Computational prediction shines light on type III secretion origins

Tatyana Goldberg [1, 2, *], Burkhard Rost [1, 3, 4] & Yana Bromberg [3, 5]

1 TUM, Department of Informatics, Bioinformatics & Computational Biology - I12, 85748 Garching, Germany
2 TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), 85748 Garching, Germany
3 Institute of Advanced Study (TUM-IAS), 85748 Garching, Germany
4 New York Consortium on Membrane Protein Structure (NYCOMPS) & Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA
5 Department of Biochemistry and Microbiology, School of Environmental and Biological Sciences, Rutgers University, New Brunswick, NJ 08901, USA
* Corresponding author: goldberg@rostlab.org, http://www.rostlab.org/, Tel: +49-289-17-850, Fax: +49-289-19414

Conflict of interest: The authors declare no conflict of interest.

Running Title: pEffect prediction of bacterial type III effectors
Document statistics: Abstract = 196, Text = 4909 words, 60 references, 6 figures; 1 table; 18 pages
Journal: The ISME Journal
Submitted: December 23rd, 2015

Abstract

The type III secretion system transports effector proteins of pathogenic and endosymbiotic Gram-negative bacteria into the cytoplasm of host cells. During infection, effectors convert host resources to work to bacterial advantage. Existing computational methods for the prediction of type III effectors mainly employ information encoded in the N-terminal protein sequence. Here we introduce pEffect, a method that predicts type III effector proteins using the entire amino acid sequence. It combines homology-based inference with *de novo* predictions, reaching 87±7% accuracy at 95±5% coverage for a large non-redundant set of proteins. This performance is up to 3-fold higher than that of other methods. pEffect also sheds new light on effector secretion mechanisms. We establish that “signals” for the recognition of type III effectors are distributed over the entire protein sequence instead of being confined to the N-terminus. Our method, therefore, maintains high performance even when used with sequence fragments like metagenomic reads, and potentially facilitates studies of microbial community interactions. Explorations into the evolutionary origins of type III secretion identify a variety of recently evolved effectors and highlight the possibility of type III secretion ancestor dating to times prior to the archaea/bacteria split. pEffect is available at http://www.bromberglab.org/services/pEffect.

Introduction

Six secretion systems have been identified in pathogenic and endosymbiotic Gram-negative bacteria (Cornelis, 2006; Holland *et al*., 2005; Leo *et al.*, 2012; Low *et al.*, 2014; Nivaskumar and Francetic, 2014). The type III secretion system mediates a wide range of bacterial infections in human, animals and plants (Buttner and He, 2009; Hueck, 1998; Marshall and Finlay, 2014). This system comprises a hollow needle-like structure localized on the surface of bacterial cells that injects specific bacterial proteins, effectors, directly into the cytoplasm of a host cell (Cornelis, 2006). During infection, effectors act in concert to convert host resources to their advantage and promote pathogenicity (Troisfontaines and Cornelis, 2005).

Advances in sequencing techniques are producing an ever-growing number of bacterial genome sequences (Wang *et al.*, 2012). As a result, the identification of bacterial type III effectors has shifted away from experimental discovery of individual proteins to whole genome computational screens. Various machine learning algorithms, including Naïve Bayes (Arnold *et al*., 2009), Support Vector Machines (SVMs) (Wang *et al.*, 2011), Artificial Neural Networks (Lower and Schneider, 2009) and Markov models (Wang *et al.*, 2013) have been deployed to identify type III effectors *in silico*. These methods use sequence similarity to experimentally known effectors as input; this similarity is defined on the basis of different features, such as GC content (coding genes), as well as, amino acid composition, secondary structure, and solvent accessibility (proteins). Methods often focus on features in the protein N-terminus, assumed to be most informative for the translocation of effectors through type III secretion (Ghosh, 2004). An independent benchmark revealed state-of-art-methods to predict type III effectors at similar levels up to 80% accuracy at 80% coverage (McDermott *et al.*, 2011); thus, there still seems to be room for substantial improvement.

Here, we introduce *pEffect*, a method that combines sequence similarity-based inference (PSI-BLAST) with *de novo* prediction using machine learning techniques (Support Vector Machines; SVM). Our method uses information about the *entire* amino acid sequence of each protein. To allow users to focus on most relevant results, it provides a score reflecting the strength of each prediction. pEffect was developed using a positive data set comprising type III effectors extracted from the literature and UniProt (UniProt Consortum, 2012) and a negative data set combining bacterial non-effector proteins and eukaryotic proteins sequence similar to bacterial effectors. It attains 87±7% accuracy at 95±5% coverage in predicting type III effectors, significantly outperforming its components (PSI-BLAST and SVM). When tested on sequence fragments similar in length to shotgun sequencing reads, pEffect's performance was not significantly different. This result suggests that the information required for distinguishing effectors is not confined to any particular part of the amino acid sequence. Our method provides a basis for the identification of exported pathogenic proteins as targets for future therapeutic treatments. We also suggest using pEffect as a starting point for studies of interactions within microbial communities, detected directly from metagenomic reads and without need for individual genome assembly.

Methods

Development data sets. Our positive data set of known type III effector proteins was extracted from scientific publications (Angot *et al.*, 2007; Arnold *et al.*, 2009; Chang *et al.*, 2005; Greenberg and Vinatzer, 2003; Gurlebeck *et al.*, 2006; Guttman *et al.*, 2002; Miao and Miller, 2000; Sato and Frank, 2004; Tobe *et al.*, 2006) and the Pseudomonas-Plant Interaction web site (http://www.pseudomonas-syringae.org/). The corresponding amino acid sequences were taken from the UniProt database (UniProt Consortum, 2012), 2012_01 release. We additionally queried UniProt with keywords 'type III effector', 'type three effector' and 'T3SS effector' and manually curated the results for experimentally identified effectors. In total, our positive (effector) data set contained 1,388 proteins.

To compile our negative data set of non-type III effectors we used the experimentally annotated Swiss-Prot proteins (Bairoch and Apweiler, 2000) from the 2012_01 UniProt release. We extracted all bacterial proteins that were NOT annotated as type III effectors and had no significant sequence similarity (BLAST (Altschul *et al.*, 1990) *e*-value > 10) to any type III effector in our positive set. We also added all eukaryotic proteins applying no sequence similarity filters. Our negative set thus contained roughly 470,000 proteins.

We removed from our sets all proteins that were annotated as 'uncharacterized', 'putative', or 'fragment'. We reduced sequence redundancy independently in each set using UniqueProt (Mika and Rost, 2003), ascertaining that no pair of proteins in one set had alignment length of less than 35 residues or a positive HSSP-value (HVAL≥0) (Rost, 1999; Sander and Schneider, 1991). After redundancy reduction our sequence-unique sets contained 115 type III effector proteins from 43 different bacterial species and 3,460 non-effector proteins (of which 37% were bacterial). Note that proteins from positive and negative sets are sometimes similar as homology reduction was only applied *within* sets and not *across* sets. Here, this set of sequences (positive and negative sets together) is termed the *Development set*. All pEffect performance results were compiled on stratified cross-validation of this Development set (five-fold cross-validation, *i.e.* we split the entire set into five similarly-sized subsets and trained five models, each on a different combination of four of these subsets, testing each model on every subset exactly once).

Additional data sets. Comparing pEffect performance to that of other methods using our cross-validation approach has only limited value due to the possible overlap between our testing and other methods' development/training sets. A more meaningful way is to use non-redundant sets of effector and non-effector proteins that have never been used for the development of any method. Toward this end, we extracted the following data sets:

(1) We collected all type III effectors added to UniProt between releases 2012_01 and 2014_08 and non-type III bacterial and eukaryotic proteins added between the corresponding releases of Swiss-Prot. These were redundancy reduced at HVAL<0 to produce the UniProt'14_{HVAL0} test set (107 effectors and 1,159 non-effectors). Note that additionally reducing this set to be sequence dissimilar to the *Development set* would retain only 30 type III effectors, too few for reliable performance estimates. However, even for this smaller and completely independent set, the performance of pEffect was higher than of other tools, making pEffect a uniquely reliable method for determining new effectors (Supplementary Table S1).

(2) To answer the question “how well will pEffect perform on protein sequences added to databases within the next six months?” we collected the proteins added to UniProt (type III effectors) and Swiss-Prot (non-effector bacterial and eukaryotic sequences) after the 2014_08 release, producing the set *UniProt'15*$_{Full}$ (498 effectors and 1,509 non-effectors).

(3) We also extracted all bacterial type III effectors from the T3DB database (Wang *et al.*, 2012) – *T3DB*$_{Full}$ set (218 effectors and 831 non-effectors). We deliberately kept the redundancy in this set (up to HVAL = 66, *i.e.* over 85% pairwise sequence identity over 450 residues aligned).

(4) Finally, we redundancy reduced T3DB set at HVAL<0. This gave the *T3DB*$_{HVAL0}$ set (66 effectors and 128 non-effectors).

T3DB Ortholog clusters of the type III secretion system (T3SS) machinery. T3DB is a database of experimentally annotated T3SS-related proteins in 36 bacterial taxa. Proteins of the same function and the same evolutionary origin are clustered in T3DB into *T3 Ortholog* clusters (http://biocomputer.bio.cuhk.edu.hk/T3DB/T3-ortholog-clusters.php). The proteins of these clusters form ten components of the T3SS. Proteins of five of these components (export apparatus, inner membrane ring, outer membrane ring, cytoplasmic ring, and ATPase) are present in all 36 taxa in T3DB (Supplementary Table S2). We thus defined the minimum number of five components necessary for the formation of the T3SS machinery. With the exception of the outer membrane ring, these components have also been defined as the core before (McCann and Guttman, 2008).

Prediction methods. We tested several ideas for prediction, including the following.

Homology-based inference. We transferred type III effector annotations by homology using PSI-BLAST (Altschul *et al.*, 1997) alignments. For every query sequence we generated a PSI-BLAST profile (two iterations, inclusion threshold *e*-value ≤ 10^{-3}) using an 80% non-redundant database combining UniProt (Bairoch and Apweiler, 2000) and PDB (Berman *et al.*, 2000). We then aligned this profile (inclusion *e*-value ≤ 10^{-3}) against all type III effectors extracted from the literature and the UniProt 2012_01 release. For known effectors, we excluded the PSI-BLAST self-hits. We transferred annotation to the query protein from the hit with highest pairwise sequence identity of all retrieved alignments.

De novo prediction. We used the WEKA (Frank *et al.*, 2004) Support Vector Machine (SVM) (Cortes and Vapnik, 1995) implementation to discriminate between type III effector and non-effector proteins. For each protein sequence, we created a PSI-BLAST profile (as described above) and applied the Profile Kernel function (Hamp *et al.*, 2013; Kuang *et al.*, 2004) to map the profile to a vector indexed by all possible subsequences of length *k* from the alphabet of amino acids; we found that *k* = 4 amino acids provides best results. Each element in the vector represents one particular *k*-mer and its score gives the number of occurrences of this *k*-mer that is below a certain user-defined threshold σ; we found that σ = 7 provides best results. This score is calculated as the ungapped cumulative substitution score in the corresponding sequence profile. Thus, the dot product between two *k*-mer vectors reflects the similarity of two protein sequence profiles. Essentially, the method identifies those stretches of *k* adjacent residues in profiles of type III effectors that are most informative for prediction and matches these to the profile of a query protein. The parameters for the SVM and the kernel function were determined separately for each fold in our 5-fold cross-validation and, thus, were never optimized for the test sets.

pEffect. Our final method, pEffect, combined sequence similarity-based and *de novo* predictions. Toward this end, over-fitting was avoided through the simplest possible combination: if any known type III effector is sequence similar to the query use this (similarity-based prediction), otherwise use the *de novo* prediction.

*Reliability index***.** The strength of a pEffect prediction is represented by a reliability index (RI) ranging from 0 (weak prediction) to 100 (strong prediction). For *de novo* predictions, we computed RI by multiplying the SVM output by 100 for positive (type III effector) predictions and subtracted this score from 100 for negative predictions. For sequence similarity-based inferences, the RI is the percentage of pairwise sequence identity normalized to the interval [50, 100], to agree with the SVM prediction range.

Existing methods. We benchmarked pEffect against three state-of-the-art publicly available methods for type III effector prediction, using their default parameters: BPBAac (Wang *et al.*, 2011), Effective T3 (Arnold *et al.*, 2009) and T3_MM (Wang *et al.*, 2013) (Supplementary Section S1).

Evolutionary distances. For the discovery of novel type III effectors in entirely sequence organisms, we extracted evolutionary distances from the phylogenetic tree of 2,966 bacterial and archaeal taxa, inferred from 38 concatenated genes and available in the Newick format (Lang *et al.*, 2013).

Results

pEffect succeeded linking homology-based and *de novo* predictions. Most functional annotations of new proteins originate from homology-based transfer, *i.e.* on the basis of their homology (shared ancestry) to proteins with experimental characterization. For type III effector prediction, homology-based inference implies finding a sequence-similar experimentally annotated type III effector (Methods).

The accuracy of homology-based inference by PSI-BLAST was comparable to that of our *de novo* prediction method on the cross-validation development set (Table 1: 91% *vs.* 92%). However, at this level of accuracy, its coverage was significantly higher (Table 1: 84% vs. 60%). This result encouraged combining these two approaches as introduced in our recent work, LocTree3 (Goldberg *et al.*, 2014): use PSI-BLAST when sequence similarity suffices (e-value ≤ 10^{-3}; Table 1: F_1 = 0.87 complete set) and the SVM otherwise (Table 1: F_1 = 0.67 on subset of proteins without PSI-BLAST hit). The combined method, pEffect, outperformed both its components, reaching an F_1 measure of 0.91 (Table 1).

>>> Table 1 <<<

pEffect outperformed other methods. We compared pEffect to publicly available methods: BPBAac (Wang *et al.*, 2011), Effective T3 (Arnold *et al.*, 2009) and T3_MM (Wang *et al.*, 2013). In contrast to pEffect, all these methods focus exclusively on N-terminal features (Supplementary Section S1). *BPBAac and T3_MM* rely solely on amino acid composition, while *Effective T3* combines amino acid composition and secondary structure information. We compared performance for UniProt proteins that had NOT been used to develop any method, and for T3DB proteins, some of which all methods (incl. pEffect) had used for development. In our hands, pEffect significantly outperformed its competitors on all data sets (Figure 1, Supplementary Table S3). The F_1 performance of pEffect exceeded the other methods by more than 0.58 when tested on any data set with eukaryotic proteins (ΔF_1 = (pEffect, T3_MM) = 0.58 for both UniProt sets, Supplementary Table S3). Thus, pEffect excelled over existing tools in distinguishing type III effectors from bacteria (F_1>0.64*) and from eukaryotes* (F_1>0.85). This improvement is particularly important to, *e.g.*, annotate results from metagenomic studies (Zhou *et al.*, 2014).

>>> Fig. 1 <<<

pEffect excelled even for protein fragments. To evaluate pEffect's ability to annotate effectors from incomplete genomic assemblies and mistakes, we fragmented the proteins from the $T3DB_{Full}$ set, *i.e.* the data set for which other methods were at their best (Figure 1, Supplementary Table S4). We started with protein rather than gene sequence fragments because we did not expect incorrect gene translations of DNA reads, even if sufficiently long, to trigger incorrect effector predictions from any method. Four different approaches were used to generate protein fragments: (i) remove the first 30 residues (N-terminus) from the full protein sequence, (ii) remove the last 30 residues (C-terminus), (iii) randomly remove residues from N- and C- terminus until two thirds of protein are left, and (iv) randomly choose from each protein a single fragment of a typical translated read length (Supplementary Figure S1).

pEffect outperformed all other methods for all fragment sets (i-iv). All methods performed best fragments with C-terminal cleavage (set ii, Figure 1, Supplementary Table S4). Performance was lowest for random fragments of typical read lengths (set iv). However, pEffect performed almost equal to full-length sequences on this set (F_1 = 0.67 on set iv vs. F_1 = 0.69 on full length, Supplementary Table S4). For all fragment sets the pEffect and PSI-BLAST performances were within the standard error of what was obtained using full-length sequences ($T3DB_{Full}$ set; Figure 1, Supplementary Table S3). Furthermore, for smaller protein fragments (sets iii and iv) using *de novo* prediction in addition to PSI-BLAST did not improve pEffect. These results suggest that the features distinguishing type III effectors are spread over the entire protein sequence and are picked up by local alignment, *i.e.* PSI-BLAST.

Reliability index identified confident predictions. pEffect provides a reliability index (RI) to measure the confidence of a prediction; the value of RI ranges from 0 (uncertain) and 100, (most reliable). For PSI-BLAST searches, RIs are normalized values of percentage pairwise sequence identities read of the alignments. For *de novo* predictions, RIs are values corresponding to SVM scores (Methods). Including predictions with low RIs gives many trusted results at reduced accuracy. Higher accuracy predictions are obtained by sampling at higher RIs, thus reducing the total number of trusted samples. For example, at the threshold of RI≥50, over 87% of all predictions of type III effectors are correct and 95% of all effectors in our set are identified (Figure 2: black arrow). On the other hand, at RI>80 effector predictions are correct 96% of the time, but only 78% of all effectors in the set are identified (Figure 2: gray arrow). Thus, users can choose the most appropriate threshold for a given study. Users can also focus on previously unidentified effectors (*de novo* predictions) or, *vice versa*, on validated homologs of known effectors (PSI-BLAST matches; Supplementary Figure S2).

>>> Fig. 2 <<<

Scanning proteomes for type III effector proteins. We used pEffect to annotate type III effectors in 862 bacterial (274 Gram-positive and 588 Gram-negative bacteria) and 90 archaeal proteomes from the European Bioinformatics Institute (EBI: http://www.ebi.ac.uk/genomes/). Our predictions are available at the pEffect website (http://www.bromberglab.org/services/pEffect/proteomes).

Each bacterium was predicted to contain at least one type III effector (Figure 3; Supplementary Table S5, a minimum of 0.8% - 2 out of all 240 proteins in a proteome are predicted as effectors). For some Gram-negative bacteria over 750 type III effectors were predicted (*e.g. Sorangium cellulosum* So ce56 – 1,207 effectors, *Stigmatella aurantiaca* DW4/3-1 – 870, *Corallococcus coralloides* DSM 2259 – 826 and *Haliangium ochraceum* DSM 14365 - 792, Supplementary Table S5). *Stigmatella aurantiaca* DW4/3-1 is hypothesized to have a type III secretion system (T3SS) and effectors (Konovalova *et al.*, 2010). We could not find any literature record for the other three species.

Overall, the number of predicted type III effectors was 1% to 10% of the whole proteome in Gram-positive bacteria, and 1% to 15% in Gram-negative bacteria (Figure 3, Supplementary Table S5). To further understand our predictions, we retrieved UniProt keywords of predicted effectors. Their annotations varied widely, with the most common for both types of bacteria being *transferase*, depicting a large class of enzymes that are responsible for the transfer of specific functional groups from one molecule to another, *nucleotide-binding*, a common functionality of effector proteins, *ATP-binding* that is also an essential component of T3SS, and *kinase*, which is necessary for the expression of T3SS genes. About one fourth (26-29% per proteomes) of predicted type III effectors are functionally 'unknown' (Supplementary Table S6).

We also predicted type III effectors in all archaeal proteomes, with over 100 effectors identified in the proteomes of *Haloterrigena turkmenica* DSM 5511 and *Methanosarcina acetivorans* C2A (126 and 105 effectors, respectively; Supplementary Table S5). On average, there were fewer effectors predicted in archaea than in bacteria: 1.9% is the overall per-organism number for archaea *vs.* 3.4% for Gram-positive and 4.6% Gram-negative bacteria (Figure 3). The most frequent annotations of predicted archaeal effectors were similar to those for predicted bacterial effectors, namely 'unknown', nucleotide-binding, ATP-binding and transferase (Supplementary Table S6). We address the unexpected predictions of effectors in Archaea further in the Discussion section.

>>> Fig. 3 <<<

T3SS is most likely to exist in organisms with ≥5% predicted effectors and five type III machinery components. We BLASTed proteins representative of five T3DB Ortholog clusters (*e*-value ≤ 10^{-3}; Supplementary Table S2) against the full proteomes of our 862 bacteria and 90 archaea set. We thus aimed to identify those proteomes likely equipped with the type III secretion system (T3SS) machinery (Figure 4).

>>> Fig. 4 <<<

We found that, as expected, archaea never contain a full T3SS (maximum three out of five components). In Gram-negative bacteria, the number of predicted effectors correlated much better with the number of type III machinery components (Pearson correlation r = 0.37) than in Gram-positive bacteria (r = 0.13). The combination of a high percentage of predicted type III effectors and a high number of conserved type III machinery components provides strong evidence for the presence of the type III secretion abilities (Figure 4). As a rule of thumb, based on our observations in archaea and Gram-positive bacteria, we suggest that these abilities can be reliably identified by the presence of the complete T3SS and ≥5% of the genome dedicated to effectors. With these cutoffs, 20% (120 species) of the Gram-negative bacteria in our set are identified as type III secreting. No archaeal species and only five Gram-positive bacteria fit these cutoffs. We searched the literature for annotation of ten randomly chosen Gram-negative bacteria from this set (Supplementary Table S7). We found evidence of type III machinery in seven of the ten organisms (Attree and Attree, 2001; Bertelli *et al.*, 2010; Block *et al.*, 2010; Brugirard-Ricaud *et al.*, 2004; Dai and Li, 2014; Mavrodi *et al.*, 2011; Salinero *et al.*, 2009). For three bacteria the secretion machinery has not been studied. Overall, our results suggest that the experimental annotation of the type III secretion in isolated and cultured organisms is incomplete, leaving significant room for improvement.

Discussion

pEffect combines homology-based and *de novo* prediction. PSI-BLAST is commonly used to annotate protein function through sequence similarity (Radivojac *et al.*, 2013). Applied to our sequence unique *Development* set, PSI-BLAST correctly annotated most type III effector proteins (F_1 = 0.87 ± 0.09) through sequence comparisons against a set of known type III effectors. The *de novo* prediction with the profile kernel SVM annotated type III effectors slightly worse (F_1 = 0.73 ± 0.11). Our new method, pEffect, successfully combined the complementary homology-based and *de novo* predictions, reaching sustained high levels of performance (F_1 = 0.91 ± 0.08), better than each of its individual components (Table 1).

Predictions succeed even for fragments from metagenomic analyses. pEffect distinguishes type III effectors from other bacterial and eukaryotic proteins using either full length proteins or protein fragments. The detection of N-terminal signals, often used as the only source of evidence for predicting type III effectors computationally, presents a special problem for metagenomic data because of the erroneous gene predictions and potentially absent reads in contig assemblies. For all fragment sets tested, pEffect performed within one standard error of the level for full-length sequences (Figure 1, Supplementary Tables S3-S4). This result suggests that the features distinguishing type III effectors are present throughout the protein sequence and are not solely confined to the N-terminal region.

The finding that the secretion signals is somehow ‘distributed’ over the entire protein was surprising and extremely relevant for the analysis of metagenomic read data. Deep Sequencing (or NGS) produces immense amounts of DNA reads, which need to be assembled and annotated to be useful. Erroneous (chimeric) gene assemblies or wrong gene predictions are common in sequencing projects (Nielsen and Krogh, 2005). To bypass the assembly errors when identifying type III secretion activity in a particular metagenomic sample it would help to annotate effectors from raw protein fragments translated directly from the DNA reads. Since pEffect succeeds in our tests on fragments, our new method might just enable such a direct analysis. To ultimately establish this point, we will have to compare predictions from raw read translations to those from translations of assembled genomes. Clearly, the results from pEffect can help establish the presence or absence of pathogenic organisms in a particular environment.

Most predictions are *de novo* without sequence similarity to known effectors. Type III effectors were predicted in all types of prokaryotes that we tested. As expected, the number of effectors in Gram-positive bacteria and archaea that are not known to utilize T3SS was lower than in Gram-negative bacteria that do use the system (Figures 3-4). Interestingly, homology searches, *i.e.* PSI-BLAST results, have identified roughly equal numbers of effectors (1%; Figure 5, Supplementary Table S5) in both types of bacterial genomes. As some effectors often co-localize with the T3SS machinery in “pathogenicity islands” (Figueira and Holden, 2012; Okada *et al.*, 2009; Reis and Horn, 2010), these findings are in line with the inheritance of the early complete secretory system, including the machinery and the secreted proteins.

>>> Fig. 5 <<<

Overall, the percentage of effectors predicted by sequence similarity (homology-based) ranged from 3%-71% for bacteria with an average of 21% (maximum for *Onion yellows phytoplasma* OY-M, an intracellular Gram-negative plant pathogen (Oshima *et al.*, 2013); Supplementary Table S5). Conversely, a significantly larger fraction (on average ~76%) of all effector predictions were based on our *de novo* method, *i.e.* could not have been identified without machine learning. The percentage of *de novo* predictions in Gram-negative bacteria was significantly larger than in Gram-positive ones (79±0.4% *vs*. 70±0.5%, respectively; Figure 3). Note, however, that 70% is still a drastically large fraction to appear in bacteria that seemingly have no use for them. Furthermore, the number of “new” effectors has grown over evolutionary time (Figure 5), suggesting functional innovation due to environmental pressures. The set of *de novo*-identified effectors found across bacteria is thus a good starting point for further investigation into effector origins.

Highest number of effectors in Gram-negative bacteria with full T3SS. The loss of type III secretion components in Gram-negative bacteria is accompanied by the loss of effectors, indicating the lack of necessity to further diversify in the absence of the complete machinery (Figure 4C). This type of correlation between the completeness of T3SS and the number of effectors in Gram-negative bacteria is not present for non-type III secreting Gram-positive bacteria (Figure 4B) or archaea (Figure 4A).

Further insight into evolution of bacterial T3SS. pEffect’s high prediction accuracy raises an interesting question about its predictions of effectors in Gram-positive bacteria, which is not known to utilize T3SS. Roughly one fourth of their predicted effectors are of yet-unknown function. Bacterial proteins of annotated function are mostly transferases, hydrolases, ATP-binding proteins or kinases (Supplementary Table S6), all of which are necessary for flagellar motility. This finding is in line with evidence of shared ancestry between bacterial flagellar and type III secretion systems (McCann and Guttman, 2008). It is not known whether T3SS evolved from the flagellar apparatus or if the two systems evolved in parallel. However, gene genealogies (Gophna *et al.*, 2003) and protein network analysis approaches (Medini *et al.*, 2006) both suggest independent evolution from a common ancestor, which comprised a subset of proteins forming a membrane-bound complex. The fact that the flagellar system can also secrete proteins (Macnab, 2004) suggests that this ancestor may have played a secretory role (McCann and Guttman, 2008). The pervasiveness of the flagellar apparatus across the bacterial space suggests that the ancestral complex existed prior to the split of the cell-walled and double-membrane organisms, indicated by the differences in gram staining. The common ancestor protein complex of T3SS and flagellar system would have then been encoded in an even earlier ancestral genome. Thus, it is not surprising that we find T3SS component homology in Gram-positive bacteria even in the absence of type III secretion functionality. Interestingly, our results show that the loss of the complete T3SS and, inherently, the associated loss in type III functionality has proceeded at a roughly similar rate in Gram-positive and Gram-negative bacteria (Figure 6A); *i.e.* once the T3SS is incomplete (4 components), and arguably non-functional, further loss of components consistently follows. A complete T3SS, however, is only visible in early Gram-positive bacteria, but preserved across time in Gram-negative bacteria (Figure 6B), further confirming the presence of the ancestral secretory complex in the last common bacterial ancestor.

>>> Fig. 6 <<<

Did T3SS exist before the archaea/bacteria split? pEffect also predicts a significant number of effectors in archaea. However, the presence of the beginnings of T3SS in the common ancestor of bacteria and archaea is neither directly supported nor negated by our results. Archaeal flagella have little or no structural similarities to bacterial flagella, but share homology with the type IV secretion system (Ng *et al.*, 2006). Some of the type IV secretion system and T3SS components are homologous, *e.g.* VirB11-like ATPases (Wallden *et al.*, 2010). However, despite this observed homology none of the archaea that we tested had the complete set of T3SS components (Figure 3). If the common ancestor of archaea and bacteria did encode the core ancestral complex, these observations would indicate a loss of functionality in archaea. Another possibility is that the T3SS in bacteria, like the flagellar apparatus (Liu and Ochman, 2007), may have been built over time from duplicated and diversified paralogous genes of the core complex after the archaea/bacteria split. In both of these scenarios, the prediction of type III effectors in archaea would then indicate re-purposing of the proteins secreted by the ancestral complex. In fact, 0.5% of an average archaeal genome is identified by homology (PSI-BLAST) to known effectors and another 0.9% *de novo* identified proteins are homologous (PSI-BLAST e-value $\leq 10^{-3}$) to predicted effectors of Gram-negative bacteria. These proteins must have been re-purposed in modern archaea; they are usually annotated as hydrolases, transferases, and metal-binding proteins (Supplementary Table S6). The use of an additional 0.5% of the archaeal proteome that is picked up by pEffect *de novo* and has no homologs in bacteria remains an enigma. While a certain level of similarity exists between archaeal proteins and bacterial type III effectors and machinery, the observed signal is insufficient to draw definitive conclusions regarding common ancestry. It is, however, significant for further exploration – if roughly one tenth of the identified effectors of Gram-negative bacteria and half of the machinery have homologs in archaea, could there have been a common ancestral secretion complex that has developed early on in evolutionary time and has given root to many systems observed today?

Availability

pEffect is accessible as an online web server (http://www.bromberglab.org/services/pEffect). Proteome scanning results, described in this manuscript, are also available for download. We expect our method framework to improve in the future as more experimental data and more sequences become available. However, pEffect's high levels of accuracy and its ability to easily handle large-scale data already place the method at the ideal starting point for annotating type III effector functionality of individual proteins, whole proteomes, or even translated metagenomes.

Acknowledgements

Thanks to Tim Karl, Guy Yachdav, Laszlo Kajan (all TUM) and Yannick Mahlich (Rutgers) for invaluable help with hardware and software; to Chengsheng Zhu (Rutgers) for helpful discussions; to Jessie Maguire (Rutgers), Marlena Drabik, Inga Weise and Lothar Richter (all TUM) for administrative support. Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases. The work was supported by a grant from the Alexander von Humboldt foundation through the German Federal Ministry for Education and Research (BMBF). Additional funding was provided to Tatyana Goldberg through Ernst Ludwig Ehrlich Studienwerk (ELES).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25:** 3389-3402.

Angot A, Vergunst A, Genin S, Peeters N (2007). Exploitation of eukaryotic ubiquitin signaling pathways by effectors translocated by bacterial type III and type IV secretion systems. *PLoS Pathog* **3:** e3.

Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S *et al* (2009). Sequence-based prediction of type III secreted proteins. *PLoS Pathog* **5:** e1000376.

Attree O, Attree I (2001). A second type III secretion system in Burkholderia pseudomallei: who is the real culprit? *Microbiology* **147:** 3197-3199.

Bairoch A, Apweiler R (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research* **28:** 45-48.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H *et al* (2000). The Protein Data Bank. *Nucleic acids research* **28:** 235-242.

Bertelli C, Collyn F, Croxatto A, Ruckert C, Polkinghorne A, Kebbi-Beghdadi C *et al* (2010). The Waddlia genome: a window into chlamydial biology. *PloS one* **5:** e10890.

Block A, Guo M, Li G, Elowsky C, Clemente TE, Alfano JR (2010). The Pseudomonas syringae type III effector HopG1 targets mitochondria, alters plant development and suppresses plant innate immunity. *Cellular microbiology* **12:** 318-330.

Brugirard-Ricaud K, Givaudan A, Parkhill J, Boemare N, Kunst F, Zumbihl R *et al* (2004). Variation in the effectors of the type III secretion system among Photorhabdus species as revealed by genomic analysis. *Journal of bacteriology* **186:** 4376-4381.

Buttner D, He SY (2009). Type III protein secretion in plant pathogenic bacteria. *Plant physiology* **150:** 1656-1664.

Chang JH, Urbach JM, Law TF, Arnold LW, Hu A, Gombar S *et al* (2005). A high-throughput, near-saturating screen for type III effector genes from Pseudomonas syringae. *Proc Natl Acad Sci U S A* **102:** 2549-2554.

Cornelis GR (2006). The type III secretion injectisome. *Nature reviews Microbiology* **4:** 811-825.

Cortes C, Vapnik V (1995). Support-vector networks. *Machine learning* **20:** 273-297.

Dai W, Li Z (2014). Conserved type III secretion system exerts important roles in Chlamydia trachomatis. *International journal of clinical and experimental pathology* **7:** 5404-5414.

Figueira R, Holden DW (2012). Functions of the Salmonella pathogenicity island 2 (SPI-2) type III secretion system effectors. *Microbiology* **158:** 1147-1161.

Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004). Data mining in bioinformatics using Weka. *Bioinformatics* **20:** 2479-2481.

Ghosh P (2004). Process of protein transport by the type III secretion system. *Microbiology and molecular biology reviews : MMBR* **68:** 771-795.

Goldberg T, Hecht M, Hamp T, Karl T, Yachdav G, Ahmed N *et al* (2014). LocTree3 prediction of localization. *Nucleic acids research* **42:** W350-355.

Gophna U, Ron EZ, Graur D (2003). Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events. *Gene* **312:** 151-163.

Greenberg JT, Vinatzer BA (2003). Identifying type III effectors of plant pathogens and analyzing their interaction with plant cells. *Curr Opin Microbiol* **6:** 20-28.

Gurlebeck D, Thieme F, Bonas U (2006). Type III effector proteins from the plant pathogen Xanthomonas and their role in the interaction with the host plant. *J Plant Physiol* **163:** 233-255.

Guttman DS, Vinatzer BA, Sarkar SF, Ranall MV, Kettler G, Greenberg JT (2002). A functional screen for the type III (Hrp) secretome of the plant pathogen Pseudomonas syringae. *Science* **295:** 1722-1726.

Hamp T, Goldberg T, Rost B (2013). Accelerating the Original Profile Kernel. *PloS one* **8:** e68459.

Holland IB, Schmitt L, Young J (2005). Type 1 protein secretion in bacteria, the ABC-transporter dependent pathway (review). *Molecular membrane biology* **22:** 29-39.

Hueck CJ (1998). Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiology and molecular biology reviews : MMBR* **62:** 379-433.

Konovalova A, Petters T, Sogaard-Andersen L (2010). Extracellular biology of Myxococcus xanthus. *FEMS microbiology reviews* **34:** 89-106.

Kuang R, Ie E, Wang K, Siddiqi M, Freund Y, Leslie C (2004). Profile-based string kernels for remote homology detection and motif extraction. *Proc IEEE Comput Syst Bioinform Conf***:** 152-160.

Lang JM, Darling AE, Eisen JA (2013). Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. *PloS one* **8:** e62510.

Leo JC, Grin I, Linke D (2012). Type V secretion: mechanism(s) of autotransport through the bacterial outer membrane. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* **367:** 1088-1101.

Liu R, Ochman H (2007). Stepwise formation of the bacterial flagellar system. *Proceedings of the National Academy of Sciences of the United States of America* **104:** 7116-7121.

Low HH, Gubellini F, Rivera-Calzada A, Braun N, Connery S, Dujeancourt A *et al* (2014). Structure of a type IV secretion system. *Nature* **508:** 550-553.

Lower M, Schneider G (2009). Prediction of type III secretion signals in genomes of gram-negative bacteria. *PloS one* **4:** e5917.

Macnab RM (2004). Type III flagellar protein export and flagellar assembly. *Biochimica et biophysica acta* **1694:** 207-217.

Marshall NC, Finlay BB (2014). Targeting the type III secretion system to treat bacterial infections. *Expert opinion on therapeutic targets* **18:** 137-152.

Mavrodi DV, Joe A, Mavrodi OV, Hassan KA, Weller DM, Paulsen IT *et al* (2011). Structural and functional analysis of the type III secretion system from Pseudomonas fluorescens Q8r1-96. *Journal of bacteriology* **193:** 177-189.

McCann HC, Guttman DS (2008). Evolution of the type III secretion system and its effectors in plant-microbe interactions. *The New phytologist* **177:** 33-47.

McDermott JE, Corrigan A, Peterson E, Oehmen C, Niemann G, Cambronne ED *et al* (2011). Computational prediction of type III and IV secreted effectors in gram-negative bacteria. *Infection and immunity* **79:** 23-32.

Medini D, Covacci A, Donati C (2006). Protein homology network families reveal step-wise diversification of Type III and Type IV secretion systems. *PLoS computational biology* **2:** e173.

Miao EA, Miller SI (2000). A conserved amino acid sequence directing intracellular type III secretion by Salmonella typhimurium. *Proc Natl Acad Sci U S A* **97:** 7539-7544.

Mika S, Rost B (2003). UniqueProt: Creating representative protein sequence sets. *Nucleic acids research* **31:** 3789-3791.

Ng SY, Chaban B, Jarrell KF (2006). Archaeal flagella, bacterial flagella and type IV pili: a comparison of genes and posttranslational modifications. *Journal of molecular microbiology and biotechnology* **11:** 167-191.

Nielsen P, Krogh A (2005). Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics* **21:** 4322-4329.

Nivaskumar M, Francetic O (2014). Type II secretion system: a magic beanstalk or a protein escalator. *Biochimica et biophysica acta* **1843:** 1568-1577.

Okada N, Iida T, Park KS, Goto N, Yasunaga T, Hiyoshi H *et al* (2009). Identification and characterization of a novel type III secretion system in trh-positive Vibrio parahaemolyticus strain TH3996 reveal genetic lineage and diversity of pathogenic machinery beyond the species level. *Infection and immunity* **77:** 904-913.

Oshima K, Maejima K, Namba S (2013). Genomic and evolutionary aspects of phytoplasmas. *Frontiers in microbiology* **4:** 230.

Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A *et al* (2013). A large-scale evaluation of computational protein function prediction. *Nature methods* **10:** 221-227.

Reis RS, Horn F (2010). Enteropathogenic Escherichia coli, Samonella, Shigella and Yersinia: cellular aspects of host-bacteria interactions in enteric diseases. *Gut pathogens* **2:** 8.

Rost B (1999). Twilight zone of protein sequence alignments. *Protein Eng* **12:** 85-94.

Salinero KK, Keller K, Feil WS, Feil H, Trong S, Di Bartolo G *et al* (2009). Metabolic analysis of the soil microbe Dechloromonas aromatica str. RCB: indications of a surprisingly complex life-style and cryptic anaerobic pathways for aromatic degradation. *BMC genomics* **10:** 351.

Sander C, Schneider R (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9:** 56-68.

Sato H, Frank DW (2004). ExoU is a potent intracellular phospholipase. *Mol Microbiol* **53:** 1279-1290.

Tobe T, Beatson SA, Taniguchi H, Abe H, Bailey CM, Fivian A *et al* (2006). An extensive repertoire of type III secretion effectors in Escherichia coli O157 and the role of lambdoid phages in their dissemination. *Proc Natl Acad Sci U S A* **103:** 14941-14946.

Troisfontaines P, Cornelis GR (2005). Type III secretion: more systems than you think. *Physiology* **20:** 326-339.

UniProt Consortum (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research* **40:** D71-75.

Wallden K, Rivera-Calzada A, Waksman G (2010). Type IV secretion systems: versatility and diversity in function. *Cellular microbiology* **12:** 1203-1212.

Wang Y, Zhang Q, Sun MA, Guo D (2011). High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics* **27:** 777-784.

Wang Y, Huang H, Sun M, Zhang Q, Guo D (2012). T3DB: an integrated database for bacterial type III secretion system. *BMC Bioinformatics* **13:** 66.

Wang Y, Sun M, Bao H, White AP (2013). T3_MM: A Markov Model Effectively Classifies Bacterial Type III Secretion Signals. *PloS one* **8:** e58173.

Zhou Q, Su X, Ning K (2014). Assessment of quality control approaches for metagenomic data analysis. *Scientific reports* **4:** 6957.

Figure captions

Fig. 1: Method performance comparison on independent test sets and protein fragments. Performance (F_1 measure, Supplementary Section S2, Eqn. 3; '±' standard error, Supplementary Section S2, Eqn. 4) was measured for BPBAac (Wang *et al.*, 2011), EffectiveT3 (Arnold *et al.*, 2009), and T3_MM (Wang *et al.*, 2013) methods. We also computed F_1 for *de novo* (SVM-based) predictions alone, PSI-BLAST homology-based look up alone, and pEffect: a combination of PSI-BLAST (if a hit is available) and *de novo* (otherwise). **Panel (A)** shows performance on evaluation data sets (Methods) including UniProt'14$_{\text{HVAL0}}$$^{\boldsymbol{1}}$ (107 effectors and 1,159 non-effector bacterial and eukaryotic proteins, added to UniProt between releases 2012_01 and 2014_08, sequence homology reduced at HVAL<0), UniProt'15$_{\text{Full}}$$^{\boldsymbol{2}}$ (498 effectors and 1,509 non-effector bacterial and eukaryotic proteins added to UniProt after 2012_08 release, NOT homology reduced), T3DB$_{\text{HVAL0}}$$^{\boldsymbol{3}}$ (66 effectors and 128 non-effector bacterial proteins from T3DB database, sequence homology reduced at HVAL<0), and T3DB$_{\text{Full}}$$^{\boldsymbol{4}}$ (218 effectors and 831 non-effector bacterial proteins from T3DB database, NOT homology reduced). **Panel (B)** shows performance on *fragments* produced from T3DB$_{\text{Ful}}$$^{\boldsymbol{4}}$ (Methods) approach i$^{\boldsymbol{5}}$: 30 N-terminal amino acids cleaved off, ii$^{\boldsymbol{6}}$: 30 C-terminal amino acids cleaved off, iii$^{\boldsymbol{7}}$: Randomly selected two thirds of the protein sequence' iv$^{\boldsymbol{8}}$: Randomly selected sequence fragments of typical translated read length (average 110 amino acids, Supplementary Figure 1).

Fig. 2: Reliable predictions are more accurate. The figure shows the cumulative percent of accuracy/coverage (Supplementary Section S2) of pEffect predictions at or above a given reliability index (RI). The graphs were obtained using the homology-reduced *Development set* of 115 type III effector and 3,460 non-effector proteins in five-fold cross-validation. At the reliability score of RI = 50 (black vertical line), 95% of type III effectors are identified at 87% accuracy (black arrow). At a higher reliability score of RI = 80 (gray vertical line), prediction accuracy increases to 97% at the cost of lower coverage of 78% (gray arrow).

Fig. 3: Percentage of predicted effectors in full proteomes. The figure shows the box-plot-and-instance representation of percentages of pEffect-predicted type III effectors (Y-axis) in 90 archaeal, 274 Gram-positive and 588 Gram-negative bacterial organisms (X-axis), which are shown as dots. At least 50% of effector predictions in all, except 11 organisms in our set were predicted *de novo*. In the figure, the color represents the percentage of *de novo* predictions for each organism: from green (50% *de novo,* 50% PSI-BLAST) to blue (100% *de novo*, 0% PSI-BLAST). While effectors predicted in archaea and Gram-positive bacteria are often picked up by PSI-BLAST, effectors in Gram-negative bacteria are mostly *de novo* predictions.

Fig. 4: Proteomes encoding some of the five components of T3SS machinery. (A) 90 archaea proteomes, **(B)** 274 Gram-positive bacteria and **(C)** 588 Gram-negative bacteria were scanned for the presence of T3SS and are shown as dots in the figure. The percentage of type III effectors predicted by pEffect (Y-axis) is compared to the number of type III secretion machinery components (max. five T3 Ortholog clusters; Methods) identified in these proteomes (X-axis). Note that effector predictions are computationally completely independent of machinery component identifications. While type III effectors compose up to 3.7% of an archaeal proteome (mean 1.9%, blue horizontal line), this number is much larger for bacteria, reaching up to 10.1% of an entire proteome for Gram-positive bacteria (mean 3.4%), and 14.9% for Gram-negative bacteria (mean 4.6%). Note that six Gram-negative bacterial species did not contain detectable homologs of any of the required machinery components (not even ATPases), indicating that their genomes are further diverged than those of other species.

Fig. 5: pEffect's whole proteome predictions identified by source. pEffect predicted type III effector proteins in the proteomes of 294 Gram-negative and 29 Gram-positive bacteria having full T3SS. The proteomes are shown as green and blue dots. Green dots indicate the percentage of proteins predicted as effectors (Y-axis) by homology searches (PSI-BLAST) and blue dots are *de novo* predictions. For each proteome, the evolutionary distance from the last common ancestor (X-axis) is extracted from (Lang *et al.*, 2013). While PSI-BLAST appears to consistently pick up ~1% of each proteome of all organisms (green horizontal trend-line), the effectors in Gram-negative bacteria diversify further over evolutionary distance, as indicated by the increase in the number of *de novo* predictions.

Fig. 6: Loss of T3SS functionality differentiates Gram-positive and -negative bacteria. 274 Gram-positive bacteria (blue dots) and 588 Gram-negative bacteria (red dots) are screened for the number of conserved components of T3SS (max. 5 T3DB Ortholog clusters; Methods) in their genomes (Y-axis) and plotted against the evolutionary distance from the most recent common ancestor (X-axis). Once the T3SS is lost **(A)**, *i.e.* less than 5 components are present, further rate of loss of components is the same for all bacteria. The number of Gram-negative bacteria with the complete system **(B)**, *i.e.* all 5 components present, however, remains constant across evolutionary time, while the number of Gram-positive bacteria declines.

Tables

Table 1: Performance of pEffect and its components on the Development set

Method	*TP*	*FN*	*FP*	*TN*	*Acc*[5]	*Cov*[5]	*F_1*[5]
PSI-BLAST[1]	*97*	*18*	*10*	*3450*	*91 ± 7*	*84 ± 8*	*0.87± 0.09*
De novo[2]	*69*	*46*	*6*	*3454*	***92 ± 8***	*60 ± 11*	*0.73 ± 0.11*
De novo$_{No\_PSI\text{-}BLAST\_hit}$[3]	*12*	*6*	*6*	*3444*	*67 ± 25*	*67 ± 28*	*0.67 ± 0.23*
pEffect[4]	*109*	*6*	*16*	*3444*	*87 ± 7*	***95 ± 5***	***0.91 ± 0.08***

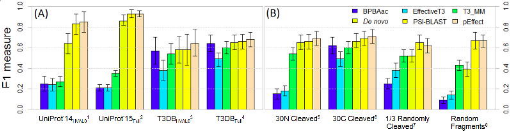
[1] PSI-BLAST: sequence similarity-based inference component of pEffect on all 3,755 proteins in the full Development set.

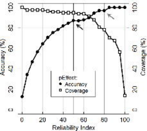
[2] *De novo*: SVM-based prediction component on the full Development set.

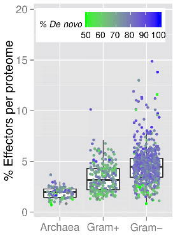
[3] *De novo*$_{No\_PSI\text{-}BLAST\_hit}$: SVM-based prediction component of pEffect tested only on the set of 3,468 proteins that did not align to any effectors using PSI-BLAST.

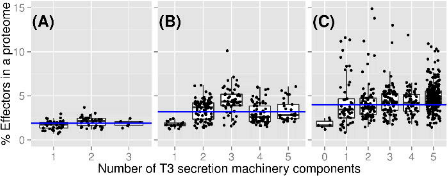
[4] pEffect: PSI-BLAST predictions, if available, and *de novo* otherwise on the full Development set.

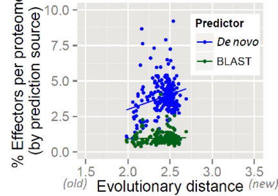
[5] Eqn. 1-4; Acc, accuracy; Cov, coverage; F_1: performance measures; '±' standard errors obtained by re-sampling the predictions (Supplementary Section S2). Highest value in each column in bold.

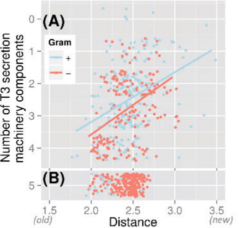

(A)
(B)
F1 measure
0.0
0.2
0.4
0.6
0.8
1.0
BPBAac
De novo
EffectiveT3
PSI-BLAST
T3_MM
pEffect
UniProt'14IVAL0 1
UniProt'15Full 2
T3DBIVAL0 3
T3DBFull 4
30N Cleaved5
30C Cleaved6
1/3 Randomly Cleaved7
Random Fragments6


Accuracy (%)
Coverage (%)
Reliability Index
0
20
40
60
80
100
pEffect:
● Accuracy
□ Coverage


% De novo
50 60 70 80 90 100
% Effectors per proteome
0
5
10
15
20
Archaea
Gram+
Gram−


(A)
(B)
(C)
% Effectors in a proteome
0
5
10
15
1
2
3
1
2
3
4
5
0
1
2
3
4
5
Number of T3 secretion machinery components


% Effectors per proteome (by prediction source)
10.0
7.5
5.0
2.5
0.0
1.5
2.0
2.5
3.0
3.5
(old)
Evolutionary distance
(new)
Predictor
De novo
BLAST


(A)
(B)
Gram
+
–
Number of T3 secretion machinery components
0
1
2
3
4
5
1.5
2.0
2.5
3.0
3.5
(old)
Distance
(new)