Redefining the boundary between self and non-self: MHC class I splicing and the immunopeptidome

Michael Ellis

Supervisor: Dr Adrian Shepherd

Department of Crystallography,
Birkbeck, University of London

Abstract

Peptide presentation by MHC-I molecules is an essential part of the immune system. Until recently, the presentation of spliced peptides - that is, peptides that have been formed by linking two or more peptide fragment into a new one - was thought to be a rare occurence. A recent paper by Liepe et al. (2016) [1], showed instead that they constitute around one quarter of all antigenic peptides. In this project, I hope to explore the distinctive properties of spliced peptides and build a classifier to predict whether a potential spliced peptide is presented.

Contents

Introduction

Background

T cells are the police force of the immune system - their role is to identify cells containing bacteria, viruses or mutant proteins and to destroy them. However, since T cells cannot directly interact with the contents of a cell, they require an alternative way of seeing what is inside a cell and detecting abnormalities. The method selected by millions of years of evolution is the presentation of thousands of small protein fragments on the surface of cells. These peptides are derived from proteins broken down in the cytosol of cells. T cells are able to recognize them as "self" - in the case of peptides from normal proteins in cells - or as "non-self" - in the case peptides derived from pathogens or mutated proteins (e.g. cancer) - and act accordingly.

The steps in this process are well understood in general, though there is still much to discover about many of the specifics. Briefly, proteins are tagged for degredation with ubiquitin and delivered to 20S proteasome. Inside the enclosed cavity of the proteasome there are three types of catalytic site, with two copies of each, which cleave the protein into smaller fragments. The cleavage occurs at particular points in the protein sequence based on the properties of surrounding residues at that point, with each site having preferences for different motifs [2]. The protein is digested until the fragments are sufficiently small, typically 3-22 residues [3], and most of these fragments are subsequently destroyed by cytosolic peptidases [4]. However, some are transported to the endoplasmic reticulum (ER) by the "transporter associated with antigen processing" (TAP) and awaiting these peptides in the ER are MHC I molecules. Of the peptides that reach the ER, a fraction bind sufficiently strongly to the MHC I molecules - if they are longer than 8-12aa they can be trimmed to fit the MHC I molecule - to be further transported - as a part of peptide-loaded MHC complex - to the surface of the cell where they are presented for recognition by T cells.

So we have three distinct steps in the path from protein to presentation that determine which peptides become antigens:

1. Cleavage in the 20S proteasome.
2. Binding to TAP transporters.
3. Binding to MHC I molecules.

Each of these adds contraints to the set of peptides that can be presented.

For a long time, it has been known that there is an optional fourth step in this process that takes place alongside cleavage in the proteasome: cleaved fragments can be joined together to form a new peptide with a sequence that does not occur in the original protein [5]. This peptide splicing was thought to be a rare event and to be unimportant to the general functioning of the immune system. In fact, until recently, only five presented spliced peptides had ever been discovered.

However, in 2016, Liepe et al. [1] used a novel approach to analyse the mass spectrometry (MS) data of peptides eluted from the MHC-I molecules on surface of cells. They discovered thousands of new antigenic spliced peptides - making up around a quarter of the total number of peptides (both spliced and nonspliced) identified by them - showing that these presumed rare objects were in fact a significant part of the immune system and that any understanding of human response to pathogens would be incomplete without an understanding of these new epitopes.

Predicting presented peptides is extremely useful - it can be used to produce peptide vaccines in which only a small fragment of a pathogen is required with obvious benefits over conventional vaccines. For nonspliced peptides there are tools available to predict each of the steps listed above (e.g. NetChop for cleavage [6], PRED(TAP) for TAP transport [7] and NetMHC for MHC binding [?]). The second and third of the three processes above are thought to be the same for both spliced and nonspliced peptides (though not the first, as we shall discuss later) and so the tools for these can be similarly applied to spliced peptides with the caveat that many of these use machine learning algorithms trained on nonspliced peptides and so can perform poorly when applied to spliced peptides [1]). By far the biggest barrier to the accurate prediction of antigenic spliced peptides is the lack of knowledge about the splicing process itself.

The mechanism for splicing is understood to occur via an acyl-enzyme intermediate. At the cleavage site, a nucleophilic attack of the peptide

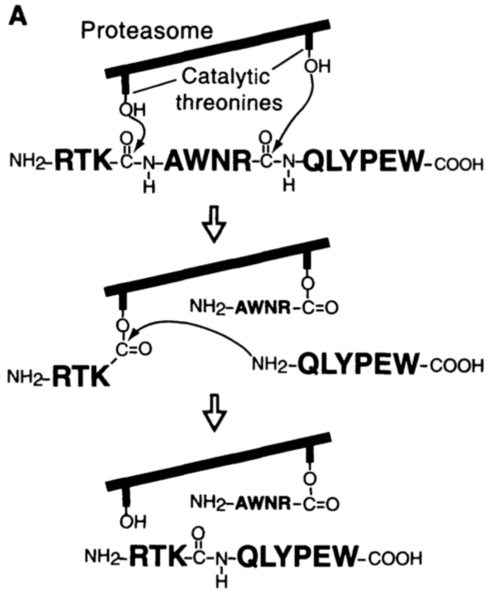


Figure 1: Figure from Vigneron et al. (2004) [8] - proposed splicing mechanism via an acyl-enzyme intermediate

bond occurs, resulting in the formation of the acyl-enzyme intermediate between the peptide on the C-terminal side of the cleavage and the catalytic site. Normally, this intermediate is rapidly hydrolyzed and the fragment released. However, in splicing, the N-terminal of another fragment successfully competes with water and forms a new peptide bond and hence a new peptide [8]. Figure 1 shows this mechanism. An important thing to note is that if this mechanism is correct, we may expect to see inefficient cleavage sites in the protein sequence to be preferred in spliced peptides, because they would allow the N-terminal peptide to remain bound as part of the acyl-enzyme intermediate for longer, giving the C-terminal peptide more time to bind.

Mishto et al. (2012) [9] propose an additional compenent to the above mechanism. They suggest that the C-terminal precursor may sit in a pocket adjacent to the catalytic site prior to splicing, allowing it to more easily out-compete water molecules and also potentially adding more contraints to the peptides that can splice - e.g if the pocket is hydrophobic this will give it a preference for particular types of residue in the C-terminal precursor. This mechanism is shown in figure 2.

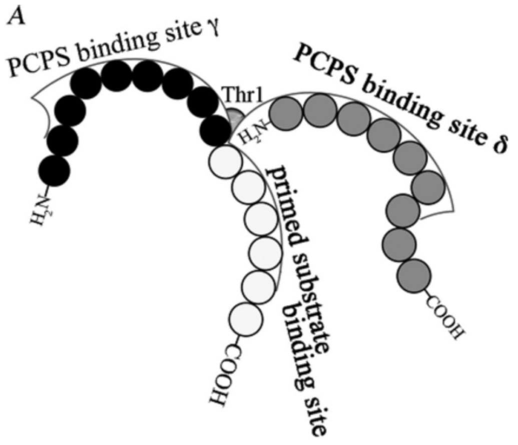


Figure 2: Figure from Mishto et al. (2012) [9], show their proposed splicing mechanism in which, prior to splicing, the C-terminal precursor sits in a pocket,δ, near the catalytic site and the N-terminal precursor sits in a pocket, γ, and is cleaved from the rest of its substrate.

In August 2017, Platteel et al. (2017) [10], proposed a multilevel strategy for identifying spliced epitopes targeted by CD8$^+$ T cells during infection. This stategy is fairly simple. Essentially, the idea is to take all possible spliced peptides from a pathogen and simply rank them by their predicted binding affinity to the MHC I molecules within the cell line used. They were able to identify two novel spliced antigens in this way.

Aim

In this project I use the large dataset of spliced peptides identified by Liepe et al. to examine the properties of spliced peptides relative to nonspliced; to explore and validate proposed splicing rules; and ultimately, to create what is - to my knowledge - the first example of a machine learning classifier for identifying spliced peptides.

Materials and Methods

Data

The set of spliced and unspliced peptides whose properties were investigated and upon which the classifers were trained consists entirely of those peptides identified by Liepe et al. (2016). These peptides were eluted from the HLA-I molecules of three unrelated cell lines - GR lymphoblastoid (GRLCL), CIR lymphoid [11] and primary human fibroblasts [12] - and analysed via liquid chromatography-mass spectrometry (LC-MS/MS). This LC-MS/MS data was then queried against both the Swissprot human proteome database, which does not include spliced peptides, and a second vast database of all possible spliced 9mer-12mer peptides with an intervening sequence of fewer than 20 residues, to generate the sets of identified unspliced and spliced peptides, respectively.

These were extracted from the supplementary materials using a web tool [13]. To analyse the residues within the peptide sequences, a python script was written to calculate the frequencies of each residue at each position for all three of the cell lines (Appendix A).

In order to properly examine the residues involved in the cleavage and splicing of peptides, a Python script was written to remove any "ambiguous" peptides from the dataset (A). That is, approximately 10% of the peptides were identified with multiple possible origins within the proteome, for which it may be the case that only one origin is correct.

To obtain the residues relevant to proteolytic cleavage that are outside the peptide sequence (e.g for the spliced peptide [DKG][VTFDID], we want the residues either side of [DKG] and [VTFDID]), I downloaded and indexed the human proteome using the python module *pyfaidx* (Appendix A), and wrote a python script to retrieve the two preceding and succeeding residues corresponding to each peptide fragment (Appendix A).

NetChop cleavage predictions for the 20S proteasome for the spliced and unspliced peptides were obtained by downloading the NetChop program and running it locally over the complete human proteome [6]. The peptide termini were then matched to the output of this to determine whether the ends of peptide fragments were indeed predicted to be cleavage sites (Appendix A).

To analyse the properties of residues either side of the splicing sites a python script was written to group the residues by property (hydrophobicity, charge and polarity, and size) (Appendix A). We obtain some measure of the importance of these properties in splicing, we compare the frequencies of pairs of residues in positions P2-P1, P1-P1', and P1'-P2' to the frequencies of all pairs of residues found between the anchor positions (position 2 and the C-terminal residue) in nonspliced peptides. The residues are grouped by property as in table 1.

Label	**Property**	**Residues**
T	Tiny	A,G,S
S	Small	C,D,N,P,T
M	Medium	E,H,Q,V
L	Large	I,K,L,M,R
V	Very Large	F,W,Y
P	Polar	C,Q,N,S,T,Y
N	Non-polar	A,F,G,I,L,M,P, V,W
-	Negatively Charged	D,E
+	Positively Charged	H,R,K
O	Hydrophobic	A,C,F,G,I,L,M, P,W,Y,V
I	Hydrophilic	D,E,H,K,N,Q,R, S,T,

Table 1: Residues grouped by size, charge and hydrophobicity.

The ic50 binding values predicted by the SMM method [14] were obtained via the software available from the IEDB. The NetMHC predictions were obtained from the CBS prediction server [15].

Classifiers

The negative data for the classifiers, i.e. potential spliced peptides not presented on the surface of cells by HLA-I molecules, were randomly generated from the indexed proteome (Appendix A). It is possible that a proportion of this set are in fact real spliced peptides not yet identified but, given the vast number of potential 9-12mer spliced peptides (1.6×10^{10}) relative to the much smaller number of HLA-I-presented spliced peptides, this proportion is almost cetrainly small enough that a classifier can still be adequately trained on the data.

Prior to all training and testing of any classifiers, relevant features were created for the spliced and invented datasets and these were transformed to a form useable by classifier (Appendix A). The features included in all my classifiers were: the amino acid at each position; the two amino acids either side of the splicing site; the length of the first peptide fragment; whether the fragments were reversed relative to the order they appear in their original protein; the distance between fragments; and the two residues preceding the N-terminus or succeeding the C-terminus of a peptide fragment (from this point refered to as *cleavage residues* for brevity). An additional feature included in some cases was the ic50 binding value predicted by the stablised matrix method (SMM) [14]. For each feature respresented by a letter for an amino acids, the letter was encoded into a binary vector of length 21, with each position in the vector representing one of the 20 amino acids and an additional position for NA (conventional sparse encoding [16]). In order to limit the number of features, I restricted my attention to 9mer peptides - the most frequent length - for all of the classification.

In all cases of training and testing the classifiers, the data was split into two sets: a training set, for fitting the classifiers and tuning the hyperparameters; and a test set, for evaluating the final model. In each case the train/test split was 80/20.

The spliced peptide data was separated by which cell line they originated from. The GRLCL cell line data was further split into groups based on the MHC allele that each peptide was likely to have been bound too. This was determined by using the SMM method to predict binding and assigning a peptide to an allele if it was predicted to bind (ic50¡500nM) and if that allele was the best scoring for that particular peptide. Only 5 of the 6 alleles present in the GRLCL cell lines were capable of being predicted by the IEDB SMM method, so there are no predicted binders for HLA-C*02:02.

The classifiers were trained and tested on the data in four different ways:

- Using data from cell lines with multiple alleles (i.e. GRLCL and fibroblasts) to train and test a classifier for each cell line.

- Training a classifier on the combined data from GRLCL and fibroblast cell lines and testing on the C1R cell line.

- Training and testing on the C1R cell line, which only contains the HLA-B*40:02 allele, including binding predictions as an additional feature.

- Training and testing on each of the GRLCL allele-specific subsets (HLA-A*01:01, -A*03:01, -B*07:02, -B*27:05, and -C*07:02)

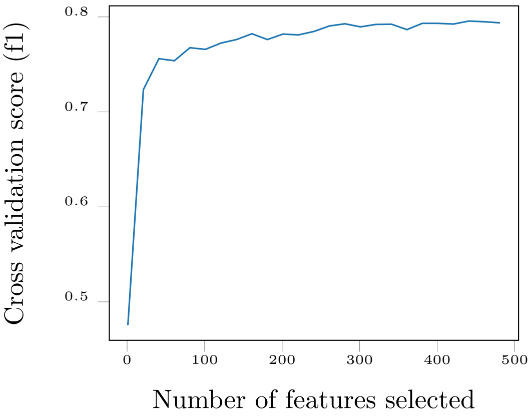


Figure 3: Feature selection for a random forest classifier on the GRLCL cell line. Performance improves as more features are added.

In each case, three different types of classifier were used - random forest, neural network, and support vector machine. For each classifier, their hyperparameters were tuned on the training set using 3-fold cross-validation and the best performing model was then applied to the relevant test set. The optimum number of features was also calculated during this process but in each case it was found to be optimal to use all available features (Figure 3). The results of the hyperparameter tuning are found in appendix B.

Results

Properties of Spliced Peptides

Examining the relative frequencies of residues between spliced and nonspliced peptides at different positions in the peptide we can immediately see that these two classes of peptide are very different from one another (see appendix C).

Proline in position 1 in the GRLCL cell line stands out in particular, it is present in 25% of the spliced peptides, compared to only 0.7% of nonspliced peptides (figure 4). We don't see this replicated to quite the same extent in the C1R and fibroblast cell lines though we do still find that pro-

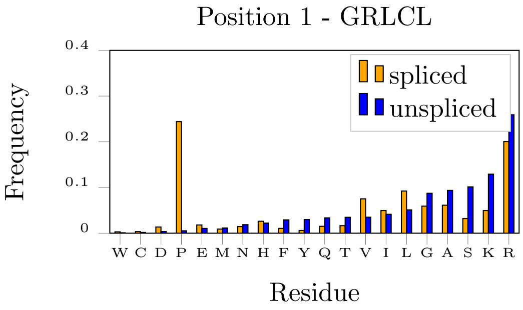


Figure 4: Comparison of frequencies of amino acids at position 1 in spliced and nonspliced peptides from the GRLCL cell line

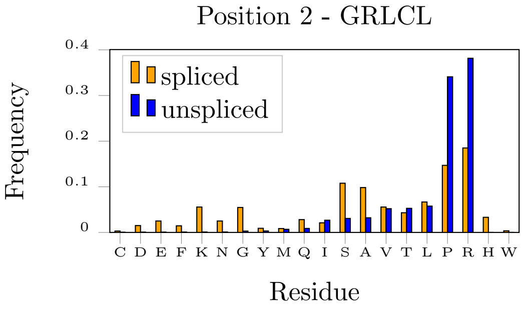


Figure 5: Comparison of frequencies of amino acids at position 2 in spliced and nonspliced peptides from the GRLCL cell line

line is over represented by a factor of 10 in each compared with the factor of 40 in GRLCL.

Beyond this, there is no clear, discernable pattern in the preferences of spliced peptides for particular amino acids. We can perhaps note that spliced peptides appear be less selective to some degree. For example, in position 2 for the GRLCL cell line D, E, F, N and G very rarely occur within unspliced peptides but they are all reasonably common in spliced peptides (figure 5).

One further interesting result we find from looking at residue frequencies at any point in the peptides is that, remarkably, while only 0.5% of unspliced peptides contain a cysteine residue, for spliced peptides this figure is 10%.

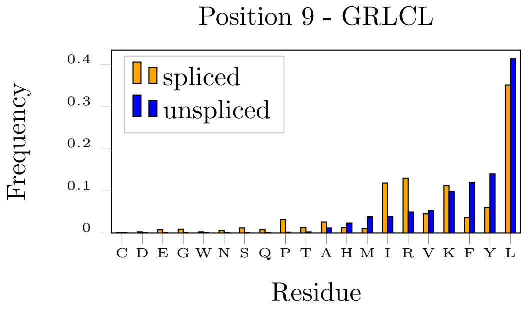


Figure 6: Comparison of frequencies of amino acids at position 9 in spliced and nonspliced peptides from the GRLCL cell line

Cleavage Sites

Residue Frequencies

By simply looking at the frequencies of different residues at cleavage site (figures 7 and 8) we notice some striking features. Most notably, proline - known to inhibit cleavage in positions P1 and P1' [17]- is found more commonly than in the proteome at both P1' for the second cleavage site and P1 for the third cleavage site, while for nonspliced peptides it occurs much less often - as we would expect from a cleavage inhibitor.

We find a few other similar curious results. At position P1', serine is seem to encourage cleavage in nonspliced peptides while being less common for spliced peptides than in the proteome. Conversely, leucine at position P1 discourages splicing in nonspliced peptides while being common in spliced peptides.

Most significantly, taking a complete view these results we can conclude that the rules that govern cleavage for nonspliced peptides don't appear to apply to spliced peptides and are sometimes completely wrong. This has implications for the tools used to predict cleavage sites, which may be ineffective for spliced peptides.

NetChop Evaluation

The tool NetChop20S 3.0 can be used to predict cleavage sites within proteins digested by the human proteasome. It has been trained on in vitro degradation data, in contrast to other tools trained on the C-terminal cleavage sites of MHC ligands.

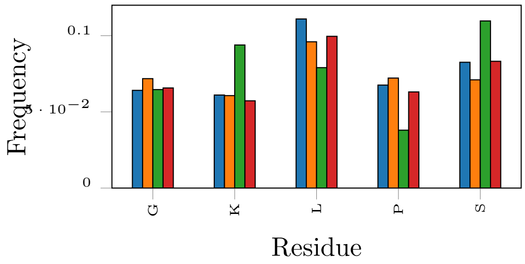


Figure 7: Frequencies of selected residues in the P1' position of C-terminal cleavage sites compared to the proteome. The frequencies for the N-terminal (blue) and C-terminal (orange) splice reactants have distinct distributions compared to unspliced peptides (green) and the proteome (red)

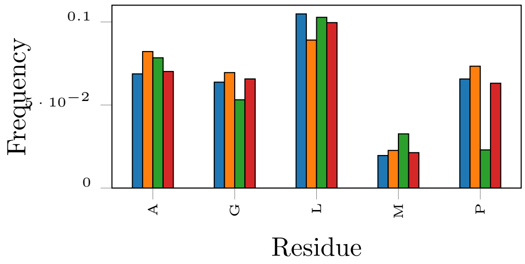


Figure 8: Frequencies of selected residues in the P1 position of N-terminal cleavage sites compared to the proteome. As with the C-terminal cleavage sites, the frequencies for the N-terminal (blue) and C-terminal (orange) splice reactants have distinct distributions compared to unspliced peptides (green) and the proteome (red)

By comparing the proportions of positive predictions made by this tool on the N- and C-terminal sites of spliced and unspliced peptides we can obtain a measure of how "cleavage-like" these sites are (Figure 9).

We find that the ends of spliced peptide precursors are predicted to be cleavage sites at a lower rate than the ends of nonspliced peptides, and hence are less likely to be typical cleavage sites. This agrees with the idea proposed by Mishto et al. (2012) that spliced peptides use uncommon cleavage sites [18]. However, this difference may also be due to incorrect labelling of the original location of the peptide precursors within the proteome - only fragments less that 20 amino acids apart were considered and this may be too restrictive to find the true locations and hence the true cleavage sites.

One thing to note in nonspliced peptides is that 70.6% of them are predicted to have a cleavage after their C-terminal residue compared with only 57.0% before their N-terminal residue. This type of difference can be explained by the existence of N-terminal peptide trimming in the endoplasmic reticulum (ER). Where peptides are too long to bind to MHC-I molecules, they are often reduced in size through the removal of N-terminal residues by ER aminopeptidases, hence abolishing the conjunction of N-terminus and cleavage site. This relationship is repeated in the spliced peptides as we would expect.

Most interestingly, of the C-terminal sites of the N-terminal precursor peptides (i.e. that right ends of the first peptide), only 42.2% are predicted to be cleavage sites. Given that, of all possible sites in the whole proteome, NetChop20 3.0 predicts 46.7% to be cleavage sites, 42.2% is significantly worse (p¡0.0001, using a binomial test) than random and suggests that these sites may be particularly atypical cleavage sites. This supports the idea proposed by Mishto et al. (2012) [?] (shown in figure **??**) that a slowly hydrolyzing N-terminal precursor peptide can increase the efficacy of splicing and, thus, the same thing that makes a site a poor cleavage site under normal conditions can make it a good site for the forming of spliced peptides.

Validation of Proposed Splicing Rules

Berkers et al. (2016)[19] propose rough rules that govern peptide splicing derived experimentally in

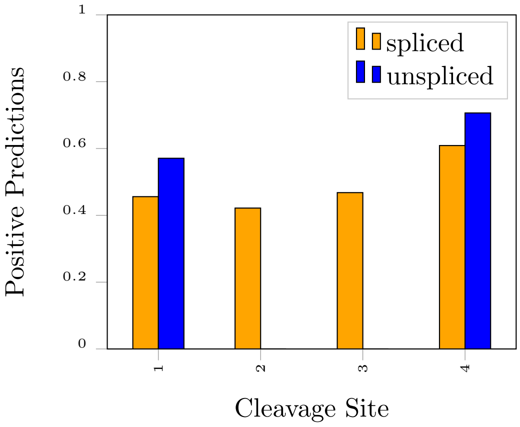


Figure 9: The proportion of positive cleavage predictions for the N- and C-terminals of peptides - 1 and 4, respectively - and for the N- a C-terminals of the splicing site within spliced peptides - 3 and 2, respectively. (For spliced peptides as they are normally represented, sites 1 to 4 are in order from left to right)

vitro. They propose that it is the N-terminal ligation precursor that primarily determines the ligation efficiency and that it is most efficient when there is a negatively charged or polar residue at P1 combined with a small or polar residue at P2, but also that it can participate in ligation if a hydrophobic residue at P1 is combined with a basic, small, or—to a lesser extent—polar or negatively charged residue at P2. (Also that it is most efficient when slowly hydrolyzed by the proteasome and therefore possesses a longer half-life but we cannot test this with the data we have). Utilising our much larger dataset we attempt to validate these rules.

In the GRLCL cell line, we find evidence that tiny and hydrophobic residues are preferred in the P2 and P1 positions of the splicing site (Figure **??**). However, we find the same pattern for the P1 and P1' positions and also for the P1' and P2' positions (Appendix C). This suggests that the differences we are observing in residue frequencies is due to general differences between the properties of spliced and unspliced peptides in this cell line rather than being specific to the positions relative to the splicing site. The GRLCL cell line has six different types of MHC-I molecules, if spliced and nonspliced peptides differ in their binding preferences - perhaps due to their anchor residues - we might expect different properties to be observed in the non-anchor residues (e.g some MHC molecules have an affinity for hydrophobic residues in these positions).

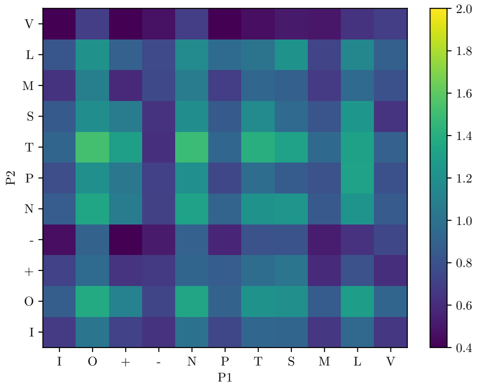


Figure 10: Relative frequencies in the GRLCL cell line of pairs of residues on the N-terminal peptide side of the splicing site compared to pairs of non-anchor in nonspliced peptides

In order to eliminate the effect of MHC alleles preferences among the peptides, we turn our attention to the C1R cell line. Our data from C1R contains fewer spliced and nonspliced than our data from GRLCL and so any statistical tests we carry out will have less power, but it also has the great benefit that the peptides within it have been eluted entirely from one allele, HLA-B*40:02.

In agreement with Berkers et al., using a chi-squared test we find that a tiny residue at P2 combined with a polar or negatively charged residue at P1 does indeed increase the efficacy of splicing (p¡0.001). However, we are unable to find a similar effect for a polar residue at P1 combined with a polar or negatively charged residue at P2. The most overrepresented pair of residues is a tiny residue followed by a tiny residue (Bonferroni-adjusted p¡0.0001). Considering each amino acid separately - rather than grouping them - and applying a Bonferroni correction, we find three pairs are, significantly overrepresented in the P2 and P1 positions (p¡0.05): GS, DA and SV.

For the P1 and P1' pairs, we observe a preference for tiny residues at P1 conbined with hydrophobic, small or non-charged residues at P1'. We find three

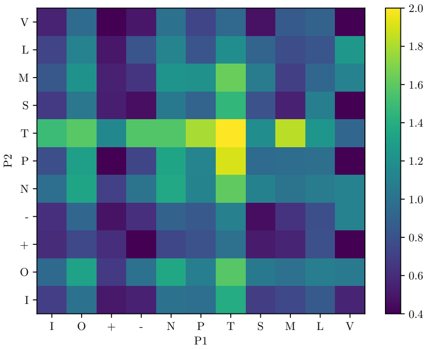


Figure 11: Relative frequencies in the C1R cell line of pairs of residues on the N-terminal peptide side of the splicing site compared to pairs of non-anchor in nonspliced peptides

pairs of residues are overrepresented (adjusted p¡0.05): LL, SL and GL.

For the pairs P1' and P2', it is clear from comparing figure 13 to figures 12 and 11 that the C-terminal precursor has less stringent requirement on the types of residues it can contain. This is in agreement with Berkers et al. (2016), that the N-terminal precursor primarily determines splicing efficiency.

Predicting Spliced Peptides

Table 2 shows the Matthews correlation coefficient (MCC) for the performance of each classifier type on the test sets. We see a remarkable consistency in the relative performance of the classifier types. The random forest approach produces the best result for all eight of the test sets, with the neural network classifier performing notably worse than both SVMs and random forests in all cases. The gap in performance between the neural network and the other approaches diminishes as the size of the datasets increase, suggesting that with a sufficiently large dataset it might be a good choice of classifier. From this point on, if the type of classifier is not mentioned, it can be assumed to be a random forest, as this was the method used for any additional tests not included in table 2.

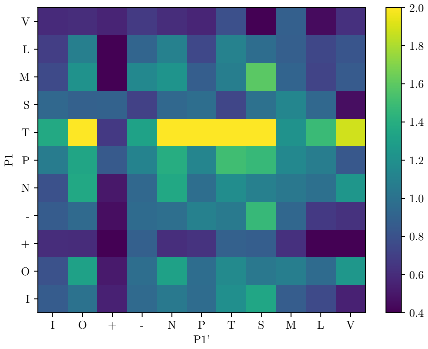


Figure 12: Relative frequencies in the C1R cell line of pairs of residues either side of the splicing site compared to pairs of non-anchor in nonspliced peptides

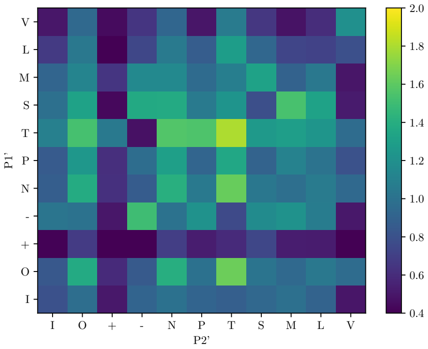


Figure 13: Relative frequencies in the C1R cell line of pairs of residues on the C-terminal peptide side of the splicing site compared to pairs of non-anchor in nonspliced peptides

Dataset	N	RF	NN	SVM
A0101	14	1.000	0.577	1.000
A0301	48	0.959	0.746	0.959
B0702	77	0.873	0.769	0.845
B2705	30	0.935	0.607	0.935
C0702	276	0.725	0.667	0.718
GRLCL	728	0.582	0.537	0.576
C1R	150	0.637	0.543	0.610
fibroblasts	166	0.575	0.471	0.527

Table 2: Matthews correlation coefficients for the best performing classifiers of each type - random forest, RF, neural network, NN, and support vector machine, SVM - evaluated on the test sets

Summary of Statistical Measures

Sensitivity, specificity and Matthews correlation coefficent are used repeatedly in this project to evaluate classifier performance. Below are their definitions in terms of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN).

$$\textbf{Sensitivity} = \frac{TP}{TP + FN}$$

$$\textbf{Specificity} = \frac{TN}{TN + FP}$$

$$\textbf{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Predicting Spliced Peptides in Cell Lines with Multiple HLA Alleles

Ideally, we ultimately want to be able to predict spliced peptides for any cell line with any HLA alleles. In training and testing a classifier on the spliced peptides from the GRLCL and fibroblast cell lines we hoped to show that it is indeed possible to predict spliced peptides.

We find the classifiers perform reasonably well and, in both cases, the level of performance is similar with a Matthews correlation coefficient just below 0.6.

Dataset	N	Sens.	Spec.	MCC
GRLCL	728	0.801	0.783	0.582
Fibroblasts	166	0.753	0.820	0.575

Table 3: Performance of the random forest classifier on GRLCL and fibroblast cell lines measure by sensitivity, specificity and Matthews correlation coefficient

Predicting Spliced Peptides in an Unrelated Cell Line

After determing that spliced peptides within a particular cell line can indeed be predicted with a reasonable degree of accuracy, I investigated whether a classifier trained on one cell line can be used to classify peptides in an unrelated cell line with distinct HLA alleles. After training and testing (figure **??**) the classifier on the spliced peptides from the GRLCL and fibroblast cell lines, we used the set of spliced peptides from the C1R cell line as the test set from an unrelated cell line. The classifier was only able to achieve 41.7% accuracy on the C1R peptides, which - while greater than the false negative rate of 19.8% (i.e. the proportion of positive predictions we would expect on a random sample of non-presented peptides) - is much worse the performance of a classifier trained and tested on the same cell line.

Predicting Spliced Peptides in a Cell Line with a Single HLA Allele

The C1R cell line has only one type of MHC-I molecule, HLA-B*40:02, for peptides to bind to. This makes it a natural choice for building a classifier to identify peptides for a particular MHC I molecule. I incorporated SMM binding predictions for B*40:02 as a feature in the classifier, a comparison of the results of these two methods can be seen in table 4.

Perhaps surprisingly, neither approach works especially well with them both scoring similarly to classifiers trained on the GRLCL and fibroblast cell lines.

More surprisingly, including the IC50 binding predictions made the classifier less effective. This may be due to the IC50 score capturing a lot of the information that the classifier would otherwise

IC50	N	Sensitivity	Specifiticity	MCC
Yes	150	0.792	0.769	0.560
No	150	0.809	0.829	0.637

Table 4: Comparison of C1R random forest classifier performance with and without the SMM IC50 binding predictions for HLA-B*40:02

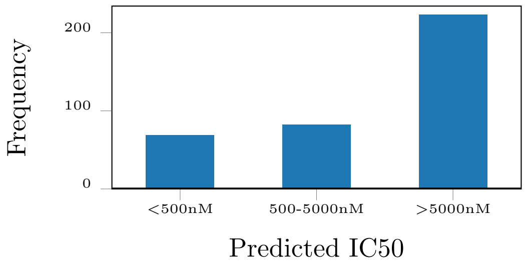


Figure 14: SMM binding predictions for spliced peptides in the C1R cell line

have to learn from patterns of residues in particular positions, leading to the classifier to overfit when trained on the same number of examples. It could also be simply due to the SMM method performing poorly for this particular allele. Figure 14 shows the distribution of binding predictions - we see that less than 20% of these spliced peptides are predicted to bind. This perhaps makes the most sense as an explanation for the poor result - incorporating an additional feature with little predictive value will inevitable lead to a worse result.

Dataset	N	RF	NN	SVM
A0101	14	1.000	0.577	1.000
A0301	48	0.959	0.746	0.959
B0702	77	0.873	0.769	0.845
B2705	30	0.935	0.607	0.935
C0702	276	0.725	0.667	0.718

Table 5: text

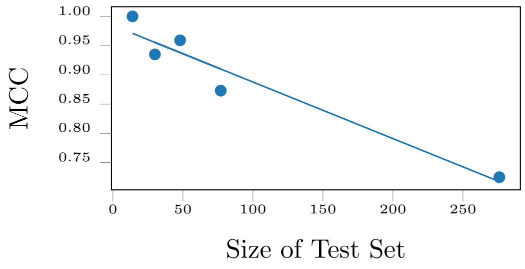


Figure 15: Relationship between size of test set and performance of classifier for the GRLCL alleles

Predicting Spliced Peptides for Specific HLA Alleles from Cell Lines with Multiple Alleles

More useful, perhaps, for immunotherapeutical and future research purposes is a tool that can predict spliced peptides for a particular allele as NetMHCpan does for nonspliced peptides. Table 5 shows the performance of the three types of classifier on the test sets for each of the five alleles from the GRLCL cell line.

We find good peformance for all five of the alleles, especially with a random forest or SVM approach, with a Matthews correlation coefficient of greater than 0.7 for all of them. Seemingly paradoxically, we find a strong relationship between the size of the test set and the performance of the classifier (figure 15). However, this can probably be explained by biases in the SMM algorithm ; if the algorithm is more selective about assigning low IC50 scores (lower is better) for a particular allele, it will result in a smaller set of predicted binders and the ones it assign will be more certain and probably have greater similarity to each other.

For both of the HLA-C*07:02 and -B*07:02 alleles (chosen because they had the largest training sets), I tested their best performing classifier on a set of randomly generated spliced peptides predicted to bind to them (IC50 ¡ 500nM) by the SMM method. This was done to test the extent to which the classifiers were acting as simply binding prediction tools (recall that the original training sets

Classifier	Sensitivity	FPR
C0702	0.862	0.267
B0702	0.972	0.896

Table 6: Comparison of the sensitivity (true positive rate) of the HLA-C*07:02 and -B*07:02 classifiers on the normal test set to the false positive rate (FPR) when tested on a set of invented spliced peptides that were predicted to bind to these alleles

for these classifiers had also been chosen by selecting spliced peptides with a binding strength of less than 500nM, making this a good test).

The relevant comparison in this test is between the sensitivity of the classifier on the original test set - i.e. the proportion of positives correctly predicted as such - and the false positive rate - i.e. the proportion of negatives incorrectly predicted as positives - in this new test set. If the classifiers are entirely predicting binding rather than anything else, we would expect these two values to be the same, as both sets consist of binding peptides according to the SMM method.

We find (table 6) excellent results for the C*07:02 classifier, just over a quarter of "invented" binding peptides are wrongly identified as spliced peptides compared with 86% of spliced peptides being accurately labelled. This strongly suggests that this classifier is doing more than just predicting binding strength.

For the B*07:02 classifier, the results are less good with a FPR of nearly 90% compared with a sensitivity of 97.2%. The test set for this allele is much smaller than for C*07:02, with only a quarter of the number of positive training examples; it may be the case that this test set is too small to effectively identify features that are typical of spliced peptides but large enough to identify the more easily distinguisable features relevant MHC-I binding (such as the residues in the anchor positions).

In table 7, we can see the relative importance of features for the C*07:02 classifier. Unsurprisingly we find the anchor positions 1,2 and 9 are the most important features as these primarily determine the strength of binding to MHC-I molecules. We find that the residues either side of the splicing site are also important - accounting for 17% of the feature weights. In addition, we see that whether the splicing is reversed or not (i.e. whether the C-terminal peptide precursor occurs before the N-terminal precursor in the original protein sequence) is by far the least important of the features included.

These results validate our findings above, that the classifier predictions are not just proxies for binding predictions, though binding will inevitably a large component of them.

Feature	Score
Position 9	0.114475
Position 1	0.086463
Position 2	0.072443
Splicing P2	0.052405
Position 4	0.041146
Position 6	0.041056
Splicing P2'	0.039159
Position 7	0.037934
Position 3	0.037084
Splicing P1	0.036972
Position 5	0.036284
Position 8	0.036176
Splicing P1'	0.035927
Cleavage 3 P2	0.034957
Cleavage 3 P1	0.033822
Length1	0.033082
Cleavage 2 P2'	0.032666
Cleavage 4 P2'	0.032569
Distance	0.032384
Cleavage 2 P1'	0.032302
Cleavage 1 P1	0.032186
Cleavage 1 P2	0.031970
Cleavage 4 P1'	0.031562
Reversed	0.004975

Table 7: Relative feature importance for each feature included in the C*07:02 random forest classifier

Dataset	N	Sensitivity	Specificity	MCC
B0702	77	0.972	0.902	0.873
C0702	276	0.862	0.862	0.725
C1R	150	0.809	0.829	0.637

Table 8: Sensitivity, specificity and Matthews correlation coefficient for the best performing random forest classifier for each dataset.

Discussion

In this project, it has been shown that spliced peptides do indeed posess many unusual qualities compared to nonspliced peptides - some of which have been explored or validated here - and, as such, current tools are insufficient to predict them. In addition, there is an absense of large amounds of data on the splicing process itself. Given these contraints at the present time, I feel the use of a classifier to identify spliced peptides is completely justified and to my knowledge, this is the first time one has been built.

This project was inevitably going to be contrained by the data available. As has been done for nonspliced peptides in the past, it would be incredibly useful to have mass spectographic data from cell lines containing only a single type of MHC I molecules. This would allow a classifier to be trained without having to attempt to filter the results by predicted binding to an allele.

There is a lot of scope in this area for future work. A natural extension of the work in this project is - similarly to Platteel et al. (2017) - to test the classifiers created here on potential spliced antigens from pathogens. Helpfully, the classifiers are able to output probabilistic predictions, which, if you set a threshold, say the 99th percentile of predictions, would narrow down the range of spliced peptides needed to be tested. It would be interesting to compare the two approached to see which is the most effective - though of course it would be restricted to the five alleles that classifiers were built for.

One of the most useful and interesting things someone could do is to carry out - following the method of Liepe et al. (2016) - a large scale mass spectrometry analysis of spliced peptides produced by the proteasome prior to TAP transport. This would allow for a much deeper and clearer analysis of the properties that determine splicing and give a much greater understranding of the underlying mechanism. From this point it would be relatiely straight forward to build a tool to predict splicing in the proteasome which could then be combined with current tools (i.e. PRED(TAP) and SMM binding predictors) to produce much better results for predicting spliced peptides and for a wider range of alleles than the classifiers in this project.

Ultimately, the immune systems remains full of mysteries, it will be fascinating to see what new approaches in peptide vaccines and in immunotherapy come from discoveries pertaining to spliced peptides over the next few years.

References

[1] J. Liepe, F. Marino, J. Sidney, A. Jeko, D. E. Bunting, A. Sette, P. M. Kloetzel, M. P. H. Stumpf, A. J. R. Heck, and M. Mishto, “A large fraction of HLA class I ligands are proteasome-generated spliced peptides,” *Science*, vol. 354, pp. 605–610, oct 2016.

[2] A. K. Nussbaum, T. P. Dick, W. Keilholz, M. Schirle, S. Stevanović, K. Dietz, W. Heinemeyer, M. Groll, D. H. Wolf, R. Huber, H. G. Rammensee, and H. Schild, “Cleavage motifs of the yeast 20S proteasome beta subunits deduced from digests of enolase 1.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, pp. 12504–9, oct 1998.

[3] A. F. Kisselev, T. N. Akopian, K. M. Woo, and A. L. Goldberg, “The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes. Implications for understanding the degradative mechanism and antigen presentation,” *Journal of Biological Chemistry*, vol. 274, pp. 3363–3371, feb 1999.

[4] E. Reits, A. Griekspoor, J. Neijssen, T. Groothuis, K. Jalink, P. Van Veelen, H. Janssen, J. Calafat, J. W. Drijfhout, and J. Neefjes, “Peptide Diffusion, Protection, and Degradation in Nuclear and Cytoplasmic Compartments before Antigen Presentation by MHC Class I,” *Immunity*, vol. 18, pp. 97–108, jan 2003.

[5] K.-i. Hanada, J. W. Yewdell, and J. C. Yang, “Immune recognition of a human renal cancer antigen through post-translational protein splicing.,” *Nature*, vol. 427, pp. 252–6, jan 2004.

[6] M. Nielsen, C. Lundegaard, O. Lund, and C. Keşmir, “The role of the proteasome in generating cytotoxic T-cell epitopes: insights

obtained from improved predictions of proteasomal cleavage," *Immunogenetics*, vol. 57, pp. 33–41, apr 2005.

[7] G. L. Zhang, N. Petrovsky, C. K. Kwoh, J. T. August, and V. Brusic, "PRED(TAP): a system for prediction of peptide binding to the human transporter associated with antigen processing.," *Immunome research*, vol. 2, p. 3, may 2006.

[8] N. Vigneron, V. Stroobant, J. Chapiro, A. Ooms, G. Degiovanni, S. Morel, P. van der Bruggen, T. Boon, and B. J. Van den Eynde, "An antigenic peptide produced by peptide splicing in the proteasome.," *Science (New York, N.Y.)*, vol. 304, no. 5670, pp. 587–90, 2004.

[9] M. Mishto, A. Goede, K. T. Taube, C. Keller, K. Janek, P. Henklein, A. Niewienda, A. Kloss, S. Gohlke, B. Dahlmann, C. Enenkel, and P. Michael Kloetzel, "Driving Forces of Proteasome-catalyzed Peptide Splicing in Yeast and Humans (SM)," *Molecular & Cellular Proteomics*, vol. 11, pp. 1008–1023, 2012.

[10] A. C. Platteel, J. Liepe, K. Textoris-Taube, C. Keller, P. Henklein, H. H. Schalkwijk, R. Cardoso, P. M. Kloetzel, M. Mishto, and A. J. Sijts, "Multi-level Strategy for Identifying Proteasome-Catalyzed Spliced Epitopes Targeted by CD8 + T Cells during Bacterial Infection," *Cell Reports*, vol. 20, pp. 1242–1253, aug 2017.

[11] E. Caron, L. Espona, D. J. Kowalewski, H. Schuster, N. Ternette, A. Alpízar, R. B. Schittenhelm, S. H. Ramarathinam, C. S. Arlehamn, C. C. Koh, L. C. Gillet, A. Rabsteyn, P. Navarro, S. Kim, H. Lam, T. Sturm, M. Marcilla, A. Sette, D. S. Campbell, E. W. Deutsch, R. L. Moritz, A. W. Purcell, H. G. Rammensee, S. Stevanovic, and R. Aebersold, "An open-source computational and data resource to analyze digital maps of immunopeptidomes," *eLife*, vol. 4, pp. 1–17, jul 2015.

[12] M. Bassani-Sternberg, S. Pletscher-Frankild, L. J. Jensen, and M. Mann, "Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation.," *Molecular & cellular proteomics : MCP*, vol. 14, no. 3, pp. 658–73, 2015.

[13] "PDF to XLS – Extract tables from PDF to XLS."

[14] B. Peters and A. Sette, "Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method.," *BMC bioinformatics*, vol. 6, p. 132, may 2005.

[15] M. Andreatta and M. Nielsen, "Gapped sequence alignment using artificial neural networks: Application to the MHC class i system," *Bioinformatics*, vol. 32, pp. 511–517, feb 2015.

[16] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *Journal of Molecular Biology*, vol. 202, pp. 865–884, aug 1988.

[17] C. Keşmir, A. K. Nussbaum, H. Schild, V. Detours, and S. Brunak, "Prediction of proteasome cleavage motifs by neural networks.," *Protein engineering*, vol. 15, pp. 287–96, apr 2002.

[18] M. Mishto, A. Goede, K. T. Taube, C. Keller, K. Janek, P. Henklein, A. Niewienda, A. Kloss, S. Gohlke, B. Dahlmann, C. Enenkel, and P. M. Kloetzel, "Driving forces of proteasome-catalyzed peptide splicing in yeast and humans.," *Molecular & cellular proteomics : MCP*, vol. 11, pp. 1008–23, oct 2012.

[19] C. R. Berkers, A. de Jong, K. G. Schuurman, C. Linnemann, J. A. J. Geenevasen, T. N. M. Schumacher, B. Rodenko, and H. Ovaa, "Peptide Splicing in the Proteasome Creates a Novel Type of Antigen with an Isopeptide Linkage.," *Journal of immunology*, vol. 195, no. 9, pp. 4075–84, 2015.

Appendices

A Programming Appendix

`www.github.com/michaelandrewellis/`

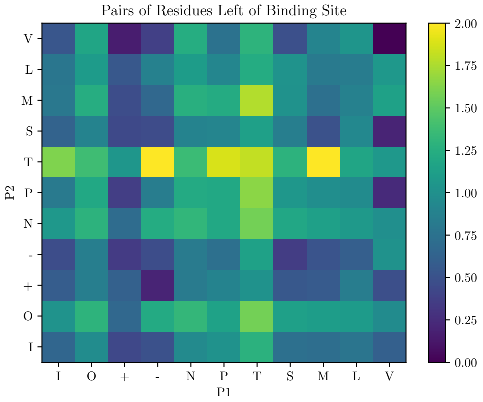


Figure 16: Binding paris left of the splice site for C1R

`MScProject`

B Table Appendix

C Graph Appendix

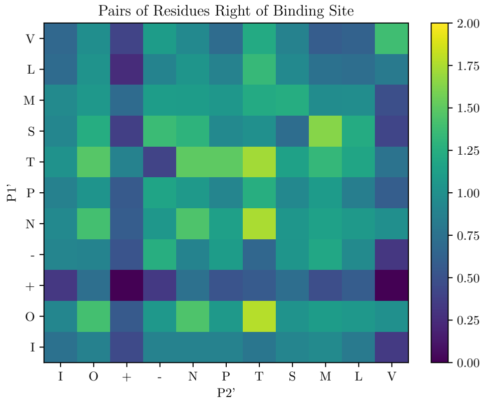


Figure 17: Binding paris right of the splice site for C1R

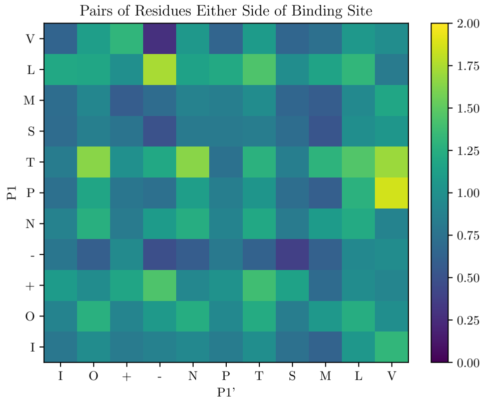


Figure 18: Relative frequency of binding pairs either side of the splicing site for the fibroblast cell line compared to non-anchor residues in nonspliced peptides

Max Features	Mean Test Score	Std Test Score
5.0	0.857143	0.048550
10.0	0.964286	0.025620
20.0	0.964286	0.025620
50.0	0.946429	0.045146
100.0	0.964286	0.051892
200.0	0.964286	0.051892
NaN	0.964286	0.051892

Table 9: Hyper-parameter tuning of random forest classifier on the A0101 dataset

C	Kernel	Mean Test Score	Std Test Score
0.1	linear	0.892857	0.046349
0.1	poly	0.517857	0.012290
0.1	rbf	0.517857	0.012290
1.0	linear	0.892857	0.046349
1.0	poly	0.517857	0.012290
1.0	rbf	0.517857	0.012290
10.0	linear	0.892857	0.046349
10.0	poly	0.517857	0.012290
10.0	rbf	0.892857	0.046349

Table 10: Hyper-parameter tuning of SVM classifier on the A0101 dataset

Activation	Hidden Layer Sizes	Mean Test Score	Std Test Score
identity	5	0.678571	0.084665
identity	10	0.660714	0.065162
identity	20	0.732143	0.043890
identity	50	0.660714	0.075552
logistic	5	0.642857	0.068605
logistic	10	0.750000	0.088794
logistic	20	0.714286	0.054349
logistic	50	0.714286	0.032774
tanh	5	0.732143	0.038163
tanh	10	0.750000	0.088794
tanh	20	0.732143	0.006828
tanh	50	0.696429	0.058720
relu	5	0.696429	0.046849
relu	10	0.678571	0.068278
relu	20	0.696429	0.060039
relu	50	0.750000	0.104962

Table 11: Hyper-parameter tuning of neural network classifier on the A0101 dataset

Max Features	Mean Test Score	Std Test Score
5.0	0.859375	0.025516
10.0	0.890625	0.022097
20.0	0.901042	0.019488
50.0	0.880208	0.019488
100.0	0.859375	0.038273
200.0	0.859375	0.038273
NaN	0.838542	0.026557

Table 12: Hyper-parameter tuning of random forest classifier on the A0301 dataset

C	Kernel	Mean Test Score	Std Test Score
0.1	linear	0.906250	0.022097
0.1	poly	0.515625	0.000000
0.1	rbf	0.515625	0.000000
1.0	linear	0.906250	0.012758
1.0	poly	0.515625	0.000000
1.0	rbf	0.515625	0.000000
10.0	linear	0.906250	0.012758
10.0	poly	0.515625	0.000000
10.0	rbf	0.916667	0.007366

Table 13: Hyper-parameter tuning of SVM classifier on the A0301 dataset

Activation	Hidden Layer Sizes	Mean Test Score	Std Test Score
identity	5	0.661458	0.019488
identity	10	0.656250	0.025516
identity	20	0.703125	0.038273
identity	50	0.671875	0.045999
logistic	5	0.703125	0.071032
logistic	10	0.661458	0.032106
logistic	20	0.666667	0.038976
logistic	50	0.671875	0.033754
tanh	5	0.692708	0.064213
tanh	10	0.666667	0.041010
tanh	20	0.713542	0.064213
tanh	50	0.671875	0.045999
relu	5	0.744792	0.026557
relu	10	0.744792	0.019488
relu	20	0.755208	0.026557
relu	50	0.729167	0.026557

Table 14: Hyper-parameter tuning of neural network classifier on the A0301 dataset

Max Features	Mean Test Score	Std Test Score
5.0	0.941368	0.000270
10.0	0.941368	0.007740
20.0	0.941368	0.007740
50.0	0.947883	0.004359
100.0	0.947883	0.004734
200.0	0.951140	0.008168
NaN	0.934853	0.012456

Table 15: Hyper-parameter tuning of random forest classifier on the B0702 dataset

C	Kernel	Mean Test Score	Std Test Score
0.1	linear	0.934853	0.016295
0.1	poly	0.508143	0.002337
0.1	rbf	0.508143	0.002337
1.0	linear	0.915309	0.016208
1.0	poly	0.508143	0.002337
1.0	rbf	0.612378	0.048808
10.0	linear	0.915309	0.016208
10.0	poly	0.508143	0.002337
10.0	rbf	0.944625	0.011994

Table 16: Hyper-parameter tuning of SVM classifier on the B0702 dataset

Activation	Hidden Layer Sizes	Mean Test Score	Std Test Score
identity	5	0.876221	0.018966
identity	10	0.856678	0.005258
identity	20	0.879479	0.026009
identity	50	0.879479	0.017124
logistic	5	0.895765	0.016946
logistic	10	0.895765	0.009469
logistic	20	0.879479	0.017124
logistic	50	0.895765	0.009469
tanh	5	0.869707	0.026037
tanh	10	0.902280	0.008364
tanh	20	0.882736	0.014337
tanh	50	0.892508	0.014292
relu	5	0.885993	0.023318
relu	10	0.872964	0.014382
relu	20	0.905537	0.011861
relu	50	0.882736	0.016414

Table 17: Hyper-parameter tuning of neural network classifier on the B0702 dataset

Max Features	Mean Test Score	Std Test Score
5.0	0.941667	0.022697
10.0	0.966667	0.031099
20.0	0.975000	0.020288
50.0	0.983333	0.023570
100.0	0.991667	0.011785
200.0	0.983333	0.023570
NaN	0.975000	0.020288

Table 18: Hyper-parameter tuning of random forest classifier on the B2705 dataset

C	Kernel	Mean Test Score	Std Test Score
0.1	linear	0.966667	0.031099
0.1	poly	0.508333	0.005898
0.1	rbf	0.508333	0.005898
1.0	linear	0.966667	0.031099
1.0	poly	0.508333	0.005898
1.0	rbf	0.508333	0.005898
10.0	linear	0.966667	0.031099
10.0	poly	0.508333	0.005898
10.0	rbf	0.991667	0.011568

Table 19: Hyper-parameter tuning of SVM classifier on the B2705 dataset

Activation	Hidden Layer Sizes	Mean Test Score	Std Test Score
identity	5	0.775000	0.058139
identity	10	0.808333	0.012473
identity	20	0.791667	0.065183
identity	50	0.825000	0.022260
logistic	5	0.808333	0.023840
logistic	10	0.816667	0.032420
logistic	20	0.825000	0.038273
logistic	50	0.800000	0.022574
tanh	5	0.850000	0.055970
tanh	10	0.825000	0.042443
tanh	20	0.791667	0.032636
tanh	50	0.816667	0.012315
relu	5	0.825000	0.038273
relu	10	0.833333	0.053097
relu	20	0.841667	0.034454
relu	50	0.841667	0.052439

Table 20: Hyper-parameter tuning of neural network classifier on the B2705 dataset

Max Features	Mean Test Score	Std Test Score
5.0	0.832428	0.015584
10.0	0.836051	0.021320
20.0	0.831522	0.021852
50.0	0.831522	0.011094
100.0	0.826087	0.008000
200.0	0.820652	0.006656
NaN	0.807065	0.013496

Table 21: Hyper-parameter tuning of random forest classifier on the C0702 dataset

C	Kernel	Mean Test Score	Std Test Score
0.1	linear	0.823370	0.008000
0.1	poly	0.612319	0.024339
0.1	rbf	0.759964	0.008967
1.0	linear	0.786232	0.008967
1.0	poly	0.612319	0.024339
1.0	rbf	0.782609	0.015531
10.0	linear	0.783514	0.010485
10.0	poly	0.612319	0.024339
10.0	rbf	0.839674	0.018161

Table 22: Hyper-parameter tuning of SVM classifier on the C0702 dataset

Activation	Hidden Layer Sizes	Mean Test Score	Std Test Score
identity	5	0.787138	0.018071
identity	10	0.790761	0.016751
identity	20	0.779891	0.013496
identity	50	0.789855	0.011167
logistic	5	0.788949	0.016946
logistic	10	0.795290	0.012810
logistic	20	0.788043	0.014549
logistic	50	0.788043	0.011740
tanh	5	0.783514	0.013557
tanh	10	0.786232	0.013557
tanh	20	0.787138	0.017796
tanh	50	0.783514	0.011167
relu	5	0.795290	0.010485
relu	10	0.804348	0.012353
relu	20	0.800725	0.013374
relu	50	0.796196	0.017329

Table 23: Hyper-parameter tuning of neural network classifier on the C0702 dataset

Max Features	Mean Test Score	Std Test Score
5.0	0.792784	0.004289
10.0	0.795876	0.008698
20.0	0.798969	0.002871
50.0	0.794845	0.002921
100.0	0.796220	0.006544
200.0	0.781443	0.006125
NaN	0.769416	0.007942

Table 24: Hyper-parameter tuning of random forest classifier on the GRLCL dataset

C	Kernel	Mean Test Score	Std Test Score
0.1	linear	0.783505	0.002055
0.1	poly	0.507904	0.000243
0.1	rbf	0.507904	0.000243
1.0	linear	0.756014	0.001488
1.0	poly	0.507904	0.000243
1.0	rbf	0.783505	0.003978
10.0	linear	0.750172	0.002328
10.0	poly	0.507904	0.000243
10.0	rbf	0.783505	0.002859

Table 25: Hyper-parameter tuning of SVM classifier on the GRLCL dataset

Activation	Hidden Layer Sizes	Mean Test Score	Std Test Score
identity	5	0.768729	0.000332
identity	10	0.765979	0.002433
identity	20	0.767354	0.001565
identity	50	0.766323	0.000331
logistic	5	0.757045	0.013359
logistic	10	0.763574	0.005518
logistic	20	0.764261	0.006424
logistic	50	0.761856	0.005288
tanh	5	0.765292	0.008583
tanh	10	0.761512	0.010683
tanh	20	0.763918	0.006835
tanh	50	0.767010	0.002199
relu	5	0.746735	0.017055
relu	10	0.760137	0.001152
relu	20	0.766323	0.002008
relu	50	0.771134	0.001788

Table 26: Hyper-parameter tuning of neural network classifier on the GRLCL dataset

Max Features	Mean Test Score	Std Test Score
5.0	0.727425	0.022839
10.0	0.744147	0.024293
20.0	0.745819	0.011940
50.0	0.762542	0.009790
100.0	0.774247	0.010306
200.0	0.764214	0.013926
NaN	0.750836	0.023820

Table 27: Hyper-parameter tuning of random forest classifier on the C1R dataset

C	Kernel	Mean Test Score	Std Test Score
0.1	linear	0.687291	0.048096
0.1	poly	0.511706	0.001209
0.1	rbf	0.511706	0.001209
1.0	linear	0.650502	0.026468
1.0	poly	0.511706	0.001209
1.0	rbf	0.513378	0.003152
10.0	linear	0.650502	0.026468
10.0	poly	0.511706	0.001209
10.0	rbf	0.735786	0.028985

Table 28: Hyper-parameter tuning of SVM classifier on the C1R dataset

Activation	Hidden Layer Sizes	Mean Test Score	Std Test Score
identity	5	0.657191	0.030015
identity	10	0.648829	0.044811
identity	20	0.663880	0.032750
identity	50	0.647157	0.038181
logistic	5	0.657191	0.032177
logistic	10	0.665552	0.032485
logistic	20	0.658863	0.036154
logistic	50	0.665552	0.040592
tanh	5	0.652174	0.046379
tanh	10	0.660535	0.024425
tanh	20	0.670569	0.040316
tanh	50	0.657191	0.032177
relu	5	0.672241	0.040629
relu	10	0.665552	0.025533
relu	20	0.663880	0.040800
relu	50	0.662207	0.038213

Table 29: Hyper-parameter tuning of neural network classifier on the C1R dataset

Max Features	Mean Test Score	Std Test Score
5.0	0.763636	0.019285
10.0	0.772727	0.024337
20.0	0.781818	0.019285
50.0	0.778788	0.014999
100.0	0.772727	0.014845
200.0	0.772727	0.013381
NaN	0.769697	0.026068

Table 30: Hyper-parameter tuning of random forest classifier on the fibroblasts dataset

C	Kernel	Mean Test Score	Std Test Score
0.1	linear	0.734848	0.024148
0.1	poly	0.509091	0.000000
0.1	rbf	0.509091	0.000000
1.0	linear	0.727273	0.011134
1.0	poly	0.509091	0.000000
1.0	rbf	0.530303	0.013034
10.0	linear	0.727273	0.011134
10.0	poly	0.509091	0.000000
10.0	rbf	0.746970	0.034484

Table 31: Hyper-parameter tuning of SVM classifier on the fibroblasts dataset

Activation	Hidden Layer Sizes	Mean Test Score	Std Test Score
identity	5	0.724242	0.008571
identity	10	0.734848	0.017539
identity	20	0.736364	0.023177
identity	50	0.734848	0.010714
logistic	5	0.740909	0.019639
logistic	10	0.740909	0.017008
logistic	20	0.736364	0.009819
logistic	50	0.736364	0.019639
tanh	5	0.733333	0.023861
tanh	10	0.725758	0.010714
tanh	20	0.727273	0.009819
tanh	50	0.733333	0.013034
relu	5	0.719697	0.005669
relu	10	0.742424	0.017539
relu	20	0.740909	0.009819
relu	50	0.740909	0.012856

Table 32: Hyper-parameter tuning of neural network classifier on the fibroblasts dataset

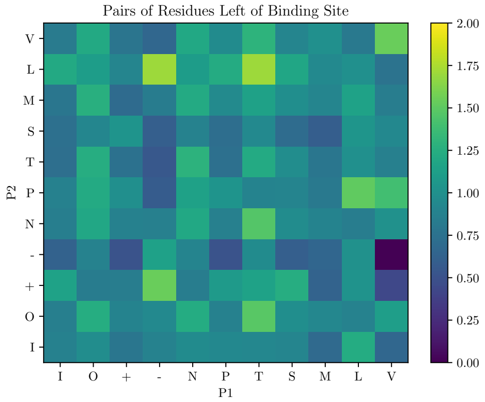


Figure 19: Relative frequency of N-terminal binding pairs for the fibroblast cell line compared to non-anchor residues in nonspliced peptides

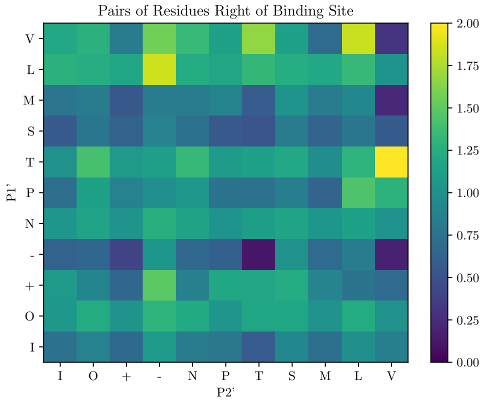
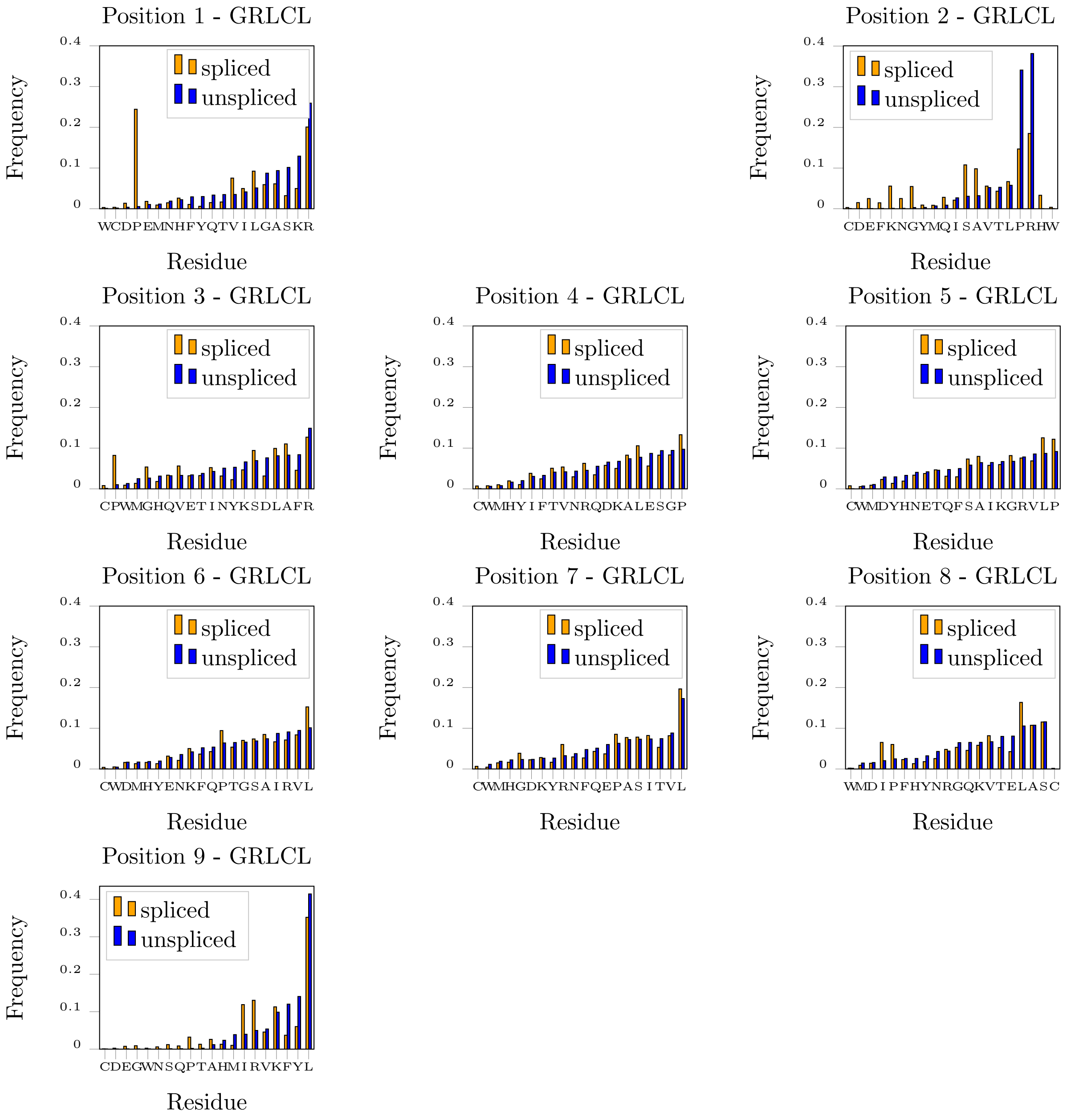


Figure 20: Relative frequency of C-terminal binding pairs for the fibroblast cell line compared to non-anchor residues in nonspliced peptides



Position 1 - GRLCL

Position 2 - GRLCL

Position 3 - GRLCL

Position 4 - GRLCL

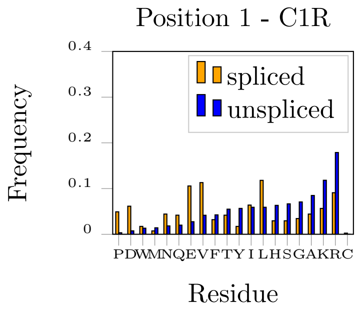
Position 5 - GRLCL

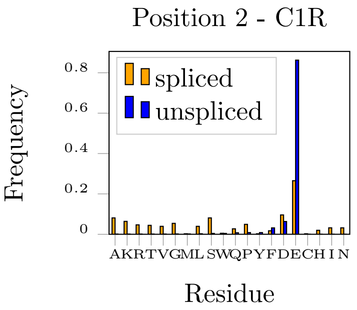
Position 6 - GRLCL

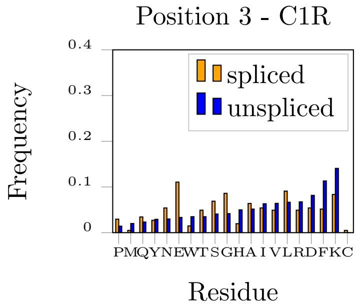
Position 7 - GRLCL

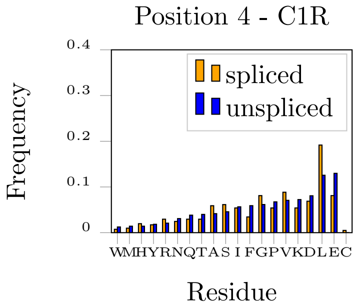
Position 8 - GRLCL

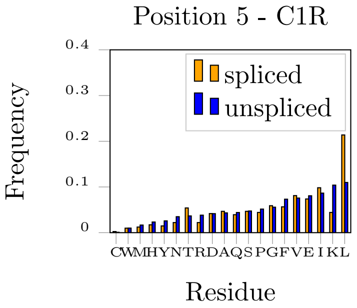
Position 9 - GRLCL

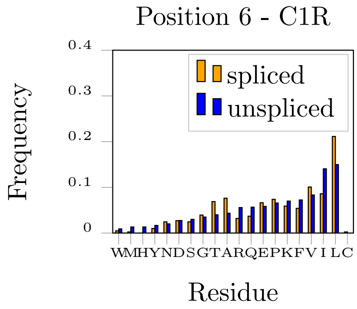

Position 1 - C1R

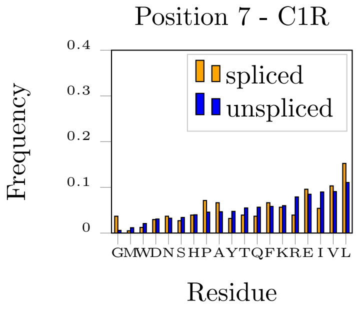

Position 2 - C1R

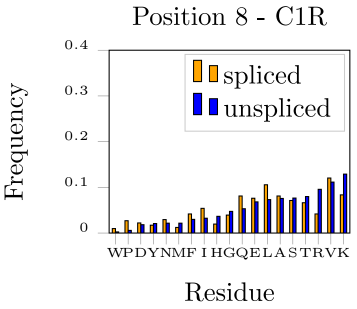

Position 3 - C1R

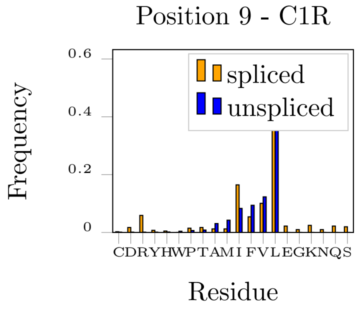

Position 4 - C1R

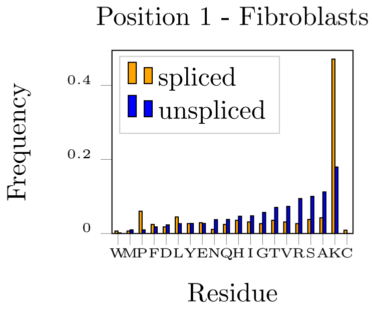

Position 5 - C1R

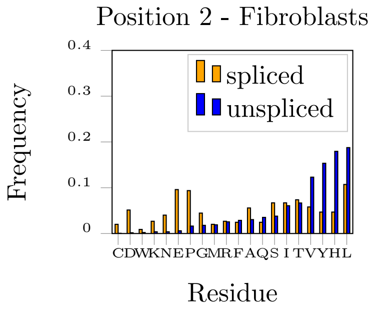

Position 6 - C1R

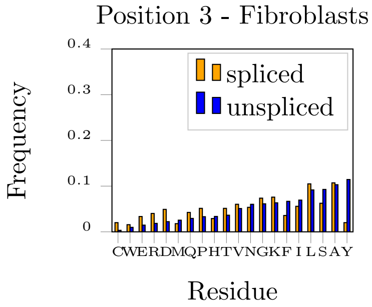

Position 7 - C1R

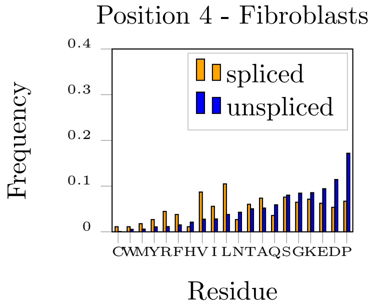

Position 8 - C1R

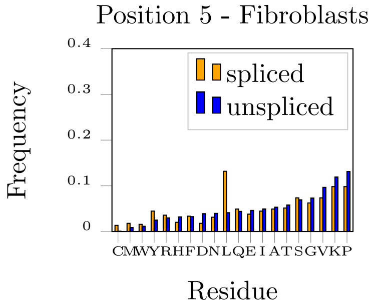

Position 9 - C1R

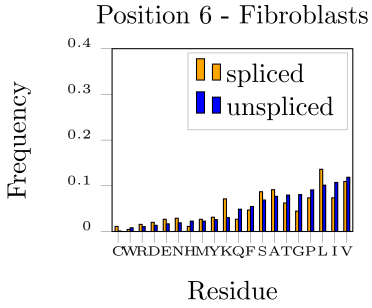

Position 1 - Fibroblasts
spliced
unspliced
Frequency
0.4
0.2
0
WMPFDLYENQHIGTVRSAKC
Residue

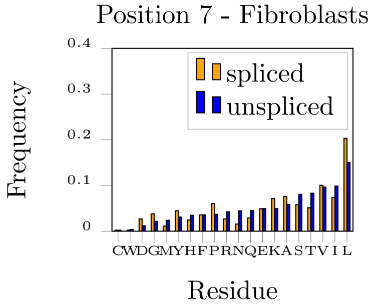

Position 2 - Fibroblasts
spliced
unspliced
Frequency
0.4
0.3
0.2
0.1
0
CDWKNEPGMRFAQSITVYHL
Residue

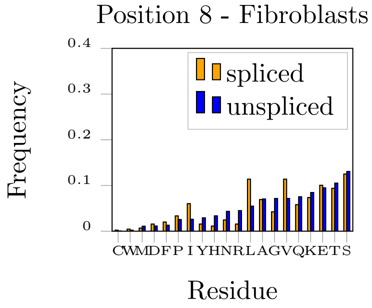

Position 3 - Fibroblasts
spliced
unspliced
Frequency
0.4
0.3
0.2
0.1
0
CWERDMQPHTVNGKFILSAY
Residue

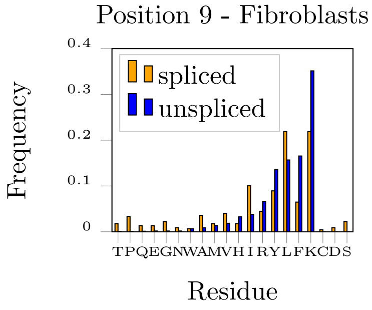

Position 4 - Fibroblasts
spliced
unspliced
Frequency
0.4
0.3
0.2
0.1
0
CWMYRFHVILNTAQSGKEDP
Residue


Position 5 - Fibroblasts
spliced
unspliced
Frequency
0.4
0.3
0.2
0.1
0
CMWYRHFDNLQEIATSGVKP
Residue


Position 6 - Fibroblasts
spliced
unspliced
Frequency
0.4
0.3
0.2
0.1
0
CWRDENHMYKQFSATGPLIV
Residue


Position 7 - Fibroblasts
spliced
unspliced
Frequency
0.4
0.3
0.2
0.1
0
CWDGMYHFPRNQEKASTVIL
Residue


Position 8 - Fibroblasts
spliced
unspliced
Frequency
0.4
0.3
0.2
0.1
0
CWMDFPIYHNRLAGVQKETS
Residue


Position 9 - Fibroblasts
spliced
unspliced
Frequency
0.4
0.3
0.2
0.1
0
TPQEGNWAMVHIRYLFKCDS
Residue