A peer-reviewed version of this preprint was published in PeerJ on 7 June 2019.

View the peer-reviewed version (peerj.com/articles/7055), which is the preferred citable publication unless you specifically need to cite this preprint.

McDermott JE, Cort JR, Nakayasu ES, Pruneda JN, Overall C, Adkins JN. 2019. Prediction of bacterial E3 ubiquitin ligase effectors using reduced amino acid peptide fingerprinting. PeerJ 7:e7055 https://doi.org/10.7717/peerj.7055

Prediction of bacterial E3 ubiquitin ligase effectors using reduced amino acid peptide fingerprinting

Jason McDermott Corresp., [1, 2] , **John Cort** [1] , **Ernesto Nakayasu** [1] , **Christopher Overall** [3] , **Joshua Adkins** [1]

[1] Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, United States of America

[2] Department of Molecular Microbiology and Immunology, Oregon Health Sciences University, Portland, Oregon, United States

[3] Center for Brain Immunology and Glia, University of Virginia, Charlottesville, United States

Corresponding Author: Jason McDermott
Email address: Jason.McDermott@pnnl.gov

Background. Although pathogenic Gram-negative bacteria lack their own ubiquitination machinery, they have evolved or acquired virulence effectors that can manipulate the host ubiquitination process through structural and/or functional mimicry of host machinery. Many such effectors have been identified in a wide variety of bacterial pathogens that share little sequence similarity amongst themselves or with eukaryotic ubiquitin E3 ligases.

Methods. To allow identification of novel bacterial E3 ubiquitin ligase effectors from protein sequences we have developed a machine learning approach, the SVM-based Identification and Evaluation of Virulence Effector Ubiquitin ligases (SIEVE-Ub). We extend the string kernel approach used previously to sequence classification by introducing reduced amino acid (RAA) alphabet encoding for protein sequences.

Results. We found that 14mer peptides with amino acids represented as simply either hydrophobic or hydrophilic provided the best models for discrimination of E3 ligases from other effector proteins with a receiver-operator characteristic area under the curve (AUC) of 0.90. When considering a subset of E3 ubiquitin ligase effectors that do not fall into known sequence based families we found that the AUC was 0.82, demonstrating the effectiveness of our method at identifying novel functional family members. Recursive feature elimination was used to identify a parsimonious set of 100 RAA peptides that provided good discrimination, and these peptides were found to be located in functionally important regions of the proteins involved in E2 and host target protein binding.Our general approach enables construction of models based on other effector functions. We used SIEVE-Ub to predict seven potential novel E3 ligases from a large set of bacterial genomes. SIEVE-Ub is available for download at https://github.com/biodataganache/SIEVE-Ub[p]

Prediction of Bacterial E3 Ubiquitin Ligase Effectors using Reduced Amino Acid Peptide Fingerprinting

Jason E. McDermott*[1,2], John R. Cort[1], Ernesto Nakayasu[1], Christopher Overall[3], Joshua N. Adkins[1]

[1]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

[2]Department of Molecular Microbiology and Immunology, Oregon Health & Sciences University, Portland, OR, USA

[3]Center for Brain Immunology and Glia, University of Virginia, Charlottesville, Virginia, USA

Corresponding Author:
Jason McDermott
902 Battelle Blvd, PO Box 999, MSIN J4-18, Richland, WA 99352
Email address: Jason.McDermott@pnnl.gov

ABSTRACT

Background. Although pathogenic Gram-negative bacteria lack their own ubiquitination machinery, they have evolved or acquired virulence effectors that can manipulate the host ubiquitination process through structural and/or functional mimicry of host machinery. Many such effectors have been identified in a wide variety of bacterial pathogens that share little sequence similarity amongst themselves or with eukaryotic ubiquitin E3 ligases.

Methods. To allow identification of novel bacterial E3 ubiquitin ligase effectors from protein sequences we have developed a machine learning approach, the SVM-based Identification and Evaluation of Virulence Effector Ubiquitin ligases (SIEVE-Ub). We extend the string kernel approach used previously to sequence classification by introducing reduced amino acid (RAA) alphabet encoding for protein sequences.

Results. We found that 14mer peptides with amino acids represented as simply either hydrophobic or hydrophilic provided the best models for discrimination of E3 ligases from other effector proteins with a receiver-operator characteristic area under the curve (AUC) of 0.90. When considering a subset of E3

ubiquitin ligase effectors that do not fall into known sequence based families we found that the AUC was 0.82, demonstrating the effectiveness of our method at identifying novel functional family members. Recursive feature elimination was used to identify a parsimonious set of 100 RAA peptides that provided good discrimination, and these peptides were found to be located in functionally important regions of the proteins involved in E2 and host target protein binding. Our general approach enables construction of models based on other effector functions. We used SIEVE-Ub to predict seven potential novel E3 ligases from a large set of bacterial genomes. SIEVE-Ub is available for download at https://github.com/biodataganache/SIEVE-Ub.

INTRODUCTION

Assignment of functional annotations for newly sequenced proteomes is accomplished largely through transference of annotations from existing proteins using sequence similarity. Many protein families exist that have shared sequence homology and functional annotation and new members can be identified through established models such as hidden Markov models (HMMs). However, there are many other groups of proteins that have closely related functions but diverse sequences. These groups can be described with multiple models that capture different regions of sequence space but may include members -- that don't have sequence similarity with other members detectable by traditional sequence methods.

Standard methods for developing sequence-based models such as HMMs rely on sequence alignment of family members as a first step. Models are then constructed using variability at specific locations established from those alignments. If sequence alignment is not possible or results in poorly aligned sequences, robust models for functionally related proteins may not exist. In these cases machine learning methods can be used to group proteins with similar function together based on sequence-derived features that do not require alignment. Such applications have developed models for problematic protein functions such as multidrug antibiotic resistance transporters [1] by us, and DNA binding proteins [2], calmodulin-binding proteins [3] and to identify subcellular localization [4] immunogenic regions of proteins [5], and

kinase specificity [6], by others. Our group previously developed a machine learning model to identify substrates of the bacterial type III secretion system, and this and similar models have been successful at identifying novel family members [7-11].

A versatile method for creation of sequence-based features for use in such models is the kmer approach, also known as string kernels. This method has been used in sequence analysis to identify distant homologs [12, 13], nucleotide-based functional features [14], and structural folds [6], and to predict antibody epitopes [15]. A current limitation of this approach is computational. Since the alphabet used by amino acids is 20, the space of possible sequences of length k expands exponentially with k, rendering even shorter kmers of length 6 with 20^6 (64 million) possible features. Additionally, as kmers increase in length they become less common resulting in feature sets that are more distinct for each protein, and thus less likely to reveal underlying relationships. This problem can be addressed using mismatch kernels [13] and similar approaches, but remains a computational and pragmatic barrier. Here we report the use of a kmer-based approach to identification of novel ubiquitin E3 ligases in pathogenic bacteria.

Ubiquitination is an abundant protein post-translation modification (PTM) in eukaryotic cells that controls many key pathways, including controlling protein turnover and innate immune signaling [16, 17]. Ubiquitination is a dynamic and reversible PTM produced by the coordinated action of three enzymes: E1 ubiquitin activating enzyme, E2 ubiquitin conjugating enzyme, and E3 ubiquitin ligase., The removal of ubiquitin units from proteins is catalyzed by deubiquitinating enzymes [18, 19]. Eukaryotic E3 ligases are mainly classified into two groups, HECT and RING, with different structural features and catalyic mechanisms. The first group is characterized by its HECT (homolog of E6-associated protein C-terminus) domain and during catalysis forms an intermediate that receives ubiquitin from the E2 conjugating enzyme before transferring to substrates [18]. The second type is characterized by the presence of a RING (Really Interesting New Gene) finger domain, which consists of a series of histidine and cysteine residues that coordinate binding to zinc ions. The RING-type E3 ligases do not form a

ubiquitin-linked intermediate, but promote the direct ubiquitin transfer from the E2 to the targeted substrate [18].

Although Gram-negative bacteria lack complete ubiquitination machinery, some pathogenic bacteria have evolved or acquired virulence effectors that can manipulate the process of ubiquitination through structural and/or functional mimicry [20, 21]. Although bacterial proteins that mimic the E1 and E2 enzymes have not been identified, a number of bacterial and viral E3 ligases have been shown to be enzymatically active and to be important for virulence [20, 21]. These E3 ligases expand the number of sequence families from eukaryotic ubiquitin ligases [22, 23], with several displaying structural mimicry, i.e. similar structure and function arising from dissimilar sequence [20]. *E. coli* expresses a class of effector proteins named NleG-like proteins, after the first characterized member of this class, that contain U-boxes, a domain similar to RING but lacking the coordination with zinc ions, and were shown to be enzymatically active E3 ligases [24]. Some Gram-negative bacteria have members of a class of E3 ligases named Novel E3 Ligases (NEL, not to be confused with NleG) that despite having a conserved cysteine residue at the catalytic site has little similarity to HECT domains [25]. Members of NELs include virulence factors, such as *Shigella* IpaH and *Salmonella* SspH1, SspH2 and SlrP [25-29].

Sequence family models have been developed as part of the popular Pfam database that can identify new members of the classes described above, but fail to identify E3 ligases that do not fall into these families. This lack of sequence similarity makes it difficult characterize new ubiquitin ligase mimics in bacteria or viruses. While experimental techniques are essential to definitively characterize a protein's function, they are time-consuming and expensive, making them unrealistic for genome-wide screening of effectors. Computational techniques are a better for choice for identifying the putative function of uncharacterized proteins, which can later verified by experimental assays. Since most protein structures have not been solved experimentally, computational techniques for identifying the function of uncharacterized protein rely upon the similarity of its amino acid sequence to that of a protein with a known function.

Here we present a novel method for alignment-free classification of proteins using kmers built from reduced amino acid alphabets. That is, physicochemical properties or other grouping strategies are used to group amino acids into sets that are then used to represent kmer feature sets. These feature sets are then used as input to an SVM using a family-wise cross-validation strategy and a classifying model is derived. Surprisingly, we found that an amino acid alphabet that represents residues as either generally hydrophobic or generally hydrophilic performed the best as features for classification yielding a classification receiver-operator characteristic (ROC) area under the curve (AUC) performance of 0.90. Feature selection identified several regions of similarity across disparate families of E3 ubiquitin ligases. We predict a number of novel E3 ubiquitin ligases from a large set of genomes with this novel approach.

MATERIALS & METHODS

Dataset

We identified a set of 168 confirmed bacterial or viral E3 ubiquitin ligase effectors from the UniProt database [30, 31]. Negative examples were 235 other bacterial effectors identified from literature [8, 20, 24, 27, 30-44]. We include details on the dataset as Supplemental Data.

To provide predictions for relevant bacterial pathogens we downloaded a set of 171 genomes that are listed as human pathogens and are representative reference genomes from PATRIC [45]. This set comprises 480,562 protein sequences excluding all of the proteins used in the training set above.

Features

Every protein sequence used for either learning or prediction is encoded by counting occurrences of peptides of varying length in the sequence in a manner similar to the previously described string kernel. The possible number of peptides greater than 4 amino acids long is very large ($20^4 = 160{,}000$ peptides). We wanted to extend this approach to identify sequence patterns based on groupings of amino acids based

on physiochemical or other properties. We therefore also encoded sequences to reduce the sequence space using one of several encodings (Table 1.) Features were then generated for a range of different peptide lengths and peptides that were observed in fewer than 10 examples were removed from consideration.

Features for each protein are generated by considering all peptides of length k in a sequence, including overlapping peptides, encoding these (optionally) using the chosen encoding scheme, then counting the occurrences of the encoded peptide.

Data Partitioning

To remove bias created by having multiple examples with very similar features (i.e. closely related effectors from different organisms) we first partitioned the examples to identify/generate clusters of related effectors. In order to achieve this partitioning, we clustered the sequences based on NCBI BLASTP [46] similarity results. Parameters of BLASTP were set to their default values. Using a lower E value threshold (for example, E = 0) groups sequences more tightly and thus results in clusters that are likely to be more similar to another cluster and thus represent a generous division of families for the classification task using our cross-validation approach (see below). Conversely, higher E value thresholds (for example, E = 0.01) yield broader, more general clusters that are less likely to be similar to any other clusters, and thus represent a conservative division of families for our classification task. We used a more conservative threshold to group the set of 407 proteins into 172 clusters of loosely related protein sequences. We examine the effect of varying the BLAST E-value threshold on the sizes of the generated protein families (Supplemental Figure 1).

Cross Validation

Cross validation (CV) is widely used to test the performance of a classification scheme on a given dataset. The entire dataset is partitioned into several non-overlapping folds. These folds are used as test sets. The corresponding training set for a particular fold consists of the remainder of the dataset. Each iteration of

cross validation involves using a training set to generate a model and testing that model on the corresponding test set. This process is repeated until every fold has been tested.

The experimental setup of our study uses a variant of CV called Family-Wise Cross Validation (FWCV) to judge the performance of our classifier. FW places all the samples belonging to a particular cluster in a single test set, while the classifier is trained using the remaining data. This prevents model overfitting by reducing the trivial similarities between testing and training sets (i.e. those similarities based on traditional sequence similiarity).

The Support Vector Machine (SVM) determines the optimally separating hyperplane between two sets of points in high-dimensional feature space each belonging to a different class [47]. We utilized the radial kernel from the e1071 R library in our implementation.

The area under the curve (AUC) and receiver-operator characteristic curve (ROC) calculation was performed using the R library pROC.

Feature Selection

Feature selection was accomplished using SVM Recursive Feature Extraction (SVM-RFE). We can obtain an ordering of the features using the absolute value of the entries of the SVM weight vector w. Each recursive feature elimination iteration involves eliminating the set of features that have the smallest absolute weight w_i until k features remain.

Implementation Details and Availability

Feature generation from sequences is performed using a standalone Python script. Training and validation of models was performed in R. The SVM-RFE algorithm used by SIEVE-Ub was implemented in R as described by GIST-RFE [48, 49].

Code for the algorithm and datasets used for training are available at https://github.com/biodataganache/SIEVE-Ub.

RESULTS

Known ubiquitin ligases fall into one of several sequence families, HECT, RING, and NEL, each of which can be identified using existing hidden Markov models (HMMs) from the Pfam database (PF00632, PF13639, PF14496). Additionally, sequence-based models exist for AvrPtoB (PF09046) and BRE1 (PF08647), which represent distinct E3 ubiquitin ligase families, and SopA (PF13981), which is a HECT-like domain. We analyzed the assembled sequences using the Pfam database and identified members of all these families (Supplemental Data). We note that each of these Pfam families map to a different sequence cluster identified by BLAST, though NEL and RING are broken into more than one sequence cluster each. The family with the most representation in our set of positive examples is the NEL family with 102 members. Taken as a whole the nine Pfam models achieve an accuracy of 95% and a precision of 98% for prediction of E3 ubiquitin ligases from the background of other virulence effectors, with 14 known ubiquitin ligases being missed. It is important to note that neither the BLAST approach we took to identify sequence clusters nor the individual Pfam models provided any predictive ability across sequence families. Our goal is to develop a generalized, alignment-free approach to predict members of this functional family capturing those not identifiable through a sequence-based model such as those in Pfam, and providing the potential to identify novel functional family members.

Dissimilar ubiquitin ligases can be detected using reduced amino acid (RAA) peptides

To provide feature sets that were specific enough to capture relationships between functionally similar proteins, yet general enough to identify regions of similarity between divergent sequences we adapted the kmer approach. Our novel extension translates each amino acid in the sequence to a smaller number of groups based on physicochemical properties or other arbitrary grouping methods- a reduced amino acid

(RAA) alphabet. Initially we chose three reduction mappings based on previously reported approaches: hydrophobicity (RAA1), standard physiochemical properties (RAA2), and solvent accessibility (RAA3) [9, 50]. The groups are listed in Table 1.

The set of positive and negative examples for E3 ubiquitin ligases was encoded using each of the RAAs and the native sequence, and peptide kmers of various lengths were counted for each. Peptides present in fewer than 10 examples were excluded from further consideration. Each dataset was then split into independent training and testing sets on a sequence cluster-wise basis (that is, clusters of similar sequences as determined by BLAST were kept together in the training or testing set), based on a conservative cluster grouping (E < 1e-2.) Cluster-wise splits and associated training and testing were performed 100 times for each model and the score (SVM discriminant) for each example averaged. Average scores were used to determine ROC AUC for each model and results are presented in Table 2 and Supplemental Figure 3.

Surprisingly, the models using RAA1, a simple division of amino acids into hydrophobic and hydrophilic residues, performed the best for nearly all peptide lengths with a maximum AUC of about 0.90. The maximum AUC observed occurs with RAA1 and a peptide length of 14 (RAA1-K14) and so we focused on characterization of this model for the remainder of the paper. Our results indicate that a simple encoding of amino acids can be used to classify effectors with E3 ubiquitin ligase function from other effectors, and from other non-effector proteins in general (see Prediction of novel E3 ubiquitin ligase mimics, below), with good confidence.

We hypothesized that the performance of the RAA1 is based on accurately representing the pattern of hydrophobic and hydrophilic residues in kmers. To examine this hypothesis we applied a family-wise cross-validation approach using ten alphabets where residues had been randomly assigned to either the hydrophobic or hydrophilic groups preserving the overall balance of hydrophobic to hydrophilic residues

in the resulting random alphabet (6:14; see Table 1). We compared the performance of these random binary RAAs at a kmer size of 14 with the true hydrophobic/hydrophilic RAA1-K14 also run ten times to show the variability in partitioning of training and testing sets inherent in our approach and show the results in Figure 1. In all cases the true RAA1 outperforms the randomized RAAs supporting our hypothesis though we note that there is a wide range of performances given with random binary RAAs. We believe this is due to some random assortments containing reasonable divisions of residues between hydrophobic and hydrophilic residues because of the very simple nature of this division.

SIEVE-Ub identifies biologically functional peptides

To identify a minimal set of features that are important for classification of E3 ubiquitin ligases from other effectors we used recursive feature elimination, a standard machine learning approach [8]. Briefly, a model is trained on all features, then weights for each feature are used to discard 50% of the features with the lowest impact on model performance. The remaining features are then used in another model training round in which this process is repeated until all the features have been eliminated. The training performance results from the RFE on the RAA1-K14 model are shown in Figure 2. We chose to keep 100 features in our final analysis given that this provided good training performance (AUC >0.9), but retained a small portion of the initial features (3%). These features are provided as Supplemental Data along with their locations in each of the positive and negative examples in our analysis set.

Though the E3 ligase examples used as our positive examples are diverse in terms of sequence many do fall into the families of E3 ligases described in the Introduction; HECT/U-box, RING, and NEL. We chose two example effectors to highlight the biological relevance of our findings. The NleL (HECT) and SspH2 (NEL) effectors have crystal structures available and in the case of NleL have also been solved in the presence of the E2 conjugating enzyme (UbcH7) [51]. In each of these structures a top-scoring peptide match was found close to the known (NleL) or presumed (SspH2) E2 binding site. The kmer peptides for both structures are directly C-terminal of the catalytic cysteine residue. The kmer peptides

matched amphipathic alpha helices with buried hydrophobic residues and exposed polar or charged residues, including a histidine for each (Figure 3).

Since a limited number of structures are available for E3 ubiquitin ligases, and some of these structures cover only small regions of the proteins, this analysis was not possible for all examples. However, RING/U-box E3 ligases have a consensus motif with two repeated zinc fingers: $Cx_2Cx_{9-39}Cx_{1-3}Hx_{2-3}/Hx_2Cx_{4-48}Cx_2C$ [20]. The first zinc finger has been found to be responsible for E2 binding and catalytic activity whereas there is evidence that the second zinc finger directs binding to host targets, such as Cdc2-like kinase 1 (Clk1) in the case of the *L. pneumophila* LubX protein [52]. We found that top-scoring peptides from our model matched the second zinc finger sequences for several RING/U-box E3 ligases including the LubX protein and the herpesvirus ICP0 protein, suggesting that these peptides participate in interactions with the host target.

Prediction of novel E3 ubiquitin ligase mimics

To predict novel E3 ubiquitin ligase mimics in a larger set of sequences we applied the model described above (kmer 14 in RAA1, top 100 most important features) to a set of over 400,000 proteins from representative human pathogens obtained from the PATRIC database [45]. We further filtered this list using a version of our previously developed type III secreted effector prediction algorithm, SIEVE [8]. The combination of these two methods provides a list of predicted E3 ubiquitin ligases that are also predicted to be secreted via type III mechanism, though we note that such effectors could be secreted via other mechanisms. These predictions are listed in Table 3. Most of these top predictions are hypothetical proteins, with the exception of the RNA endonuclease, which could be a false positive barring any unusual and unexpected dual functionality. Though two of the predictions are quite short in length at around 40 amino acids, this is consistent with the length of, for example, the RING zinc finger motif of E3 ubiquitin ligases, so these predictions should not be immediately discarded, though the involvement of additional protein machinery would be stipulated if a novel E3 ligase were to be presumed to at least have

similar requirements for binding the ubiquitin and host target substrates.

DISCUSSION

We note that the intent of our study was to develop a model that could identify E3 ubiquitin ligases based on protein sequence with reasonable accuracy and precision, which we demonstrated clearly. As such, we did not fully explore the range of possible parameters such as choice of SVM kernel, or other machine learning approaches that would work on our input features, to determine an optimal model. Our results show that we can use models based on highly divergent sequences to robustly predict E3 ubiquitin ligase function in bacterial and viral effectors. It is unclear how many E3 ubiquitin ligases that may exist but have not yet been discovered, and this question will only be answered through experimental validation of predictions made by our method, similar to the validation we have done for the original SIEVE [8].

CONCLUSIONS

The general approach we describe, using peptides with reduced amino acid alphabets as features for machine learning, could be easily applied to other problems of functional classification given appropriate positive and negative example sets. We show that this approach can be used to discriminate effectors with E3 ubiquitin ligase activity from other effectors with good confidence and present a single model that is able to identify E3 ubiquitin ligases from different sequence families. Importantly, development of this model does not require sequence alignment of any kind. From this analysis we have presented an example of this approach identifying functionally important regions with dissimilar sequences, but similar structures. However, further work is necessary to explore the possibility that this is a more general property of the approach. This is the first algorithm dedicated to prediction of E3 ligase function in non-eukaryotic proteins. In combination with our existing SIEVE algorithm for prediction of Type III secreted effectors our SIEVE-Ub algorithm can be used to predict novel effectors with E3 ligase activity as we've shown in Table 3. Combining this approach with type IV prediction algorithms would allow similar results for type IV secretion systems.

ACKNOWLEDGEMENTS

We would like to recognize anonymous reviewer #2 from a previous submission of this work whose diligence and thoroughness helped shape the current version for the better.

REFERENCES

1. McDermott JE, Bruillard P, Overall CC, Gosink L, Lindemann SR: **Prediction of multi-drug resistance transporters using a novel sequence analysis method**. *F1000Research* 2015, **4**:60.
2. Qu K, Han K, Wu S, Wang G, Wei L: **Identification of DNA-Binding Proteins Using Mixed Feature Representation Methods**. *Molecules* 2017, **22**(10).
3. Abbasi WA, Asif A, Andleeb S, Minhas F: **CaMELS: In silico prediction of calmodulin binding proteins and their binding sites**. *Proteins* 2017, **85**(9):1724-1740.
4. Tung CH, Chen CW, Sun HH, Chu YW: **Predicting human protein subcellular localization by heterogeneous and comprehensive approaches**. *PLoS One* 2017, **12**(6):e0178832.
5. Kuksa PP, Min MR, Dugar R, Gerstein M: **High-order neural networks and kernel methods for peptide-MHC binding prediction**. *Bioinformatics* 2015, **31**(22):3600-3607.
6. Wang D, Zeng S, Xu C, Qiu W, Liang Y, Joshi T, Xu D: **MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction**. *Bioinformatics* 2017, **33**(24):3909-3916.
7. McDermott J, Corrigan A, Peterson E, Oehmen C, Niemann G, Cambronne E, Sharp D, Adkins J, Samudrala R, Heffron F: **Computational prediction of type III and IV secreted effectors in Gram-negative bacteria**. *Infection and Immunity* 2010, **In Press**.
8. Samudrala R, Heffron F, McDermott JE: **Accurate prediction of secreted substrates and identification of a conserved putative secretion signal for type III secretion systems**. *PLoS Pathog* 2009, **5**(4):e1000375.
9. Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S, Niinikoski A, Mewes HW, Horn M, Rattei T: **Sequence-based prediction of type III secreted proteins**. *PLoS Pathog* 2009, **5**(4):e1000376.
10. Niemann GS, Brown RN, Gustin JK, Stufkens A, Shaikh-Kidwai AS, Li J, McDermott JE, Brewer HM, Schepmoes A, Smith RD *et al*: **Discovery of novel secreted virulence factors from Salmonella enterica serovar Typhimurium by proteomic analysis of culture supernatants**. *Infect Immun* 2011, **79**(1):33-43.
11. Hovis KM, Mojica S, McDermott JE, Pedersen L, Simhi C, Rank RG, Myers GS, Ravel J, Hsia RC, Bavoil PM: **Genus-optimized strategy for the identification of chlamydial type III secretion substrates**. *Pathogens and disease* 2013.
12. Leslie C, Eskin E, Noble WS: **The spectrum kernel: a string kernel for SVM protein classification**. *Pac Symp Biocomput* 2002:564-575.
13. Leslie CS, Eskin E, Cohen A, Weston J, Noble WS: **Mismatch string kernels for discriminative protein classification**. *Bioinformatics* 2004, **20**(4):467-476.
14. Li H, Jiang T: **A class of edit kernels for SVMs to predict translation initiation sites in eukaryotic mRNAs**. *J Comput Biol* 2005, **12**(6):702-718.
15. Sher G, Zhi D, Zhang S: **DRREP: deep ridge regressed epitope predictor**. *BMC Genomics* 2017, **18**(Suppl 6):676.

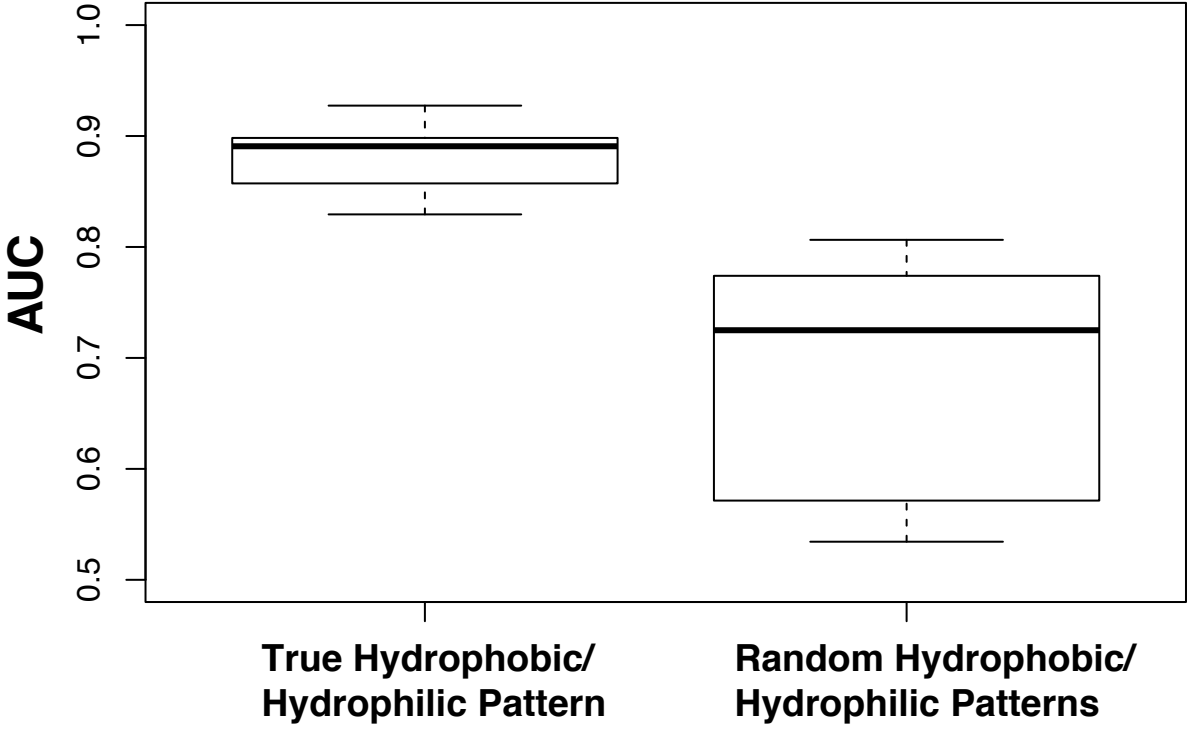
16. Bhoj VG, Chen ZJ: **Ubiquitylation in innate and adaptive immunity**. *Nature* 2009, **458**(7237):430-437.
17. Kravtsova-Ivantsiv Y, Ciechanover A: **Non-canonical ubiquitin-based signals for proteasomal degradation**. *Journal of cell science* 2012, **125**(Pt 3):539-548.
18. Metzger MB, Hristova VA, Weissman AM: **HECT and RING finger families of E3 ubiquitin ligases at a glance**. *Journal of cell science* 2012, **125**(Pt 3):531-537.
19. Komander D, Clague MJ, Urbe S: **Breaking the chains: structure and function of the deubiquitinases**. *Nature reviews Molecular cell biology* 2009, **10**(8):550-563.
20. Hicks SW, Galan JE: **Hijacking the host ubiquitin pathway: structural strategies of bacterial E3 ubiquitin ligases**. *Curr Opin Microbiol* 2010, **13**(1):41-46.
21. Rytkonen A, Holden DW: **Bacterial interference of ubiquitination and deubiquitination**. *Cell host & microbe* 2007, **1**(1):13-22.
22. Catic A, Misaghi S, Korbel GA, Ploegh HL: **ElaD, a Deubiquitinating protease expressed by E. coli**. *PloS one* 2007, **2**(4):e381.
23. Cui J, Yao Q, Li S, Ding X, Lu Q, Mao H, Liu L, Zheng N, Chen S, Shao F: **Glutamine deamidation and dysfunction of ubiquitin/NEDD8 induced by a bacterial effector family**. *Science* 2010, **329**(5996):1215-1218.
24. Wu B, Skarina T, Yee A, Jobin MC, Dileo R, Semesi A, Fares C, Lemak A, Coombes BK, Arrowsmith CH *et al*: **NleG Type 3 effectors from enterohaemorrhagic Escherichia coli are U-Box E3 ubiquitin ligases**. *PLoS pathogens* 2010, **6**(6):e1000960.
25. Singer AU, Rohde JR, Lam R, Skarina T, Kagan O, Dileo R, Chirgadze NY, Cuff ME, Joachimiak A, Tyers M *et al*: **Structure of the Shigella T3SS effector IpaH defines a new class of E3 ubiquitin ligases**. *Nature structural & molecular biology* 2008, **15**(12):1293-1301.
26. Rohde JR, Breitkreutz A, Chenal A, Sansonetti PJ, Parsot C: **Type III secretion effectors of the IpaH family are E3 ubiquitin ligases**. *Cell host & microbe* 2007, **1**(1):77-83.
27. Quezada CM, Hicks SW, Galan JE, Stebbins CE: **A family of Salmonella virulence factors functions as a distinct class of autoregulated E3 ubiquitin ligases**. *Proc Natl Acad Sci U S A* 2009, **106**(12):4864-4869.
28. Levin I, Eakin C, Blanc MP, Klevit RE, Miller SI, Brzovic PS: **Identification of an unconventional E3 binding surface on the UbcH5 ~ Ub conjugate recognized by a pathogenic bacterial E3 ligase**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(7):2848-2853.
29. Bernal-Bayard J, Ramos-Morales F: **Salmonella type III secretion effector SlrP is an E3 ubiquitin ligase for mammalian thioredoxin**. *The Journal of biological chemistry* 2009, **284**(40):27587-27595.
30. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M *et al*: **The Universal Protein Resource (UniProt)**. *Nucleic acids research* 2005, **33**(Database issue):D154-159.
31. **UniProt**. In.
32. Lee VT, Mazmanian SK, Schneewind O: **A program of Yersinia enterocolitica type III secretion reactions is activated by specific signals**. *J Bacteriol* 2001, **183**(17):4970-4978.
33. Anderson DM, Frank DW: **Five mechanisms of manipulation by bacterial effectors: a ubiquitous theme**. *PLoS pathogens* 2012, **8**(8):e1002823.
34. Buchko GW, Niemann G, Baker ES, Belov ME, Smith RD, Heffron F, Adkins JN, McDermott JE: **A multi-pronged search for a common structural motif in the secretion signal of Salmonella enterica serovar Typhimurium type III effector proteins**. *Mol Biosyst* 2010, **6**(12):2448-2458.

35. Burstein D, Zusman T, Degtyar E, Viner R, Segal G, Pupko T: **Genome-scale identification of Legionella pneumophila effectors using a machine learning approach**. *PLoS Pathog* 2009, **5**(7):e1000508.
36. Collins CA, Brown EJ: **Cytosol as battleground: ubiquitin as a weapon for both host and pathogen**. *Trends in cell biology* 2010, **20**(4):205-213.
37. Dean P: **Functional domains and motifs of bacterial type III effector proteins and their roles in infection**. *FEMS microbiology reviews* 2011, **35**(6):1100-1125.
38. Deslandes L, Rivas S: **Catch me if you can: bacterial effectors and plant targets**. *Trends in plant science* 2012.
39. Lin DY, Diao J, Zhou D, Chen J: **Biochemical and structural studies of a HECT-like ubiquitin ligase from Escherichia coli O157:H7**. *The Journal of biological chemistry* 2011, **286**(1):441-449.
40. McDermott JE, Corrigan A, Peterson E, Oehmen C, Niemann G, Cambronne ED, Sharp D, Adkins JN, Samudrala R, Heffron F: **Computational prediction of type III and IV secreted effectors in gram-negative bacteria**. *Infect Immun* 2011, **79**(1):23-32.
41. Price CT, Kwaik YA: **Exploitation of Host Polyubiquitination Machinery through Molecular Mimicry by Eukaryotic-Like Bacterial F-Box Effectors**. *Frontiers in microbiology* 2010, **1**:122.
42. Spallek T, Robatzek S, Gohre V: **How microbes utilize host ubiquitination**. *Cellular microbiology* 2009, **11**(10):1425-1434.
43. Stebbins CE, Galan JE: **Structural mimicry in bacterial virulence**. *Nature* 2001, **412**(6848):701-705.
44. Xin DW, Liao S, Xie ZP, Hann DR, Steinle L, Boller T, Staehelin C: **Functional analysis of NopM, a novel E3 ubiquitin ligase (NEL) domain effector of Rhizobium sp. strain NGR234**. *PLoS pathogens* 2012, **8**(5):e1002707.
45. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, Conrad N, Dietrich EM, Disz T, Gabbard JL *et al*: **Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center**. *Nucleic Acids Res* 2017, **45**(D1):D535-D542.
46. **NCBI BLASTP**. In.
47. Noble WS: **What is a support vector machine?** *Nat Biotechnol* 2006, **24**(12):1565-1567.
48. Edwards JS, Palsson BO: **Robustness analysis of the Escherichia coli metabolic network**. *Biotechnol Prog* 2000, **16**(6):927-939.
49. Edwards JS, Palsson BO: **Multiple steady states in kinetic models of red cell metabolism**. *J Theor Biol* 2000, **207**(1):125-127.
50. Bacardit J, Stout M, Hirst JD, Valencia A, Smith RE, Krasnogor N: **Automated alphabet reduction for protein datasets**. *BMC Bioinformatics* 2009, **10**:6.
51. Lin DY, Diao J, Chen J: **Crystal structures of two bacterial HECT-like E3 ligases in complex with a human E2 reveal atomic details of pathogen-host interactions**. *Proc Natl Acad Sci U S A* 2012, **109**(6):1925-1930.
52. Quaile AT, Urbanus ML, Stogios PJ, Nocek B, Skarina T, Ensminger AW, Savchenko A: **Molecular Characterization of LubX: Functional Divergence of the U-Box Fold by Legionella pneumophila**. *Structure* 2015, **23**(8):1459-1469.

Figure 1(on next page)

Amino acid reduction based on physicochemical properties is important.

Models were evaluated using the standard hydrophobic/hydrophilic reduction alphabet (RED0) and randomly divided sets of amino acids (RND0) with a kmer length of 14. Performance was evaluated using 100 fold family-wise cross validation and AUC. The plot shows that a division of amino acids into hydrophobic and hydrophilic residues outperforms a random division of amino acids.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27292v1 | CC0 Open Access | rec: 20 Oct 2018, publ: 20 Oct 2018

Figure 2(on next page)

Model performance with varying numbers of features.

Recursive feature elimination (RFE) was applied to all examples using 14mers and the RAA1 and AUC assessed for each model. The plot shows that very good performance can be achieved with 100 features.

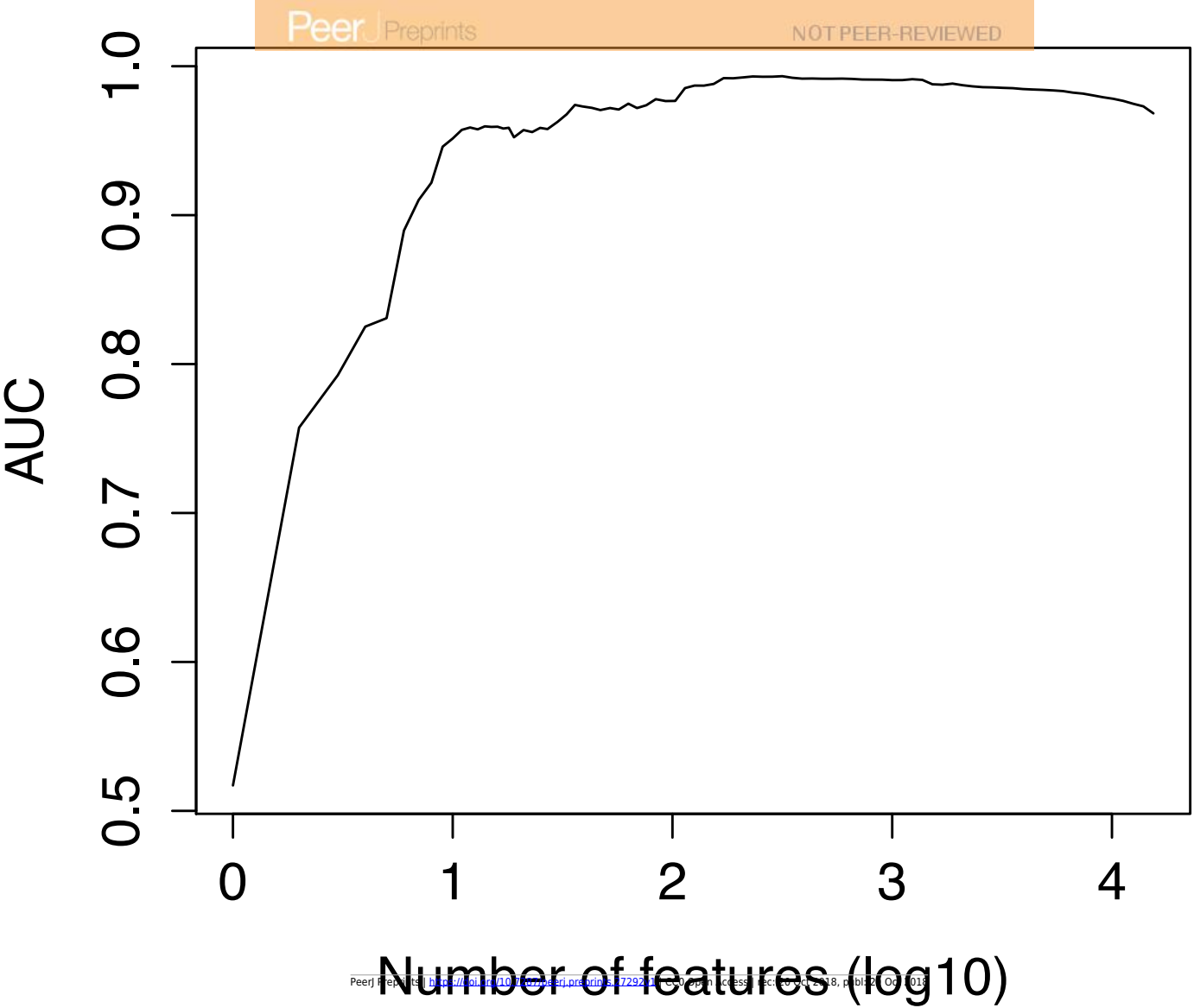


Figure 3

Discriminating peptides in E3 ligase structures.

Ribbon cartoon diagrams of known bacterial ubiquitin E3 ligase mimics *E. coli*NleL and *Salmonella*SspH2 (NEL), as well as NleL homologue SopA from *Salmonella*which was not identified by SIEVE-Ub but has sequence similarity at the site of the kmer peptide of NleL. In NleL the kmer peptide is a helix (depicted as light blue/red spheres) that interacts alternately with either E2 (in open form) or the hinge linking the N-term and C-term domains (in closed form), as if mediating the two structural forms. For SspH2, there is no structure with bound E2 available, but the helix is similarly positioned relative to the LRR-domain and the catalytic Cys. The catalytic Cys in each structure near the N-term of the kmer helix is indicated as red spheres. No structural information about a presumed ubiquitin binding site is available for either of these structures.

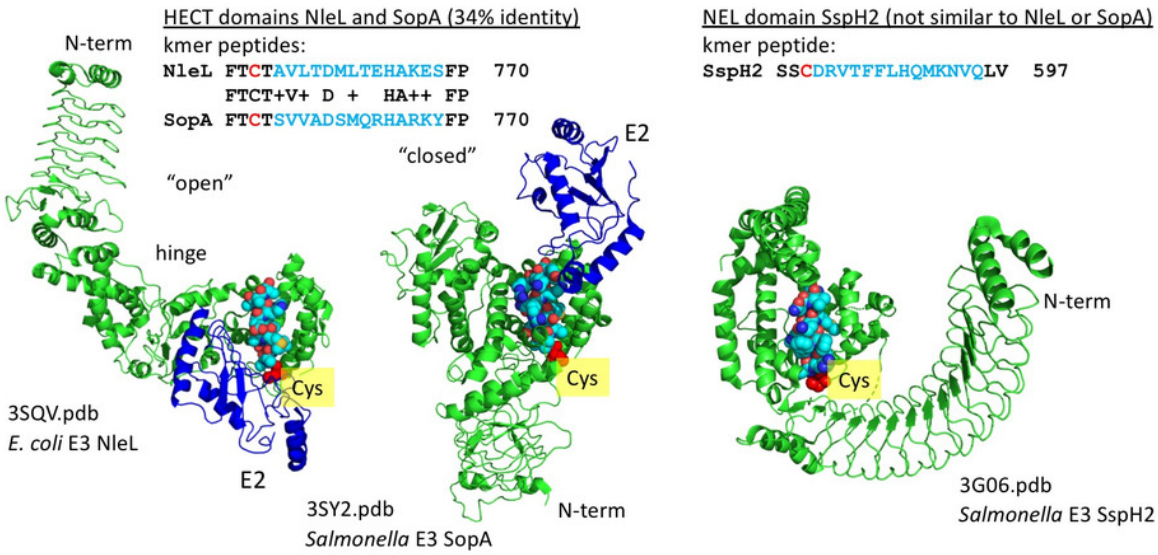


Table 1(on next page)

Reduced amino acid (RAA) encodings

Table 1. Reduced amino acid (RAA) encodings

Name	Groups	Notes	Reference
NAT (Natural)	ACDEFGHIKLMNPQRSTVWY	No encoding	
RAA1 (Hydrophobicity)	SFTNKYEQCWPHDR AGILMV	Hydrophilic Hydrophobic	[9]
RAA2 (Physiochemical)	AGILMV PH FEY NQST DE KR CY	Hydrophobic Hydrophilic Aromatic Polar Acidic Basic Ionizable	[9]
RAA3 (Solvent accessibility)	CILMVFWY AGHST PDEKNQR	Low Medium High	[50]
RAA4 (Hydrophobicity and charge)	SFTNYQCWPH AGILMV KEDR	Hydrophobic Hydrophilic Charged	This study
RAA5 (Hydrophobicity and structure)	SFTNKYEQCWHDR AILMV PG	Hydrophilic Hydrophobic Structural	This study

Table 2(on next page)

Best model performance

Table 2. Best model performance

	Kmer Length	AUC
NAT	17	0.851
RAA1	14	0.903
RAA2	6	0.803
RAA3	8	0.742
RAA4	6	0.884
RAA5	13	0.814

Table 3(on next page)

Proteins predicted to be similar to ubiquitin ligase mimic set.

*annotation based on sequence comparison only

Table 3. Proteins predicted to be similar to ubiquitin ligase mimic set. *annotation based on sequence comparison only

Genbank ID	SIEVE score	SIEVE-Ub score	Genome	Length	Gene	Description
WP_012732629.1	0.50	0.82	Corynebacterium kroppenstedtii	360		hypothetical protein
WP_082022266.1	0.31	0.71	Rickettsia conorii	43		hypothetical protein
ABE96403.1	0.30	0.87	Bifidobacterium breve	1021	rne	Ribonuclease E (EC 3.1.26.12)*
AMD88982.1	0.30	0.58	Desulfovibrio fairfieldensis	159		hypothetical protein
AMD99888.1	0.24	0.65	Actinomyces oris	428		GNAT family acetyltransferase*
KDS45810.1	0.23	0.87	Bacteroides cellulosilyticus	45		hypothetical protein
WP_012742696.1	0.21	0.65	Eubacterium rectale	551		Iron-sulfur flavoprotein multimeric flavodoxin WrbA*

5