Two-stage ML Classifier for Identifying Host Protein Targets of the Dengue Protease

Jacob T. Stanley, Alison R. Gilchrist, Alex C. Stabell, Mary A. Allen, Sara L. Sawyer, Robin D. Dowell[†]

Molecular, Cellular and Developmental Biology, BioFrontiers Institute, University of Colorado,
Boulder, CO 80305, USA
[†]*E-mail: robin.dowell@colorado.edu*

Flaviviruses such as dengue encode a protease that is essential for viral replication. The protease functions by cleaving well-conserved positions in the viral polyprotein. In addition to the viral polyprotein, the dengue protease cleaves at least one host protein involved in immune response. This raises the question, what other host proteins are targeted and cleaved? Here we present a new computational method for identifying putative host protein targets of the dengue virus protease. Our method relies on biochemical and secondary structure features at the known cleavage sites in the viral polyprotein in a two-stage classification process to identify putative cleavage targets. The accuracy of our predictions scaled inversely with evolutionary distance when we applied it to the known cleavage sites of several other flaviviruses—a good indication of the validity of our predictions. Ultimately, our classifier identified 257 human protein sites possessing both a similar target motif and accessible local structure. These proteins are promising candidates for further investigation. As the number of viral sequences expands, our method could be adopted to predict host targets of other flaviviruses.

Keywords: viral protease; dengue; SVM; target discovery

1. Introduction

Flaviviruses are a genus of positive-strand RNA viruses under the family Flaviviridae, most of them insect-borne. The clade includes many viruses that are important for human health, such as dengue, Zika, West Nile, yellow fever, and tick-borne encephalitis virus. Dengue virus alone infects approximately 390 million people every year, and is endemic in at least 100 countries.[1] The genome of dengue virus encodes a protease (composed of the non-structural proteins NS2B and NS3) the action of which is essential for viral replication. It recognizes an eight amino acid motif, cleaving between the fourth and fifth position.[2,3] As well as cleaving the viral polyprotein (Fig. 1), the dengue virus protease cleaves at least one host protein during replication.[4,5] Stimulator of Interferon Genes (STING) is an ER-resident protein that is activated by ER stress or cytoplasmic nucleotides, and in turn activates the interferon response. The dengue virus protease can cleave STING between amino acid positions 78 and 79, thereby down-regulating the interferon response and increasing virus replication. STING is also cleaved by Zika, Japanese encephalitis, and West Nile viruses, indicating that overlap exists in the set of motifs these divergent protease target.[6] Furthermore, the identification of STING as a target of the flavivirus proteases suggests there may be other human protein targets that have yet to be discovered. In this work, we present a method for identifying protease targets computationally.

The motifs recognized by any given protease can be identified based on high-throughput screening assays or computational predictions.[3,7] High-throughput screens based on randomized peptide generation or mass spectrometry vary in their specificity and accuracy and can be costly or time consuming. Alternatively, there exist a number of computational methods (referred to as “motif scanning”) for identifying putative target motifs in primary sequence. Most of these rely on probabilistic models using position-weight matrices (PWMs), i.e. frequency at which an amino acid appears in known "reference" motifs at a given position.[8–15] PWM methods are broadly successful, but can be ill-suited for certain applications. Firstly, because they assess similarity based only on the frequency of each monomer or dimer, these methods can be insensitive to less frequent—but still successfully targeted—reference motifs. Published sequences of dengue viruses, for example, reveal motifs within the viral population that are highly dissimilar yet still effectively cleaved by the viral protease. Secondly, since they are constrained to the lengths of the reference motifs (i.e. the list of motifs from which the position-weight matrix is created), they are blind to the properties of the surrounding sequence that may influence accessibility to a positively identified site (conversely, extending the effective size of this motif would require a prohibitively large training data set). For example, secondary or tertiary structure might make a site inaccessible. A motif scanning method that is based on the motif's biochemistry instead of amino acid frequency, but also incorporates the structural context surrounding the motif, would circumvent these shortcomings and produce more biologically meaningful results.

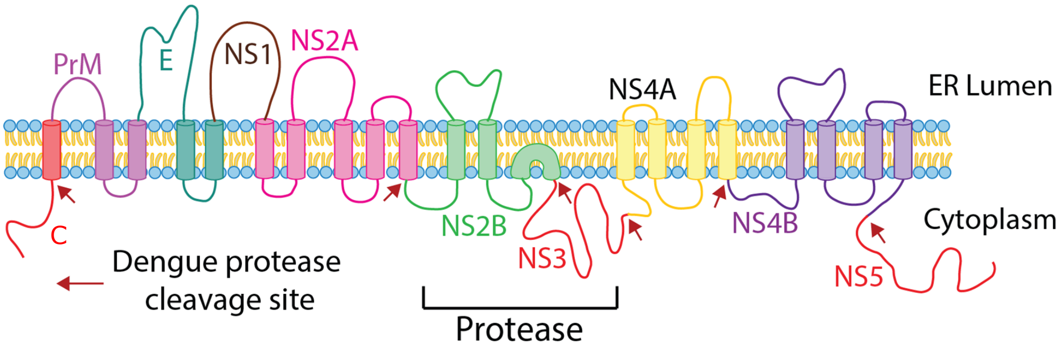

Fig. 1. The dengue virus polyprotein is translated by the host machinery in the ER. It is a multi-pass transmembrane sequence that is then cleaved by its native protease NS2B–NS3, which is located on the cytoplasmic side of the membrane. The six cleaved sites are indicated by the red arrows. (adapted from Umareddy et al.[16])

Here we propose a new computational method to predict targets of flavivirus proteases. This method consists of a two-stage supervised-learning classifier, trained on motifs from the dengue virus polyprotein known to be cleaved by the dengue virus protease. The first stage identifies putative target motifs, informed by biochemical features of the amino acids comprising the cleavage motifs. The second stage identifies which of the target sites predicted by the first stage have a structural context that makes the site accessible to the protease, thus acting as a filter. Unlike other motif scanning methods, this two-stage method uses biochemical properties of each amino acid within the motif, not its frequency, and evaluates the motif's surrounding secondary structure. We show that the motif stage predicts a more diverse range of targets in host proteins than a PWM method, and the structure stage of the classifier greatly reduces the number of sites predicted to be cleaved, suggesting most are not accessible. Additionally, we show that our dengue-trained classifier, when applied to other flavivirus motifs, recapitulates the evolutionary divergence of the cleavage motifs of this virus clade, providing further validation of our method.

2. Methods

An overview of our classifier method can be seen in Fig. 2. Supplementary materials available at http://dowell.colorado.edu/pubs/dengue/

2.1. *Curation of dengue polyprotein sequence and training data*

To capture as much diversity for our training data as possible, we downloaded all available polyprotein sequences (length 3400 residues) for the known dengue serotypes (DENV1-4) from the Virus Pathogen Resource database.[17] These sequences were then grouped by serotype and collectively aligned using the MUSCLE alignment method.[18] Alignments were manually curated to remove those that aligned poorly or were significantly incomplete. A total of 7955 instances of the polyprotein remained (see supplemental info for list of NCBI accession numbers). The well-conserved sequence positions of the six known cleavage sites (C/prM, NS2A/NS2B, NS2B/NS3, NS3/NS4A, NS4A/NS4B, and NS4B/NS5;[2,19] Fig. 1, red arrows) were identified for each of the serotype alignments and the 50 residues to either side of each site were extracted from all of the polyprotein sequences. Duplicate 100 residue sequences were removed, resulting in 2297 unique sequences. From this final set we sourced the positive training data for both classifiers. To train the first-stage classifier, the eight-residue peptides that are centered on the cleavage sites represented the positively cleaved target motifs.[2,3] To train the second-stage classifier, positive (i.e. "accessible") structural context was derived from the full 100 residue sequence. Negative training data for the first stage was also derived from the full-length polyprotein sequences by randomly sampling eight-residue peptides that were both from the cytoplasmic portions of the polyprotein and farther than four residues from any cleavage site.[20] We excluded transmembrane and ER lumenal regions because they are physically inaccessible to the protease—thus sequences in these regions do not constitute motifs that are definitively uncleavable.

Our curation resulted in $N_+ = 195$ unique positive cleavage motifs and $N_- = 35{,}702$ unique negative cleavage motifs, with no overlap between the two sets. The positive cleavage motifs, with the cleavage site being between positions four and five (see "Stage 1: Training Data" in Fig. 2), show a common serine protease signal: strong preference for basic amino acid residues (arginine or lysine) at positions three and four and polar amino acids at position five of the substrate.[2,3] The negative cleavage motifs show an essentially uniform distribution of amino acid residues across all positions of the motif substrate (see Fig. 3).

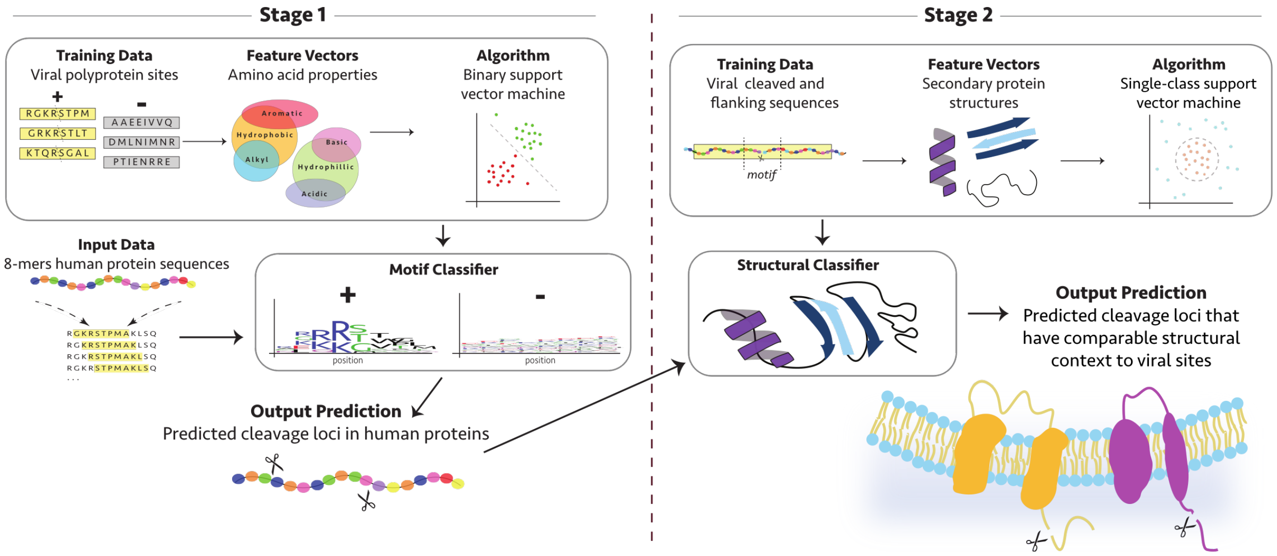

Fig. 2. Schematic of our two-stage method, showing the feature selection and training for both stages. The output predictions of the first stage ("Motif Classifier") are the inputs to the second stage ("Structural Classifier").

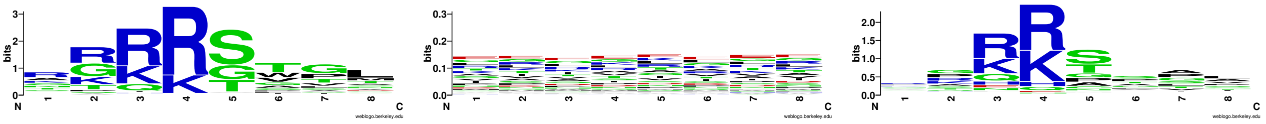

Fig. 3. Sequence logos[21] for curated positive (left) and negative (middle) motif training data. Also, sequence logo for the 3991 putative cleavage sites predicted by the first stage of classification (right). The color scheme corresponds to chemical properties: polar (green); hydrophobic (black); basic (blue); and acidic (red). Background assumes all amino acids are equally likely.

2.2. *Curation of human proteome*

Since the dengue protease is ER-resident and the polyprotein it targets is a multi-pass transmembrane sequence, we narrowed the human proteome search space to only membrane-associated proteins, using the longest isoform in each case (the list of NCBI accession numbers we used can be found in the supplemental information). We identified transmembrane domains using the transmembrane topology prediction program Phobius 1.01[22] on each of the proteins in our list (using a membrane domain fractional probability threshold of >0.5). These domains were then excluded from consideration by the classifier, as they would not be accessible to the protease. From the remaining protein sequence, we evaluated the stage one classifier on every eight amino acid peptide therein, using a sliding window approach (see “Stage 1: Input Data” in Fig. 2).

2.3. *Stage 1 features: amino acid properties*

For our method, we employed a support vector machine (SVM) algorithm, which performs best with features that are bounded, ordinal, and continuous. We wanted a feature representation that captured key biological features, was interpretable, and was easy to compute. We sought to capture the catalytic mechanism of the protease by representing each motif residue with multiple biochemical properties that affect protease activity (see “Stage 1: Features” in Fig. 2). Namely, to represent each individual amino acid, we used polar requirement,[23,24] hydropathy,[25] molecular volume,[26] isoelectric point,[27] and four principal components generated from PCA on 134 physio-chemical properties of amino acids that correlate with biological activity (called Sneath’s Index).[28] At least one other method exists that leverages biological similarity between sequences,[29] however, it can be expensive to compute and difficult to extract meaning from the features. Our chosen features are biologically relevant and require no additional computation. Furthermore, since these features are ordinal and approximately continuous, they are well-suited for an SVM. These eight features were used to represent each of the eight residues in the motif sequence, resulting in a 64-length feature vector for the first-stage classifier.

To visualize our curated motif training data (described in Sec. 2.1) with this feature representation, we performed PCA on the positive and negative training data and plotted the first two principal components (combined, they explain approximately 25% of the variance). As shown on the left of Fig. 4, the positive and negative training data sets show clear separability, a good indication that they constitute distinct categories with this feature representation.

2.4. *Stage 1 classifier: training on length eight amino acid motifs*

For this supervised binary-classification task, we employed the support vector machine (SVM) algorithm with a radial basis function (RBF) kernel (see “Stage 1: Algorithm” in Fig. 2).[30] Our classification method was implemented using scikit-learn (version 0.20.3).[31,32] The eight-residue motif training

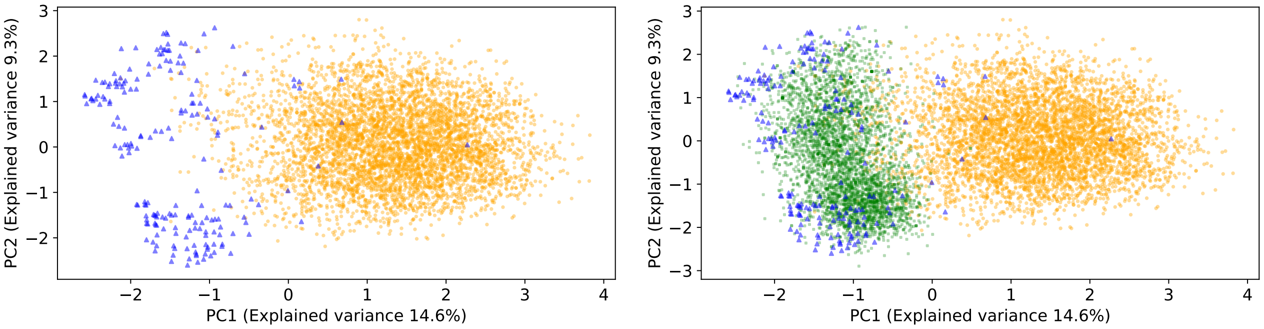

Fig. 4. Left: Principal components (PC1 and PC2) for our training data extracted from the dengue polyprotein sequence, expressed with our biochemical feature representation. The two sets of training data clearly separate along PC1. Higher order components do not demonstrate separability. Right: The same principal components, including the predicted human motifs. Positive cleavage motifs—blue triangles; negative cleavage motifs—orange points; and predicted motifs—green squares.

data curated from the dengue polyproteins (described in Sec. 2.1) were converted using the feature representation in Sec. 2.3 and then each feature was scaled to the range [−1,+1] within the data sets.

There was a significant difference in the size of the two training data sets ($N_+ = 195$ and $N_- = 35{,}702$), which can have a detrimental effect on classifier performance.[33] To address this imbalance we combine subsamples of the larger class plus the entirety of the smaller class.[33] In this way, a collection of equivalent classifiers are generated, each trained on a distinct subset of the negative class. Each equivalent classifier (we created 100 non-overlapping subsamples) is then applied to the target data and results are combined to form a final classification. For each individual classifier we left out a randomly sampled 15% for testing (equal parts positive and negative class), then performed 5-fold cross-validation with the remaining 85% in order to optimize the hyperparameters C and γ. Optimization of hyperparameters used an initial logarithmic grid search, further refined by a local, linear grid search, resulting in final values $C = 1.1$ and $\gamma = 0.12$. The average test performance across the 100 trained classifiers was $f1{-}score = 0.98 \pm 0.02$. The resulting 100 binary-SVM motif classifiers, were subsequently applied to the curated human protein data, described in Sec. 2.2 (see "Motif Classifier" in Fig. 2). In order to identify a target site as cleaved, we required that all 100 classify the site as such.

2.5. *Stage 2 features: secondary structure*

The purpose of the second stage of classification is to determine if each motif resides within the appropriate structural context, as defined by the known cleavage sites in the dengue polyprotein, and therefore would be accessible to the protease. We employed an SVM-based algorithm for this second stage as well.

We based our feature representation for this stage on the local protein secondary structure (within 50 residues of the cleavage site). For features, we focused on the three principle types of secondary structure—the α-helix, β-pleated sheet, and unstructured sequence, referred to as "random coil" (see "Stage 2: Features" in Fig. 2).[34] To predict these secondary structures we used the protein structure prediction server JPred4 (Jnet v.2.3.1).[35] For each residue of the input sequence, the server returns fractional probabilities that the residue was contained in an α-helix, β-pleated sheet, or random coil.[35] Thus, this feature representation is well suited for many classifier algorithms as it is numerical, ordinal, and scaled to the range [0,1]. Structural features were determined using the entire 100-residue amino acid sequences.

Only the central 30 residues (15 to either side of the cleavage site) were used as the training

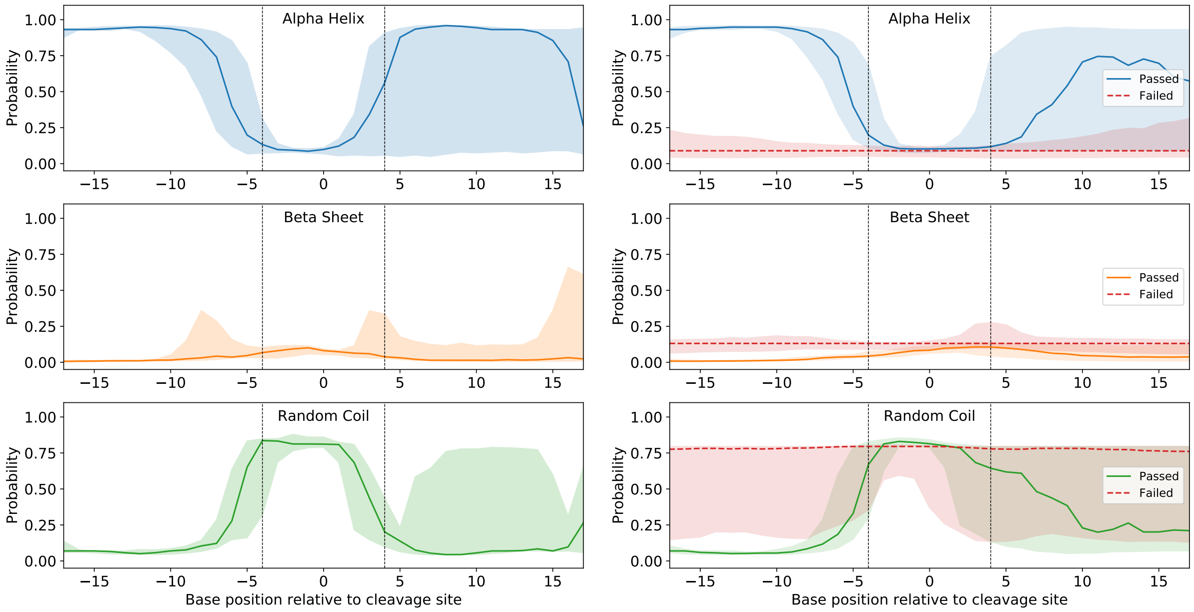

Fig. 5. Secondary structure for training data (left) and Stage 2 predictions (right) – Median probabilities (solid line) of the feature as a function of position relative to cleavage site (shaded regions indicated 2nd and 3rd quartiles). The dotted lines indicate the bounds of the motif (i.e., dotted lines in "Stage 2: Training Data" in Fig. 2).

data for the second stage classifier, since secondary structures are typically less than 30 residues in length[34] and shorter feature vectors are easier to compute. Using the full 100-residue sequences for structure prediction reduced the impact of edge effects on the central 30-residue region of interest. The 30-residue training sequences were converted to linear feature vectors of dimension 90: 3 features (α-helix, β-pleated sheet, random coil) times 30 residues. These final linear feature vectors were those used for the second stage classification.

A clear structural pattern emerges in the distributions of the three features (Fig. 5, left) in the training data. As expected, in the approximate region of the motif (residue range [−4,+4]) the probability of random coil is consistently high, while the probability of α-helix or β-pleated sheet is consistently low. Furthermore, the probability of finding β-pleated sheet anywhere in the vicinity of a cleavage site appears to be uniformly low. Beyond the immediate region of the motif, the random coil and α-helix probabilities are roughly complimentary to one another, with the median probability of finding an α-helix being around 90% or greater. However, the variability of these two differ dramatically between the N- and C-terminus sides of the motif region, with the N-terminus side showing lower variance.

2.6. *Stage 2 classifier: training on secondary structure*

The classification for this second stage was equivalent to an outlier detection task. For this we employed the One-Class SVM algorithm with a RBF kernel (see "Stage 2: Algorithm" in Fig. 2), implemented in scikit-learn (version 0.20.3).[31,32,36]

The training data for this classifier was the set of all 30-residue amino acid sequences centered on the dengue cleavage sites (2297 unique sequences in total), described in Sec. 2.1. These sequences were then represented with the structure features described in Sec. 2.5. The hyperparameter ν, which represents the upper bound on the fraction of margin errors,[36,37] was set to $\nu = 10^{-4}$. For testing, 10%

Table 1. Loadings (expressed as percentages: $0.XX \rightarrow XX$) for PC1 in Fig. 4. +/- loading indicates +/- correlation with PC1.

Motif pos.	1	2	3	4	5	6	7	8
Polar req.	−03	−05	−12	−13	+06	+13	+08	+20
Hydrop.	+11	+14	+25	+27	−03	−12	−07	−24
Iso. point	−13	−18	−29	−30	+05	+03	+01	+00
Mol. vol.	−02	−00	−15	−16	+18	+03	+14	−03
Sneath 1	+00	+09	+03	+04	+02	−00	+10	−20
Sneath 2	+02	−02	+13	+16	−10	−14	−20	−12
Sneath 3	−01	−03	−08	−08	+08	+00	+03	+01
Sneath 4	+00	+03	+11	+11	−29	−08	−06	−02

of the training data was held out, on which the trained classifier scored accuracy = 97.6%. The second stage classifier was then applied to the sites identified by the first stage classifier (see "Structure Classifier" in Stage 2 of Fig. 2), effectively functioning as an outlier detection filter on the first stage.

3. Results

3.1. *Stage 1: predicted motifs in human proteins*

The first stage classifier positively identified 3991 putative cleavage sites within the 6060 membrane-associated proteins. The amino acid distribution for each position in the motif substrate can be seen in the sequence logo (Fig. 3, right).[21] As expected, the same prevailing motif signatures are present in the training and predicted sets (compare left and right logos Fig. 3)—namely, the highest information content residues are those in positions three, four and five, with the first two being predominantly basic amino acids (arginine and lysine) and position five being predominantly polar amino acids (serine and threonine). However, there are some notable differences. For one, the total information content at every position in the motif substrate is lower for the predicted sites compared to that of the positive training data in Fig. 3. This indicates that the predicted cleavage motif sequences demonstrate greater diversity than that found in the training data. Furthermore, at positions three, four, and five of Fig. 3 there is clear signal of amino acids with differing biochemical properties—namely, acidic (glutamic acid), basic (histidine), and polar (asparagine, glutamine)—which are not evident in the sequence logo of the training data. The source of these variants are several low occurrence motifs found at the NS4A/NS4B and NS4B/NS5 cleavage sites. Though these variant cleavage motifs represent only about 1% of the training data, the classifier is still capable of leveraging these to identify such motifs within the target proteins.

To further explore how the predicted sites compare to the training data, we applied the same PCA transformation to the predicted motifs that was used for the training data (Fig. 4). The predicted motifs roughly cluster along the same dimensions as the positive and negative training data sets, as indicated by PC1 (not shown: the predicted motifs are not distinguishable from the negative sites for PC3 and higher). In order to interpret the biochemical features encoded in PC1, we must examine the component's loadings, shown in Table 1 organized by substrate position and biochemical feature. Loadings, which range from -1 to +1, are the weights that the term contributes to the principal component. They represent how well the given term correlates with the principle component value—a -1 indicates anti-correlation, 0 indicates uncorrelated, and +1 indicates positive correlation.

Since PC1 captures the distinction between cleaved and uncleaved data points (Fig. 4), the larger the magnitude of the loading the more informative that component is for discriminating between uncleaved and cleaved motifs. For example, hydropathy at position 4 has a loading value of +0.27

indicating that PC1 is directly correlated with hydropathy, i.e. having a hydrophobic amino acid at this position is correlated with the motif being uncleaved. Looking at Table 1 we see that all positions in the motif contribute significantly, in at least one biochemical property, to discriminating between cleaved and uncleaved (e.g. has a loading $>10\%$). Furthermore, every biochemical property contributes significantly in at least one substrate position.

By adding the magnitudes of all the loadings at a given substrate position, we see that positions three, four, five, and eight are the most important. This is consistent with the information content per position seen in the sequence logos (Fig. 3) with the notable exception of position eight. The most significant biochemical property for the motif classifier is hydropathy, for which the loadings of positions 1–4 are significantly positively correlated—consistent with the abundance of polar and charged amino acids located upstream of the cleavage sites, seen in the positive motif plot in Fig. 3, far left. Ultimately, the loadings for PC1 appear to capture and explain the cleavage signature present in the training data.

3.2. *Stage 2: predicted sites after filtering for secondary structure*

The second stage classifier identified 257 sites (of the 3991 passing the first stage) as having secondary structure similar to the known cleavage sites in the dengue polyprotein. The right plot of Fig. 5 shows the distribution of the three structure features as a function of position for both the 257 that passed the second stage ("Passed") and the 3734 that failed ("Failed").

Unsurprisingly, the secondary structure was consistent between the "Passed" and "Failed" sites in the region of the motif. Hence the second stage is not discriminating between the two groups based on the structure of the motif sequence itself. Beyond the region of the motif, however, differences in secondary structure between the two are evident. Additionally, for both groups the presence of β-sheets was uniformly low (with the "Passed" sites being marginally lower than the "Failed" sites). The "Failed" sites have uniformly low probability of α-helix (median probability $\sim 10\%$) and uniformly high probability of random coil (median probability $\sim 75\%$) in the vicinity of the motif. This suggests these sites may not be as proximal to transmembrane domains as were those of the dengue polyprotein. Ultimately, as can be seen by comparing the two plots of Fig. 5, the group of sites that passed the second stage clearly demonstrate local secondary structure similar to the dengue polyprotein training data.

3.3. *Classification of the STING cleavage site*

As only one human target is known, we examined our classifier on the human STING protein.[5] The known site, SRYRGSYW, is only weakly cleaved,[5] and hence is considered a particularly difficult case. In fact, the standard PWM approach to classifying motifs fails to identify this site as cleaved (see Sup. Fig. 1 and 4). Unfortunately, our classifier's first stage also indicated this site would not be cleaved. Importantly, there exists two non-human primate variants at this site in STING (marmoset: SRYQGSYW and chimp: SRYWGSYW), known to not be cleaved by the dengue protease.[5] Our first stage classifier correctly identifies these as negatives. For insight into these results, we plot the principal components for these three sites and compare them to the training data (see Sup. Fig. 1). We observe that the known cleaved site is indeed not supported by any positive training data points in its vicinity. Additionally, the two negative motif variants are found to be closer to the center of the negative training data (larger PC1) than the human STING motif (see Sup. Fig. 1), as indicated by the black arrow. So while our training data fails to lead to a model that captures STING, the result suggests our interpretation of PC1 as discriminating between cleaved and not cleaved is correct. Furthermore, it also suggests our model represents a conservative estimate of the diversity in the proteases cleavage specificity.

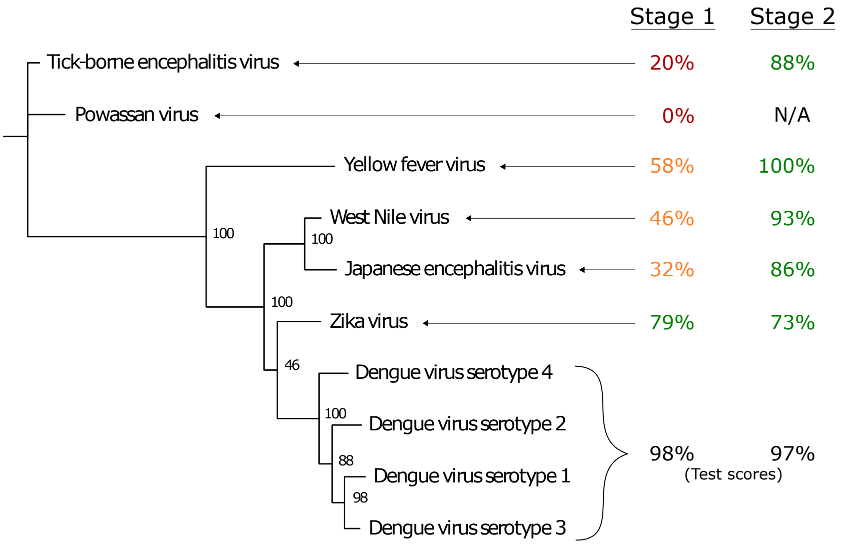

Fig. 6. Phylogeny of other flaviviruses with comparable protease—generated from the NS3 protein sequence. Here we show the performance of the dengue-trained classifier when classifying known cleavage sites in other flavivirus polyproteins. Stage 1 values represent the percentage of correctly classified motifs in each of the other viruses. Stage 2 values represent the percentage of those sites that pass Stage 1 that also pass the second stage—red $0-30\%$, orange $30-60\%$, green $>60\%$.

3.4. *Classification of other flaviviruses*

An alternative way of validating our method is to see how it performs on other flaviviruses at various evolutionary distance from dengue. The conservation of cleavage motifs would be expected to increase with decreasing evolutionary distance. Thus, motifs from flaviviruses that are more closely related to dengue would be predicted to be cleaved more often by the first stage than those of more distantly related viruses.

In this manner, we applied our classifier to other flaviviruses with a polyprotein encoding a protease comparable to that of dengue. We found complete polyprotein sequences for six other flaviviruses. Using the NS3 protein sequence from each, we generated a phylogenetic tree of all seven viruses (Fig. 6), which matches the relationship of the flavivirus family predicted using whole genome alignments. Indeed, the percentage of cleavage motifs that are correctly classified by the dengue-trained first stage is inversely correlated with evolutionary distance from dengue, consistent with our expectation.

Furthermore, since the secondary structure of the viral polyprotein is important for viral replication, there should be less variability in this secondary structure between flaviviruses than would be present in the primary sequence of the cleavage motifs. In other words, the secondary structure in the vicinity of each cleavage motif should be mostly unchanged between flaviviruses, regardless of evolutionary distance. Therefore, we would expect the second stage classifier to positively classify all sites that pass the first stage, since the secondary structure of the other viruses would be comparable to that of dengue. As Fig. 6 shows, the percentage of sites that pass the second stage of classification is roughly equal to that of the first stage, for all six flaviviruses, consistent with our expectation.

4. Discussion

Flaviviruses manipulate the host cellular environment in several ways during infection.[38] By creating protective structures in the ER membrane, they evade recognition by immune sensing proteins in the cytoplasm.[39,40] By actively inhibiting host proteins, including by action of the viral protease, they hinder the immune system response.[4,5] We have developed a method to predict new host-virus interactions computationally; specifically, to predict targets of the dengue virus protease using a two-step classifier trained on the motifs in the dengue polyprotein that are internally cleaved.

One of our explicit design goals was to leverage the known protease cleavage sites towards a better understanding of the underlying biochemical properties and local structure that results in protease cleavage. Hence we utilized a biochemical encoding of the amino acids rather than a typical sequence-based approach and a second stage that captured site's local structure. The first stage of our approach identified hydropathy as the biochemical property that most influenced classifier performance, though all properties were influential in accurate classification. In the second stage, we found a preference for random coiled regions at the site of cleavage. Additionally, the C terminus side of the motif appears more loosely constrained than the N terminus side, possibly illuminating some structural preference of the protease.

Furthermore, we sought a method that was sensitive to the variation contained within a small training dataset, as most viruses have only a limited set of known targets. For example, the one known human target of the dengue protease, STING, is unlike any of the known dengue cleaved sites. We believe this speaks to how the flexibility of the protease has evolved— it can still be active even when mutations arise in the viral genome that change the motif sequence in a biochemically consistent way. If we include the identified STING site into our training data, our classifier method recovers this site and others that are sufficiently similar (data not shown). Importantly, in that scenario STING represents only 0.5% of the training data set, and probabilistic methods do not alter the underlying probabilities in a manner that meaningfully alters the predictions. In contrast, our SVM classifier is highly responsive to individual new instances within the training data. Here we relied on the known dengue cleavage sites in order to generate a conservative estimate of what the protease could target in order to identify high confidence candidates (257 total) for subsequent experimental testing. Moving forward, as additional sites are validated as being cleaved, they can be included in the training data and our prediction list can be easily refined.

Our method has a number of limitations. First, secondary structure, though informative, provides an incomplete picture of accessibility—full protein structure is needed to fully determine site accessibility. Second, our method does not account for the likelihood of co-localization of any target protein and the protease in the cell. Third, the scope of our predictions is limited by the small training data set. Finally, our predictions are not informative about cleavage efficiency.

As a means of computational validation, we tested whether cleaved motifs from other flaviviruses (known positives, cleaved by each virus's own protease) were correctly classified using the dengue training data. We found that the accuracy of the motif classifier decreases with evolutionary distance, e.g. motifs from flaviviruses that are more closely related to dengue are predicted to be cleaved more often than those that are more distantly related. This result is consistent with the specificity of the protease evolving as the viruses diverge. However, it is possible the dengue protease is able to cleave the critical positions of the other viruses, despite not being represented in the dengue virus training data. Experimentally testing whether yellow fever virus motifs (or motifs from any other flaviviruses diverged from dengue) are cleaved by the dengue virus protease would tease apart these two scenarios and refine our training data.

At present, the number of sequences available for other flavivirus polyproteins is too small to build models for each virus. However, as the catalogue of flavivirus genome sequences continues to expand, our method could be applied more broadly. This method may one day be able to predict the targets of other flaviviruses such as Zika, in a similar manner. Predicting host-virus interactions in flaviviruses may have major implications for future drug discovery or treatment options. Ultimately, though we have only considered this computational method in the context of viral protease targets, in principle, this method can be applied to any protein-protein targeting application in which the biochemistry of the targeted site and the surrounding secondary structure play a role in the success of the protein-protein interaction.

5. Acknowledgements

We would like to thank Jessica Westfall, Jonathan Rubin, and Rutendo Sigauke for all of their input. This project was funded by a 2017 Innovative Seed Grant (CU Boulder).

References

1. S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh, M. F. Myers, D. B. George, T. Jaenisch, G. R. W. Wint, C. P. Simmons, T. W. Scott, J. J. Farrar and S. I. Hay, The global distribution and burden of dengue, *Nature* **496**, 504 (Apr 2013).
2. S. A. Shiryaev, I. A. Kozlov, B. I. Ratnikov, J. W. Smith, M. Lebl and A. Y. Strongin, Cleavage preference distinguishes the two-component NS2BNS3 serine proteinases of Dengue and West Nile viruses, *Biochemical Journal* **401**, 743 (Feb 2007).
3. J. Li, S. P. Lim, D. Beer, V. Patel, D. Wen, C. Tumanut, D. C. Tully, J. A. Williams, J. Jiricek, J. P. Priestle, J. L. Harris and S. G. Vasudevan, Functional profiling of recombinant NS3 proteases from all four serotypes of dengue virus Using tetrapeptide and octapeptide substrate libraries, *Journal of Biological Chemistry* **280**, 28766 (Aug 2005).
4. S. Aguirre, A. M. Maestre, S. Pagni, J. R. Patel, T. Savage, D. Gutman, K. Maringer, D. Bernal-Rubio, R. S. Shabman, V. Simon, J. R. Rodriguez-Madoz, L. C. F. Mulder, G. N. Barber and A. Fernandez-Sesma, DENV inhibits type I IFN production in infected cells by cleaving human STING, *PLoS Pathogens* **8**, p. e1002934 (Oct 2012).
5. A. C. Stabell, N. R. Meyerson, R. C. Gullberg, A. R. Gilchrist, K. J. Webb, W. M. Old, R. Perera and S. L. Sawyer, Dengue viruses cleave STING in humans but not in nonhuman primates, their presumed natural reservoir, *eLife* **7**, 1 (Mar 2018).
6. Q. Ding, J. M. Gaska, F. Douam, L. Wei, D. Kim, M. Balev, B. Heller and A. Ploss, Species-specific disruption of STING-dependent antiviral cellular defenses by the Zika virus NS2B3 protease, *Proceedings of the National Academy of Sciences* **115**, E6310 (jul 2018).
7. J. Song, F. Li, A. Leier, T. T. Marquez-Lago, T. Akutsu, G. Haffari, K.-C. Chou, G. I. Webb and R. N. Pike, PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy, *Bioinformatics* **34**, 684 (Feb 2018).
8. J. H. Korhonen, K. Palin, J. Taipale and E. Ukkonen, Fast motif matching revisited: high-order PWMs, SNPs and indels, *Bioinformatics* **33**, p. btw683 (dec 2016).
9. C. E. Grant, T. L. Bailey and W. S. Noble, FIMO: scanning for occurrences of a given motif, *Bioinformatics* **27**, 1017 (apr 2011).
10. R. C. McLeay and T. L. Bailey, Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data, *BMC Bioinformatics* **11**, p. 165 (Dec 2010).
11. F. Zambelli, G. Pesole and G. Pavesi, Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes, *Nucleic Acids Research* **37**, W247 (jul 2009).
12. D. S. Chekmenev, C. Haid and A. E. Kel, P-Match: transcription factor binding site search by combining patterns and weight matrices, *Nucleic Acids Research* **33**, W432 (jul 2005).
13. A. Kel, MATCHTM: a tool for searching transcription factor binding sites in DNA sequences, *Nucleic Acids Research* **31**, 3576 (Jul 2003).
14. G. D. Stormo, Modeling the specificity of protein-DNA interactions, *Quantitative Biology* **1**, 115 (2013).
15. G. D. Stormo, T. D. Schneider, L. Gold and A. Ehrenfeucht, Use of the Perceptron' algorithm to distinguish translational initiation sites in E. coli, *Nucleic Acids Research* **10**, 2997 (1982).
16. I. Umareddy, O. Pluquet, Q. Wang, S. G. Vasudevan, E. Chevet and F. Gu, Dengue virus serotype infection specifies the activation of the unfolded protein response, *Virology Journal* **4**, p. 91 (2007).
17. B. E. Pickett, E. L. Sadat, Y. Zhang, J. M. Noronha, R. B. Squires, V. Hunt, M. Liu, S. Kumar, S. Zaremba, Z. Gu, L. Zhou, C. N. Larson, J. Dietrich, E. B. Klem and R. H. Scheuermann, ViPR: an open bioinformatics database and analysis resource for virology research, *Nucleic Acids Research* **40**, D593 (jan 2012).

18. R. C. Edgar, MUSCLE: A multiple sequence alignment method with reduced time and space complexity, *BMC Bioinformatics* **5**, 1 (2004).
19. C. E. Stocks and M. Lobigs, Signal peptidase cleavage at the flavivirus C-prM junction: dependence on the viral NS2B-3 protease for efficient processing requires determinants in C, the signal peptide, and prM., *Journal of virology* **72**, 2141 (Mar 1998).
20. F. Meng, R. A. Badierah, H. A. Almehdar, E. M. Redwan, L. Kurgan and V. N. Uversky, Unstructural biology of the dengue virus proteins, *FEBS Journal* **282**, 3368 (Sep 2015).
21. G. E. Crooks, G. Hon, J.-m. Chandonia and S. E. Brenner, WebLogo: a sequence logo generator., *Genome research* **14**, 1188 (Jun 2004).
22. L. Käll, A. Krogh and E. L. Sonnhammer, A combined transmembrane topology and signal peptide prediction method, *Journal of Molecular Biology* **338**, 1027 (May 2004).
23. D. C. Mathew and Z. Luthey-Schulten, On the physical basis of the amino acid polar requirement, *Journal of Molecular Evolution* **66**, 519 (May 2008).
24. C. Woese and D. Dugre, The molecular basis for the genetic code., *Proceedings of the National Academy of Sciences of the United States of America* **55**, 966 (1966).
25. J. Kyte and R. F. Doolittle, A simple method for displaying the hydropathic character of a protein, *Journal of Molecular Biology* **157**, 105 (may 1982).
26. G. R., Amino acid difference formula to help explain protein evolution., *Science.* **185**, 862 (1974).
27. C. Alff-Steinberger, The genetic code and error transmission., *Proceedings of the National Academy of Sciences of the United States of America* **64**, 584 (Oct 1969).
28. P. H. Sneath, Relations between chemical structure and biological activity in peptides, *Journal of Theoretical Biology* **12**, 157 (1966).
29. R. Thomson, T. C. Hodgman, Z. R. Yang and A. K. Doyle, Characterizing proteolytic cleavage site activity using bio-basis function neural networks, *Bioinformatics* **19**, 1741 (sep 2003).
30. B. E. Boser, I. M. Guyon and V. N. Vapnik, A training algorithm for optimal margin classifiers, in *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*, (ACM Press, New York, New York, USA, 1992).
31. L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt and G. Varoquaux, *API design for machine learning software: experiences from the scikit-learn project*, Sep 2013).
32. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, Scikit-learn: Machine Learning in Python, *Journal of Behavioral & Applied Management* **70**, 1 (Jan 2012).
33. R. Batuwita and V. Palade, *Imbalanced Learning* (John Wiley & Sons, Inc., Hoboken, NJ, USA, Jun 2013).
34. W. Kabsch and C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* **22**, 2577 (Dec 1983).
35. A. Drozdetskiy, C. Cole, J. Procter and G. J. Barton, JPred4: a protein secondary structure prediction server, *Nucleic Acids Research* **43**, W389 (Jul 2015).
36. C.-C. Chang and C.-J. Lin, Training v-support vector classifiers: theory and algorithms, *Neural Computation* **13**, 2119 (Sep 2001).
37. B. Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola and R. C. Williamson, Estimating the support of a high-dimensional distribution, *Neural computation* **13**, 1443 (2001).
38. M. S. Diamond, Evasion of innate and adaptive immunity by flaviviruses, *Immunology and Cell Biology* **81**, 196 (jun 2003).
39. J. A. den Boon and P. Ahlquist, Organelle-Like Membrane Compartmentalization of Positive-Strand RNA Virus Replication Factories, *Annual Review of Microbiology* **64**, 241 (oct 2010).
40. B. L. Fredericksen and M. Gale, West nile virus evades activation of interferon regulatory factor 3 through RIG-I-dependent and -independent pathways without antagonizing host defense signaling, *Journal of Virology* **80**, 2913 (2006).