*, and *Halomicrobium mukohataei* using SignalP and TatP approaches, respectively (only the data for Tat, and not for Sec, are shown here). Tat substrates in the three secretomes were grouped based on COG functional categories, which were then plotted and ranked according to their numbers of substrates within each of the categories. These numbers were statistically tested for category enrichment in protein secretion by comparing the observed number of secreted proteins within each category to the expected number. Expected number of secreted proteins within each of the categories was calculated by assuming that the percentage of secreted proteins is proportional to the total number of proteins within COG categories for that organism (see equations 3 and 4). Interestingly, in *Halobacterium* sp. NCR-1, the functional category P—annotated as Inorganic ion transport and metabolism—was significantly enriched for secreted proteins and was ranked the largest group amongst all the COG functional categories (fig. 2A). Similar patterns were also observed in *H. volcanii* (fig. 2B) and *H. mukohataei* (fig. 2C) and in Sec-dependent secretomes (data not shown). It is known that halophilic archaea accumulate potassium ions to counterbalance the high salt concentration in their environment (Albers et al. 2006). It is not surprising therefore to see that a large fraction of Tat-dependent substrates is related to inorganic...
ion transport and metabolism (supplementary table S7, Supplementary Material online). To determine if FD in TatC from haloarchaea may be correlated with its adaptation to halophilic environments and not to other indirect causes, we also identified the Sec- and Tat-dependent secretome of three species from the other three nonhalophilic lineages under FD (fig. 1A): Leptospira interrogans Lai, Bartonella quintana, and Rickettsia rickettsii. Conversely to the case of haloarchaea, Tat-dependent secretome analysis of these bacteria did not rank the COG P category amongst the best represented in the secretome (fig. 2D–F, supplementary table S7, Supplementary Material online).

The Sulfolobus clade was ranked among the top ones regarding the enrichment for FD in TatA (fig. 18). Sulfolobus is known to live in hot environments with temperatures ranging between 60 and 90 °C (Huber and Prangishvili 2006). Ribosomal proteins (RPs) are shown to bind zinc (Makarova et al. 2001) and mounting evidence point to their alternative extracellular location (Trost et al. 2005; Tjalsma et al. 2008; Joshi et al. 2010), in where they likely perform nonribosomal functions (Warner and McIntosh 2009). Interestingly, prediction of Tat-dependent substrate proteins from one of the Sulfolobus islandicus strains (M.14.25) and subsequent COG grouping of the secreted proteins (supplementary table S8, Supplementary Material online) identified the functional category J—which comprises proteins involved in translation, ribosomal structure, and biogenesis—as the most enriched for secreted proteins from the secretome ($\chi^2 = 7.81, P < 0.01$). 13 RPs were predicted to be Tat substrate proteins, 6 of which were predicted to contain zinc-binding sites. Moreover, four of the six RPs were predicted to have multiple zinc-binding sites (one has four zinc-binding sites; three have five zinc-binding sites) (Shu et al. 2008). To test the significance of Tat-dependent zinc-containing RPs, we analyzed the 64 RPs from S. islandicus M.14.25. We found that there are 19 RPs having at least one zinc-binding site, 4 of which have at least 5 zinc-binding sites. Interestingly, three of the four RPs are predicted to be Tat-dependent substrates. The same analysis was performed in other clades to determine whether this observation was restricted to Sulfolobus or was general to thermophiles. Our results suggest that this pattern was unique to Sulfolobus. RP with binding zinc was proposed to be more stable, and therefore, they are more frequently used in thermophilic bacteria (Makarova et al. 2001). In a thermophilic environment, proteins are likely to go denaturation. Zn-binding proteins are more stable in environments that are bound to Tat-dependent RPs, which can subsequently mediate the homeostasis of Zn(II) in the cytosol.

In pathogenic bacteria, COG functional category J was the largest, although not always significantly enriched, Tat-dependent substrate group with FD (fig. 2D–F). This was the case for all the pathogenic bacterial strains we analyzed, such as Leptospira borgpetersenii JB197 (TatC, supplementary fig. S9A, Supplementary Material online), L. interrogans (TatC, fig. 2D), B. quintana (TatB and TatC, fig. 2E), Rickettsia prowazekii Madrid E (TatC, supplementary fig. S9B, Supplementary Material online) and R. rickettsii (TatC, fig. 2F) (supplementary table S7, Supplementary Material online), Corynebacterium diphtheriae (TatB, supplementary fig. S9C and table S9, Supplementary Material online), N. meningitidis Z2491 (TatB, supplementary fig. 5D and supplementary table S9, Supplementary Material online). As shown above, zinc-containing proteins are known to play vital role in bacterial physiology, including DNA polymerases, proteases, and RPs, among others. During pathogenesis, bacterial pathogen can experience zinc starvation, which can interfere with the normal function of zinc-containing proteins. A feasible solution would be using a homologous protein, which does not need, or that would require very little, zinc to function properly (Panina et al. 2003). We found three of eight secreted RPs from the obligate intracellular bacterial pathogen R. rickettsii bearing predicted zinc-binding sites, although prediction scores were weak. Among the identified Tat-dependent secreted proteins in L. borgpetersenii and L. interrogans, we found the 30S ribosomal protein S3 (RPS3). Recently, a study demonstrated that RPS3 is an active component of NF-κB transcription factor complex (Wan et al. 2007). NF-κB plays an important role in the host immune response to pathogens by regulating the expression of genes involved in different immune pathways (Vallabhapurapu and Karin 2009). Bacterial secreted Tat substrate RPS3 can be important to interfere with the normal function of host RPS3, which would impair NF-κB dependent gene regulation and expression. These pathogenic bacteria may therefore evade host immune response by the Tat-dependent secretion of RPS3 and other RPs to the immune cells. This conclusion, however, should be taken with caution and requires further investigation.

**Supplementary Material**

Supplementary tables 1–11 and figures 1–12 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

This work was supported by Science Foundation Ireland, under the Research Frontiers Program (10/RFP/Gen2685).
and a grant from Ministerio de Ciencia e Innovación (BFU2009-12022) to M.A.F. X.J. is supported by Irish Research Council for Science, Engineering and Technology Government of Ireland Postgraduate Scholarship in Science, Engineering and Technology. We would like to thank both the editor and reviewers for their important contribution in helping us to improve the clarity and presentation of this manuscript. Finally, we are thankful to all our colleagues who contributed to improve the quality of the writing of this manuscript.

References


