Computational prediction shines light on type III secretion origins

Tatyana Goldberg1,2, Burkhard Rost1,3,4 & Yana Bromberg3,5

Type III secretion system is a key bacterial symbiosis and pathogenicity mechanism responsible for a variety of infectious diseases, ranging from food-borne illnesses to the bubonic plague. In many Gram-negative bacteria, the type III secretion system transports effector proteins into host cells, converting resources to bacterial advantage. Here we introduce a computational method that identifies type III effectors by combining homology-based inference with de novo predictions, reaching up to 3-fold higher performance than existing tools. Our work reveals that signals for recognition and transport of effectors are distributed over the entire protein sequence instead of being confined to the N-terminus, as was previously thought. Our scan of hundreds of prokaryotic genomes identified previously unknown effectors, suggesting that type III secretion may have evolved prior to the archaea/bacteria split. Crucially, our method performs well for short sequence fragments, facilitating evaluation of microbial communities and rapid identification of bacterial pathogenicity – no genome assembly required. pEffect and its data sets are available at http://services.bromberglab.org/peffect.
any method, and for T3DB11 proteins, some of which all methods (incl. pEffect) had used for development. In
set. (5) Eqn. 1–4; Acc, accuracy; Cov, coverage; F1: performance measures; ‘de novo’ PSI-BLAST. (4) pEffect: PSI-BLAST predictions, if available, and
prediction component of pEffect tested only on the set of 3,468 proteins that did not align to any effectors using
PSI-BLAST. (1) pEffect outperformed other methods. (2) Method TP FN FP TN Acc(5) Cov(5) F1(5)

| Method                          | TP   | FN   | FP   | TN   | Acc(5) | Cov(5) | F1(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 → PSI-BLAST 34(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 |

Table 1. Performance of pEffect and its components on the Development set. (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 → PSI-BLAST 34: 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; F1: performance measures; ‘±’ standard errors obtained by re-
sampling the predictions (Supplementary S1 Text). Highest value in each column in bold.

analysis approaches21 suggest evolution of both systems from a common ancestor. For archaea common ancestry
is less clear. However, predominance in the number of predicted effectors in Gram-negative bacteria, as opposed
to the number of predicted effectors in Gram-positive bacteria and archaea suggests repurposing of effector-like
proteins independent of organism secretory abilities. In addition to pEffect’s application to evolutionary infer-
ce, we find that the time and T3SS completeness–driven results, which follow expectations for correlation with
quantities of predicted effectors, are reassuring of our method’s performance.

Our method provides a basis for rapid identification of T3SS-utilizing bacteria and their exported effector
proteins as targets for future therapeutic treatments. The method also proposes interesting directions in which the
evolution of bacterial pathogenicity can further be explored. Finally, we suggest using pEffect as a starting point
for studies of interactions within microbial communities, detected directly from metagenomic reads and without
the need for individual genome assembly.

Results
pEffect succeeded linking homology-based and de novo predictions. Most functional annota-
tions of new proteins originate from homology-based transfer, i.e. on the basis of shared ancestry with experi-
mentally characterized proteins22. For type III effector prediction, homology-based inference implies finding a
sequence-similar experimentally annotated type III effector (‘Methods’ section), e.g. via PSI-BLAST.

The accuracy of homology-based inference by PSI-BLAST was comparable to that of our de novo predic-
tion method on the cross-validation Development set (Table 1: 91% vs. 92%). However, at this level of accu-
cy, its coverage was significantly higher (Table 1: 84% vs. 60%). This result encouraged combining these two
approaches as introduced in our recent work, LocTree323: use PSI-BLAST when sequence similarity suffices
(Table 1: F1 = 0.87 complete set) and the SVM otherwise (Table 1: F1 = 0.67 on subset of proteins without PSI-BLAST hit). The combined method, pEffect, outperformed both its components, reaching an F1
measure of 0.91 (Table 1).

pEffect outperformed other methods. We compared pEffect to publicly available methods: BPBAac13,
EffectiveT324, T3_MM24, Modlab15 and BEAN 2.014. BPBAac, EffectiveT3, T3_MM and Modlab focus exclu-
sively on N-terminal features, while BEAN 2.0 and pEffect are not confined to those regions only (Methods,
Supplementary S2 Text). BPBAac, T3_MM and Modlab rely solely on amino acid composition; EffectiveT3
combines amino acid composition and secondary structure information; BEAN 2.0 uses BLAST18 and PFAM25
domain searches, evolutionary information encoded in N- and C-termini, as well as information from an inter-
mediate sequence region. We compared performance for UniProt26 proteins that had NOT been used to develop
any method, and for T3DB11 proteins, some of which all methods (incl. pEffect) had used for development. In
our hands, pEffect significantly outperformed its competitors on UniProt sets containing eukaryotic proteins
(Fig. 1, Supplementary Table S1). The F1 performance of pEffect exceeded the other methods by more than 0.35
(ΔF1 = (pEffect, BEAN 2.0) = 0.35 for both UniProt sets, Supplementary Table S1). On the bacterial T3DB data
sets, pEffect performed within one standard error of the prediction performance achieved by its best perform-
ing competitor BEAN 2.0. Thus, pEffect performed as well or better when benchmarked with existing tools in
distinguishing type III effectors from bacteria (F1 > 0.64) and from eukaryotes (F1 > 0.88). This improvement is
particularly important to, e.g. annotate results from contaminated metagenomic studies27.

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 homology reduced T3DBreduced
set containing bacterial proteins only. 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 one third of residues, and (iv) randomly choose from each protein a single fragment of a typical
translated read length (Supplementary Fig. S1).

pEffect compared favourably to all other methods for all fragment sets (i–iv). Most methods performed
best on fragments with C-terminal cleavage (set ii, Fig. 1, Supplementary Table S2). Performance was lowest
for random fragments of typical read lengths (set iv). For pEffect it dipped least (F1 = 0.59 ± 0.14 on set iv vs.
F1 = 0.64 ± 0.14 on full length, Supplementary Table S1). For all fragment sets, performances of homology-based
The number of predicted type III effectors was 1–10% of the whole proteome in Gram-positive bacteria and 1–15% in Gram-negative bacteria (Fig. 2, Supplementary Table S3). 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 – an essential component of T3SS, and kinase – necessary for the expression of T3SS genes. About one fourth (26–29% per proteomes) of predicted type III effectors are functionally ‘unknown’ (Supplementary Table S4).

We also predicted type III effectors in all archaeal proteomes, with over 100 effectors identified in the proteomes of *H. turkenica* DSM 5511 and *M. acetivorans* C2A (126 and 105 effectors, respectively; Supplementary Table S3). 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 (Fig. 2). 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 S4).
Evaluation of predictions for proteomes. We BLASTed proteins representative of five T3DB Ortholog clusters (e-value \( \leq 10^{-3} \); Supplementary Table S5) against the full proteomes of our 862 bacteria and 90 archaea set. We thus aimed to identify those proteomes likely equipped with the T3SS machinery (Fig. 3).

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 (Fig. 3). 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 \( \geq 5\% \) of the genome dedicated to effectors. With these cutoffs, 20\% (115 species) of the Gram-negative bacteria in our set are identified as type III secreting. We randomly picked ten species from these 20\% and found evidence in the literature for T3SS presence for seven of them (Supplementary Table S6). No archaeal species and only five Gram-positive bacteria fit these cutoffs. Note that our rule does not imply that organisms with full T3SS and over 5\% predicted effectors necessarily have the complete ability to use the system. Instead, we suggest that organisms without the necessary components cannot use the system. Overall, our results indicate that the experimental annotation of the type III secretion in isolated and cultured organisms is incomplete, leaving significant room for improvement, possibly with assistance from pEffect.

Finally, we extracted from the HAMAP database available annotations of pathogenicity and symbiotic relationships for 115 Gram-negative bacteria in our set with a complete T3SS and \( \geq 5\% \) of the genome dedicated to effectors. We compared the number of predicted effectors in organisms that infect eukaryotic cells in general and mammalian cells in particular with those that are currently not known to be symbiotic or pathogenic. Note that further manual curation of currently not annotated bacteria still highlights possibility of type III secretion for a large fraction of them. Our analysis showed that while the distributions of numbers of effectors across the different types of bacteria were not significantly different, mammalian pathogens carried, on average, more effectors than pathogens of other taxa. Those, in turn, carried more effectors than bacteria not currently annotated as pathogenic or symbiotic (Supplementary Fig. S4). Thus, we believe that pEffect can be used to pinpoint for future exploration of the type III secretion-mediated pathogenicity of newly sequenced organisms.

Discussion

pEffect successfully combines complementary approaches for the prediction of type III effector proteins: homology-based and \textit{de novo}. Specifically, it uses PSI-BLAST for a high accuracy (precision) mode of prediction and SVM for improved coverage (recall). The resulting single method pEffect outperforms both of its individual components (Table 1) and other methods (Fig. 1, Supplementary Tables S1 and S2). When tested on samples...
contaminated with eukaryotic proteins, pEffect predicts effectors with a performance level that is significantly higher than that of any other currently available method (Fig. 1, Supplementary Tables S1 and S7). Similar to the results of Arnold et al.\textsuperscript{12}, we find that there is no significant difference in performance across different species of bacteria (pEffect: \( F_1 = 0.54 \pm 0.31 \) on a data set with no proteins of the same species shared between training and test sets vs. \( F_1 = 0.91 \pm 0.08 \) on the Development set). pEffect was trained on a sequence homology reduced data set at H\( \text{VAL} = 0 \) (i.e. there is no pair of sequences in our data set with over 20\% sequence identity that have over 250 amino acid residues aligned) that to our knowledge presents the largest and most complete set of effector proteins currently available. The data set can be downloaded from pEffect’s website.

pEffect uses information stored in the entire protein sequence and performs on sequence fragments just as well as on full-length protein sequences (Fig. 1, Supplementary Table S2). This result made us conclude that signals discriminating effector proteins are distributed across the entire protein sequences and are not confined to the N-terminus, as it is currently anticipated. This finding 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\textsuperscript{33}. To bypass the assembly errors when identifying type III secretion activity in a particular metagenomic sample it would help to annotate effectors from peptides translated directly from the DNA reads. pEffect facilitates this type of direct analysis of metagenomic sequence data, establishing the level of type III secretion activity and, by proxy, the endosymbiotic interactions and the potential presence or absence of pathogenic organisms in a particular environment.

We applied pEffect to over 900 prokaryotic proteomes with the aim of annotating those organisms that are likely to utilize a T3SS. We validated our results using three different metrics: (i) percentage of predicted effector proteins per proteome, (ii) evolutionary age of an organism and (iii) the number of conserved T3SS elements. As expected, pEffect predicted a higher percentage of effector proteins per proteome in Gram-negative bacteria with full T3SS (five conserved T3SS elements) than in Gram-positive bacteria and archaea that are not known to utilize the system (Figs 2 and 3). This indicates a possible acquisition of a larger effector repertoire in Gram-negative bacteria, which was unnecessary for other organisms. Incorporating the independently established evolutionary age estimate, effector proteins of T3SS-using Gram-negative bacteria appear to further diversify with the increasing evolutionary distance from the last common ancestor (Fig. 4a). This correlation could not be expected at random, as the age of bacteria and their effector quantities are independently established and are not correlating for other organisms.

Interestingly, homology searches have identified roughly equal numbers of effectors (on average, 1\% of each respective proteome; Supplementary Table S3) across both types of bacteria. As their percentage per proteome remains stable over time (Supplementary Table S3) and as they are found in almost all organisms with PSI-BLAST, we suggest these effectors to be the older ones that had the time to spread throughout different species. On the other hand, the increasing number of new effectors, recognized by the SVM, in relationship to organism age (as long as organism is using T3SS, Fig. 4b), indicates likely new “inventions” that accumulate over time of T3SS use. These results are in line with potential ancestral presence of the early complete secretory system\textsuperscript{10,34}, including
The machinery and the secreted proteins, and further diversification of effectors exclusively in T3SS-utilizing Gram-negative bacteria. The set of de novo-identified effectors found across bacteria is a good starting point for further investigation into effector origins. Due to T3SS significance in pathogenicity of Gram-negative bacteria, the de novo identified effectors are also potentially interesting as drug targets.

\[ \text{pEffect's high prediction accuracy raises an interesting question about its false positive predictions of effectors in Gram-positive bacteria, which is not known to utilize T3SS. Roughly one fourth of these predicted effectors are of yet-unknown function. Those that are annotated include enzymes necessary for flagellar motility (Supplementary Table S6). This finding is in line with evidence of shared ancestry between bacterial flagellar and type III secretion systems. Gene genealogies and protein network analysis approaches suggest independent evolution of both systems from a common ancestor, comprising a set of proteins forming a membrane-bound complex. The fact that the flagellar system can also secrete proteins suggests that this ancestor may have played a secretory role. The pervasiveness of the flagellar apparatus across the bacterial space also 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. Thus, it is not surprising that we find T3SS component homology in Gram-positive bacteria even in the absence of type III secretion functionality. Curiously, our results show that the loss of the type III secretion functionality, indicated by the loss of the complete T3SS, has proceeded at a roughly similar rate in Gram-positive and Gram-negative bacteria (Fig. 5a); i.e. once the T3SS becomes incomplete (4 components) and, arguably, non-functional, further loss of components consistently follows. Notably, a complete T3SS is only visible in early Gram-positive bacteria, but preserved across time in Gram-negative bacteria (Fig. 5b), further confirming the likely presence of the ancestral secretory complex in the last common bacterial ancestor.

\[ \text{pEffect also predicts a significant number of false positive effectors in archaea, inspiring the question: did T3SS exist before the archaea/bacteria split? Unfortunately, the presence of the beginnings of T3SS in the common ancestor of bacteria and archaea is neither directly supported nor negated by our results. Archaeal flagella have little or no structural similarities to bacterial flagella and none of the archaea that we tested had the complete T3SS (Fig. 2). If the common ancestor of archaea and bacteria did encode the core ancestral complex, the latter observation would indicate a loss of functionality in archaea. Another possibility is that the T3SS in bacteria may have been built over time from duplicated and diversified paralogous genes of the core complex after the archaea/bacteria split. In both of these scenarios, the prediction of type III effectors in archaea would indicate re-purposing of the proteins secreted by the ancestral complex.} \]

\[ \text{In fact, 0.5% of an average archaeal genome is identified by homologs in archaea, could there have been a common ancestral secretion complex that has developed early on in evolutionary time and has given root to many systems observed today?} \]
pEffect immediately and importantly contributes to the study of type III secretion mechanisms. It allows for rapid identification of type III secretion abilities within unassembled genomic and metagenomic read data. Moreover, the quantity of identified effectors seems to correspond with bacterial pathogenicity, potentially contributing to the tracking of infectious strains. We believe that pEffect will facilitate future experimental insights in microbiological research and will significantly contribute to our understanding and management of infectious disease.

Methods

Development data sets. Our positive data set of known type III effector proteins was extracted from scientific publications12,36–43 and the Pseudomonas-Plant Interaction web site (http://www.pseudomonas-syringae.org/). We additionally queried Swiss-Prot with keywords ‘type III effector’, ‘type three effector’ and ‘T3SS effector’ and manually curated the results for experimentally identified effectors (removing entries with “potential”, “probable” and “by similarity” annotations). All corresponding amino acid sequences were taken from UniProt26, 2012_01 release. 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 from the 2012_01 UniProt release. We extracted all bacterial proteins that were NOT annotated as type III effector and had no significant sequence similarity (BLAST 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 UniqueProt44, ascertaining that no pair of proteins in one set had alignment length of less than 35 residues or a positive HSSP-value45,46 (HVAL ≥ 0). 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 were 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’ training sets, and can lead to an overestimate of other methods’ performance. 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 after the 2014_02 release and non-type III bacterial and eukaryotic proteins added to Swiss-Prot after the same release. These were redundancy reduced at HVAL < 0 to produce the UniProt T3SS effector set (51 effectors and 691 non-effectors, of which 53% were of bacterial origin).
Note that additionally reducing this set to be sequence dissimilar to the Development set would retain only 10 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 S7).

(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 UniProt15\textsuperscript{HVAL} (498 effectors and 1,509 non-effectors, of which ).

(3) We also extracted all bacterial type III effectors from the T3DB database\textsuperscript{11} – T3DB\textsuperscript{HVAL} 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). Note that some proteins from this set are contained in the training sets of all compared methods, including pEffect.

(4) Finally, we redundancy reduced T3DB set at HVAL < 0. This gave the T3DB\textsuperscript{HIVAL} 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\textsuperscript{9}.

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\textsuperscript{18} alignments. For every query sequence we generated a PSI-BLAST profile (two iterations, inclusion threshold $e$-value $\leq 10^{-5}$) using an 80% non-redundant database combining UniProt\textsuperscript{47} and PDB\textsuperscript{48}. We then aligned this profile (inclusion $e$-value $\leq 10^{-5}$) 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\textsuperscript{49} Support Vector Machine (SVM)\textsuperscript{50} 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\textsuperscript{51} 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 $\sigma$; we found that $\sigma = 7$ 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 $k$ adjacents 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.

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\textsuperscript{52}.

References
1. Holland, I. B., Schmitt, L. & Young, J. Type 1 protein secretion in bacteria, the ABC-transporter dependent pathway (review). Molecular membrane biology 22, 29–39 (2005).
2. Nivaskumar, M. & Franetic, O. Type II secretion system: a magic beanstalk or a protein escalator. Biochimica et biophysica acta 1843, 1568–1577, doi: 10.1016/j.bbamac.2013.12.020 (2014).
3. Cornels, G. R. The type III secretion injectisome. Nature reviews. Microbiology 4, 811–825, doi: 10.1038/nrmicro1526 (2006).
4. Low, H. H. et al. Structure of a type IV secretion system. Nature 508, 550–553, doi: 10.1038/nature13081 (2014).
5. Leo, J. C., Grin, I. & Linke, D. 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, doi: 10.1098/rstb.2011.0208 (2012).
47. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research* **28**, 45–48 (2000).
48. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic acids research* **28**, 235–242 (2000).
49. Frank, E., Hall, M., Trigg, L., Holmes, G. & Witten, I. H. Data mining in bioinformatics using Weka. *Bioinformatics* **20**, 2479–2481, doi: 10.1093/bioinformatics/bth261 (2004).
50. Kuang, R. *et al.* Profile-based string kernels for remote homology detection and motif extraction. *Proc IEEE Comput Syst Bioinform Conf*, 152–160 (2004).
51. Hamp, T., Goldberg, T. & Rost, B. Accelerating the Original Profile Kernel. *PloS one* **8**, e68459, doi: 10.1371/journal.pone.0068459 (2013).
52. Lang, J. M., Darling, A. E. & Eisen, J. A. Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. *PloS one* **8**, e62510, doi: 10.1371/journal.pone.0062510 (2013).

**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. T.G. and B.R. were supported by the Alexander von Humboldt foundation through the German Ministry for Research and Education (BMBF: Bundesministerium fuer Bildung und Forschung). Additional funding was provided to T.G. by the Ernst Ludwig Ehrlich Studienwerk (ELES). Y.B. was supported by the NSF CAREER Award 1553289, NIH U01 GM115486, USDA-NIFA 1015:0228906, and the Technische Universität München – Institute for Advanced Study Hans Fischer Fellowship, funded by the German Excellence Initiative and the European Union Seventh Framework Programme, grant agreement 291763.

**Author Contributions**

T.G. and Y.B. designed the project. T.G. developed the prediction method, its web server, and performed whole proteome predictions. T.G., Y.B and B.R. analysed the results and wrote the manuscript.

**Additional Information**

**Supplementary information** accompanies this paper at http://www.nature.com/srep

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Goldberg, T. *et al.* Computational prediction shines light on type III secretion origins. *Sci. Rep.* **6**, 34516; doi: 10.1038/srep34516 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2016