Peptide Vaccine Prediction Using Multiple Biological Parameters

Jeffrey Bagdis
Group 5 - Konrad Karczewshi, Jasdave S Chahal, Xiaojuan Ma

January 16, 2007

1 Abstract

Most current models used to predict vaccine targets which can trigger an adaptive immune response focus only on the affinity of the antigen peptide sequence for binding the Major Histocompatability Complex (MHC). MHC is, however, only one link in the pathway required to trigger and adaptive immune response. Other factors, such as T cell receptor (TCR) and B cell receptor (BCR) binding affinity affect the peptide sequences's ability to trigger an adaptive immune reponse. We attempted to construct a model based on TCR and BCR affinities, as well as MHC binding affinity, to more accurately predict potential vaccine targets. Our model appears not to reduce the number of false negatives in its predictions. It does, however, display the potential to reduce the number of false positives in its predictions, although further research is needed to verify this claim.

2 Introduction

Major Histocompatability Complex (MHC) is a protein complex in the cell membrane of almost all cell's in the human body. Small peptide sequences from the interior of the cell are presented on the surface of the cell membrane by MHC. From there, T cells receptors (TCR), which activate an adaptive immune response, can interact with the protein fragment. Additionally, B cells collect small peptide sequences from the extra-cellular region and present them on MHC.

Only antigen sequences that bind well to B cell receptors (BCR) will be collected by circulating B cells[5]. Only sequences that bind well to MHC will be presented on the surface of the cell, and only sequences that are able to interact with TCR will be able to trigger an adaptive immune response[10]. Most current models for predicting a protein sequence's proclivity to activate an immune response focus solely on the protein binding affinity for MHC and completely disregard the TCR and BCR interactions[11].

Proclivity for MHC binding is necessary, but not sufficient, to initiate an adaptive immune response[3]. We attempt to produce a better model for predicting the proclivity of a protein to induce an immune response by considering not just MHC binding affinity, but TCR and BCR binding affinity as well.

3 Method

Our model consists of three Position Specific Scoring Matrices (PSSMs), which score a peptide sequence according to its predicted binding affinity with MHC, TCR, or BCR independently. Each PSSM is constructed from a set of sequences known to have strong interaction with the respective protein complex.

3.1 Data Collection

All data used to create this model came from the Immune Epitope Database[6]. This database contains data on peptide sequences' interaction with MHC, TCR, and BCR. We found no other database containing all three sets of data in one collection. Although there are separate databases containing data of each type individually, we decided that it would be more consistent to use data from a uniform source.

All of the data we collected for MHC and TCR was for the MHC HLA-DR locus - an MHC Class II gene. We collected several subsets - one containing data for all HLA-DR alleles, and another or only the two most prevalent HLA-DR alleles in the database.

We fetched three sets of data from the Immune Epitope Database, each consisting of a list of amino acid sequences and their respective interactions with one of the protein complexes (MHC, TCR, or BCR). For MHC, the database contains of qualitative result (*positive* or *negative*) for a sequence binding with MHC, and (in some cases) a quantitative result (the *concentration* of that particular sequence required to elicit a bonding response). For TCR and BCR binding, the database only contains a qualitative result. Additionally, even for MHC, the database contains many sequences which have either no quantitave data, or quantitative data with no units specified. Because of this, we decided not to use any quantitative weighting of the input data at this stage in our investigation. We do believe, however, that quantitative weighting of the input data would help improve the model. This is an avenue of research which we intend to pursue in the future.

3.2 Data Preprocessing

The raw data was processed in several steps before it was used to generate a PSSM. These steps apply equally to the data for MHC, TCR, and BCR interaction, except where noted.

3.2.1 Duplicate Removal

The MHC dataset contains scores for a single sequence's interaction with many different alleles of MHC. To prevent a preponderance of duplicates from skewing the resulting matrix, we removed all duplicate sequences from the dataset, keeping only the best result. [1]

The TCR dataset also contains many duplicates. Because TCR binds to a peptide sequence while it is being held within the MHC complex, the binding behavior of TCR and a given peptide sequence might be slightly dependent upon which allele of MHC is holding that peptide[10]. We removed duplicates from the TCR dataset in the same manner and for the same reasons as described for the MHC dataset.

The BCR dataset does not have duplicates resulting from multiple alleles. However, it is still possible that the data contain some duplicate entries resulting from different experimenters studying the same sequence. Therefore, we removed duplicates from the BCR dataset in the same manner as described for the MHC and TCR datasets.

[1] Even if a sequence does not bind well with many common alleles of MHC, as long as it binds well to *some* allele of MHC then it could be a useful vaccine target in some percentage of the population.

3.2.2 Purging

Once the duplicates were removed from the datasets, each dataset was split into a *positive* and *negative* component.[2] Although these datasets no longer contain duplicate sequences, they do contain groups of closely related sequences. To prevent a preponderence of similar sequences from skewing the matrix, we purged our data of sequences that were too closely related to another sequence in the dataset.

To accomplish the purge operation, we used a component program of the Aligned Segment Statistical Evaluation Tool (ASSET)[4]. This purge program compares sequences using the BLOSUM62 matrix, and removes sequences from the dataset which have a relatedness score above a certain cutoff. The makers of the ASSET package recommend a cutoff score in the range of 100 - 200. However, they provided no justification for that assertion. Another group decided to use a cutoff of 30, but again provided little justification for that decision[7].

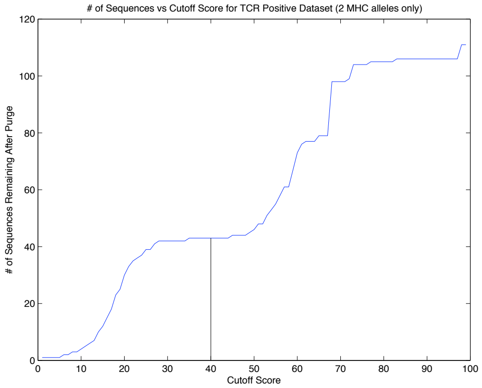


Figure 1: Number of Sequences Remaining after Purge versus Cutoff Score for TCR (2 allele) positive dataset.

To determine an appropriate cutoff score for the purge program, we analyzed the number of sequences remaining after purge with varying cutoffs. The graphs in Figure 1 compare the number of sequences remaining after

[2] We only use the *positive* portion of the dataset to generate PSSMs. The negative dataset was only used in the validation phase of the PSSMs, to provide a pool of known negative sequences to test against.

purge to the cutoff score.[3] In each case except the BCR data, the resulting curves appear to have a plateau region roughly centered around a cutoff score of 40. Our interpretation of these curves is that for cutoff values above the plateau region, somewhat highly related sequences are not being fully purged, and that for cutoff values below the plateau, relatively unrelated sequences are being overly aggressively purged. Therefore, we decided to use a cutoff score of 40 for purging our datasets. Graphs of the purge curves for each dataset are included in Appendix A

The cutoff curve for the BCR dataset did not display the same plateau region found in the curves for all the other datasets. One possible explanation for this phenomenon is that the BCR dataset is larger than the other datasets, and the plateau region may exist but be too small to observe on the graph. Additionally, although we could not detect a plateau region in the BCR graph, the chosen cutoff score of 40 still falls roughly in the vicinity of a point of inflection. Nothing else about the BCR curve suggests an alternative cutoff value, so we decided to continue using the chosen cutoff of 40 for all datasets.

3.2.3 Splitting

After overly-similar sequences were purged from the datasets, each dataset was randomly divided into ten subsets. Nine of these subsets were used for constructing the PSSMs (the “training” sets), and the tenth subset was set aside for validating the PSSMs (the “testing” set). Sequences in the testing sets were then separated according to length to facilitate the construction of the PSSMs.

3.3 PSSM Generation

Research suggests that the average length of the motif that determines MHC binding affinity is 9 residues.[8]. However, a peptide sequence must be at least several residues longer than 9 in order to fit properly in the MHC complex. The binding affinity of the sequence seems to depend predominantly on the peptides in the main 9-residue motif, but the identity of the ancillary peptides may also contribute slightly[13]. Therefore we decided to generate multiple

[3] Only the positive datasets are analyzed here, because only the positive datasets are used for generating the PSSMs. The negative datasets, therefore, do not need to be purged.

Position Specific Scoring Matrices (PSSMs), with lengths varying between 9 and 15 residues.

We used the motif discovery tool MEME[1] to generate our PSSMs. For each dataset (MHC, TCR, and BCR) we generated PSSMs of several lengths: 9, 10, 12, 14, and 15 residues long. A PSSM of length N is generated from all sequences of length $\geq N$. The MEME program was run with these command line options `meme -mod oops -w x` where x is the length of the desired PSSM (in this case $x \in \{9, 10, 12, 14, 15\}$), and `oops` stands for "One Occurence Per Sequence" - a mode which includes each input sequence exactly once in the PSSM.

3.4 Scoring

An input sequence is scored against our PSSMs according to the following procedure. A score is generated for the sequence by application of each PSSM. These raw scores are converted into percentage scores: for each PSSM, the sequence's percentage score is its raw score expressed as a percentage of the maximum theoretical score that could be returned by that PSSM. Finally, the sequence's final MHC score is computed as the maximum percentage score returned by all of the MHC PSSMs. The final TCR and final BCR scores are computed similarly as the maximum percentage scores from the TCR and BCR PSSMs, respectively.

4 Results

4.1 Logos

The logos in Figure 2 represent the 10 residue PSSMs for MHC, TCR, and BCR produced by our model. Additional logos for alternate lengths are included in Appendix B. This MHC logo seems to indicate the MHC requires several fairly specific amino acids in certain locations within the motif (3 and 8 in from the left, in particular.) The logos for TCR indicate less peculiarity, and the logo for BCR indicates a fairly promiscuous motif.

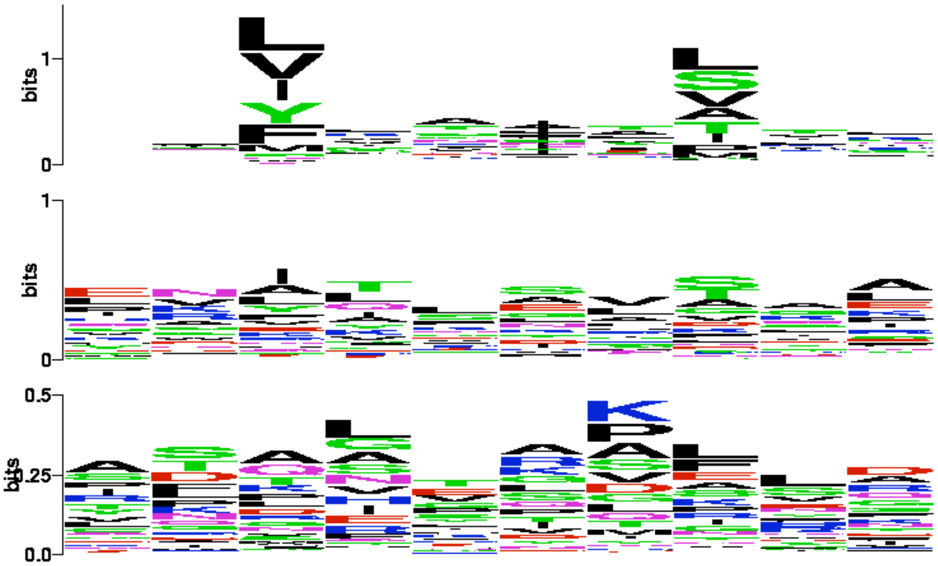


Figure 2: PSSM Logos (length 10) for MHC (all MHC alleles), TCR (2 MHC alleles), and BCR, from top to bottom.

4.2 T-Test

Our model displays a statistically significant separation between the set of known *positive* sequences and a set of random sequences. We performed a T-test on the distribution of scores returned by our model. The results of that test are in Table 1.

Dataset	**P-value**
MHC 2 alleles (positive vs. random)	$2.38 \cdot 10^{-7}$
MHC all alleles (positive vs. random)	$2.40 \cdot 10^{-8}$
TCR all alleles (positive vs. random)	$2.16 \cdot 10^{-2}$
BCR (positive vs. negative)	$4.58 \cdot 10^{-2}$

Table 1: P-values for the separation of *positive* and *negative* data by our model

4.3 ROC Curves

Curve	**Area Under Curve**
MHC 2 alleles (positive vs. random)	0.793
MHC all alleles (positive vs. random)	0.761
TCR 2 alleles (positive vs. random)	0.600
TCR all alleles (positive vs. random)	0.766
BCR (positive vs. random)	0.601
BCR (positive vs. negative)	0.782

Table 2: Area Under Curve

The ROC curve measures the ratio of false positives[4] to true positives[5]. We are using a set of random sequences to represent the negative testing set in most cases. However, for the BCR curve, we have also used the set of negative sequences (which do not bind well to BCR) retrieved from the database.

The ratio of the area under the ROC curve to the area above it (AUC) is a good measure of the specificity of a model. It is a direct representation of the ratio of true positives to false positives over a wide range of cutoff scores[6]. An AUC of 0.5 is equavalent to an algorithm which choosed by random chance. In all cases, our model performs at or above an AUC of 0.6 which is decent. In some cases, it approaches an AUC of 0.8 which is indicative of excellent specificity. Table 2 contains AUC ratios for the different versions of our model. The actual graphs are contained in Appendix C.

The set of known negative sequences is used for the BCR model but not for the other models. For BCR, the negative set if consistently identified as negative by the model. However, for MHC and TCR, sequences in the negative set are very frequently identified as positives. We believe that this abnormally high false positive rate is due to the fact that the only sequences in the database are sequences which exhibited characteristic of a good binding sequence. Some of these, upon experimentation, proved to bind poorly, and

[4] False positives in this case are sequences which we know to be *negative* (non-binding), but which our model predicts will bind strongly.

[5] True positives in this case are sequences which we know to have a binding affinity, and which our model also predicts to have such an affinity.

[6] Scores above which a sequence is considered to be *positive* by our model

were classified at *negative*. Because these sequences were initially identified as likely binding sequences, however, they probably posess at least some of the features that distinguish a good binding sequence (and a *positive* result in our model), even though they also posess some unknown characteristics that make them poor binding sequences. Therefore, we chose to use a set of random sequences as the *negative* set for the MHC and TCR models. This set is much more consistently identified as *negative* by the models.

Conversely, the BCR models are less consistent at identifying random sequences as negative. This could be because B cell receptors are less discriminating than MHC and TCR, and thus actually bind with a greater percentage of the random set. Therefore, we have used the BCR negative set, as well as the random set, as negative testing sets for the BCR model.

4.4 Correlation

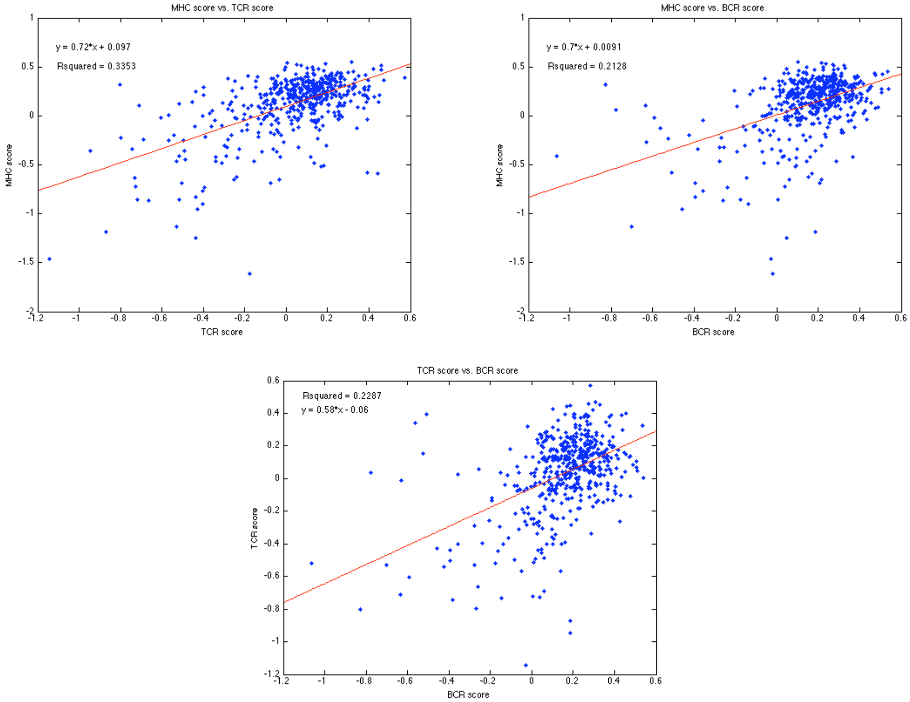


Figure 3: MHC-TCR, MHC-BCR, and TCR-BCR Correlations

Although this is somewhat tangential to our hypothesis, we decided to examine the correlation between the MHC, TCR, and BCR binding scores under our model. The plots in Figure 3 relate MHC score to TCR score, MHC score to BCR score, and TCR score to BCR score. We attempted to fit a line to these data using linear regression. The R^2 value of the regression is displayed on the charts. Their appears to be at least a small amount of correlation in the data, especially between MHC and TCR scores.

5 Discussion

The T-tests and ROC areas described in the results seciton indicate that our model is able to distinguish positive and negative sequences with a fair amount of success. However, we would also like to compare its effectiveness to other, existing models, and test its usefulness in identifying actual vaccine targets.

5.1 Comparison with RankPep

We decided to compare our model with RankPep, an existing PSSM-based MHC binding prediction algorithm. RankPep uses a separate PSSM for each individual MHC HLA-DR allele. Potential vaccine targets are identified by RankPep based solely on thei affinity for binding to MHC [8].

To compare the specificity of our model versus RankPep, we conducted a T-test of the RankPep model in the same manner as the T-test we conducted on our own model. Since RankPep does not take TCR or BCR binding affinity into account, those aspect of our model are useless in this comparison. Also, since our model considers binding affinity for multiple alleles of MHC, and was built with the assumption that a sequence that exhibited a binding affinity for *any* allele of MHC would be considered positive, the RankPep model was extended to handle multiple alleles of MHC simultaneously by simply taking the maximum score returned by any of its constituent PSSMs for a given input sequence. Table 3 compares the P-values of our model with the P-values of the RankPep model, both for our 2-allele set, and the set encompassing all alleles of HLA-DR MHC.

The P-value from the T-test is roughly the probability of the 2 datasets (*positive* and *negative*) having the same mean score. A P-value of 1 would be equivalent to a model that produced no score separation between *positive* and

Dataset	**Our P-value**	**RankPep's P-value**
MHC 2 alleles (positive vs. random)	$2.38 \cdot 10^{-7}$	$2.08 \cdot 10^{-2}$
MHC all alleles (positive vs. random)	$2.40 \cdot 10^{-8}$	$1.26 \cdot 10^{-2}$

Table 3: P-values for the separation of *positive* and *negative* data by our model and the RankPep model.

negative datasets - i.e. a model that was completely unable to differentiate the two sets. Lower P-values indicate a higher specificity of the model. The P-values in Table 3 indicate that our model is approximately 5 orders of magnitide more specific for the set of only 2 MHC alleles, and approximately 6 times more specific for the set of all HLA-DR MHC alleles.

The substantially greater specificity of our model could potentially be explained by the fact that we build one PSSM based on sequence interaction with multple alleles of MHC. This could potentially allow our model to identify broader features of the peptide chain that govern interaction with all alleles of MHC in general, and allow it to ignore peculiar features of the peptide chain that only influence interaction with a single allele of MHC.

5.2 Vaccines

To evaluate the effectiveness of our model in predicting actually useful vaccine targets, we selected several peptide sequences that have been demonstrated to be good vaccines targets. These sequences, and their scores under RankPep and our models are shown in Table 4 (for only 2 alleles of MHC) and in Table 5 (for all alleles).

The set of random sequences is shows in the table to give a baseline cutoff reading (the average score of all sequences in the random set) for each of the models. Since these models, on average, give this score to random sequences, the cutoff score must be higher that ths value. The

In this mode of analysis, our model appear to perform approximately the same as RankPep. Taking the average score of the random set as the cutoff value for the model, both models correctly identify the 2^{nd}, 4^{th}, and 5^{th} vaccines, and fail to identify the 1^{st} and 3^{rd}. Furthermore, the TCR and BCR components of our model agree with the MHC component, and do not provide any additional specificity as we had hoped. Our model does not seem

Name	**Sequence**	**RankPep Max Score**	**Our MHC Score**	**Our TCR Score**	**Our BCR Score**
MART-1 tumor[14]	AAGIG ILTV	−0.148	0.054	−0.113	−0.334
Strep Vacc.[12]	TGAQT IKGQK LYFKA NGQQV KG	0.226	0.347	0.450	0.251
Melanoma Vacc. g209-2M mod with extra Gs on ends[9]	GIMQV PFSVG	0.123	0.043	0.066	−0.213
C4-V3MN[2]	KQIIN MWQEV GKAMY ATRPN YNKRK RIHIG PGRAF YTTK	0.183	0.283	0.228	0.335
C4-V3MN[2]	KQIIN MWQEV GKAMY ATRPG NNTRK SIPIG PGRAF IATS	0.197	0.283	0.258	0.297
341 random sequences	Even distribution of Amino Acids	0.156	0.120	0.058	0.191
BCR negative sequences	-	-	-	-	0.079

Table 4:

Name	**Sequence**	**RankPep Max Score**	**Our MHC Score**	**Our TCR Score**	**Our BCR Score**
MART-1 tumor	AAGIG ILTV	0.083	−0.166	−0.207	−0.334
Strep Vacc.	TGAQT IKGQK LYFKA NGQQV KG	0.584	0.220	0.364	0.251
Melanoma Vacc. g209-2M mod with extra Gs on ends	GIMQV PFSVG	0.160	−0.127	−0.216	−0.213
C4-V3MN	KQIIN MWQEV GKAMY ATRPN YNKRK RIHIG PGRAF YTTK	0.379	0.180	0.157	0.335
C4-V3MN	KQIIN MWQEV GKAMY ATRPG NNTRK SIPIG PGRAF IATS	0.379	0.182	0.129	0.297
341 random sequences	Even distribution of Amino Acids	0.347	0.156	0.130	0.191
BCR negative sequences	-	-	-	-	0.079

Table 5:

to be able to effectivey reduce the number of false negatives.[7]

5.3 Correlation

The TCR and BCR scores may still increase the overall specificity of the model in certain circumstances. Specifically, they may help reduce the number of false positives that would be generated by a model based solely on MHC affinity. A hypothetical false positive that would be eliminated by the TCR component of our model is a peptide sequence that shows a high affinity for binding with MHC, but a low affinity for binding with TCR. A model based solely on MHC would likely identify this peptide as a good vaccine target. However, even though this peptide would indeed bind will with MHC, if it didn't also bind to a T-cell receptor, it would not trigger an adaptive immune response. We were thus far unable to find well-researched and publicized false-positives to an MHC-only model - simply because there is a dirth of published material about sequences that didn't work. However, the existence of such sequences can be demonstrated by our correlation graph (Figure 3) that compares the MHC and TCR scores of individual peptide sequences. There is an outlying point in the upper-left region of the MHC-TCR correlation graph, which corresponds to an MHC score of about 0.4 (a *positive* result) and a TCR score of about -0.8 (definitely a *negative* result). If our models are correct, then this sequence binds well to MHC, but does not activate TCR, and thus does not trigger an adaptive immune response. This sequence represents an input that would be a false positive on an MHC-only model, but would be a negative on our combined model.

5.4 Conclusion

In attempting to design a more specific predictor of the prediliction of a peptide sequence to trigger an adaptive immune response, we have attempted to design a model based on the TCR and BCR binding affinity of a peptide sequence in addition to MHC binding affinity (which is all that most mainstream models seem to use).

Our combined model was unable to decrease the number of false negatives in its predictions. However, it is likely that our combined model will reduce the number of false positives (for effective vaccine targets) in its pre-

[7] effective vaccine sequences which are not scored above the cutoff

dictions compared to MHC-only models. Preliminary evidence supports this assertion, and an avenue of future research would be to further explore the false-positive behavior of the model.

References

[1] Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006) *Nucl. Acids Res.* **34**(suppl_2), W369-373

[2] Bartlett, J. A., Wasserman, S. S., Hicks, C. B., Dodge, R. T., Weinhold, K. J., Tacket, C. O., Ketter, N., Wittek, A. E., Palker, T. J., Haynes, B. F., and Group, T. D. S. (1998) *AIDS* **12**(11), 1291-1300

[3] Heemels, M.-T., and Ploegh, H. L. (1994) *Immunity* **1**(9), 775-784

[4] A. F. Neuwald and P. Green, (1994) *JMB* **239**,698-712.

[5] Patterson, H. C. K., Kraus, M., Kim, Y.-M., Ploegh, H., and Rajewsky, K. (2006) *Immunity* **25**(1), 55-65

[6] Peters, B., Sidney, J., Bourne, P., Bui, H.-H., Buus, S., Doh, G., Fleri, W., Kronenberg, M., Kubo, R., Lund, O., Nemazee, D., Ponomarenko, J. V., Sathiamurthy, M., Schoenberger, S., Stewart, S., Surko, P., Way, S., Wilson, S., and Sette, A. (2005) *PLoS Biology* **3**(3), e91

[7] Reche, P. A., Glutting, J.-P., and Reinherz, E. L. (2002) *Human Immunology* **63**(9), 701-709

[8] Reche, P. A., Glutting, J.-P., Zhang, H., and Reinherz, E. L. (2004) *Immunogenetics* **56**(6), 405-419

[9] Rosenberg, S. A., Yang, J. C., Schwartzentruber, D. J., Hwu, P., Marincola, F. M., Topalian, S. L., Restifo, N. P., Dudley, M. E., Schwarz, S. L., Spiess, P. J., Parkhurst, M. R., Kawakami, Y., Seipp, C. A., Einhorn, J. H., and White, D. E. (1998) *Nat Med* **4**(3), 321-327

[10] Derek B. Sant'Angelo, E. R. C. A. J. J. L. K. D. (2002) *European Journal of Immunology* **32**(9), 2510-2520

[11] Sette, A. (2000) *Science* **290**(5499), 2074b-2075

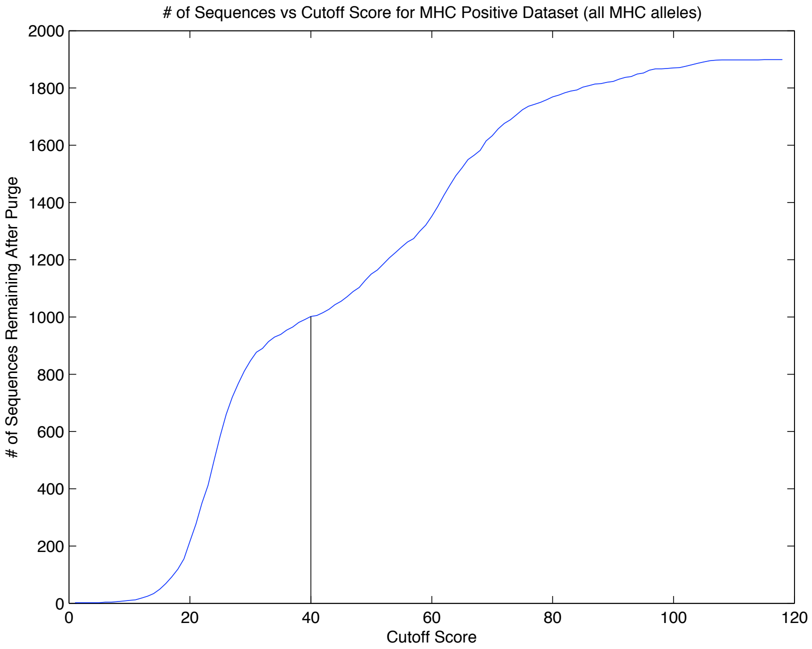
[12] Smith, D. J., Taubman, M. A., Holmberg, C. F., Eastcott, J., King, W. F., and Ali-Salaam, P. (1993) *Infect. Immun.* **61**(7), 2899-2905

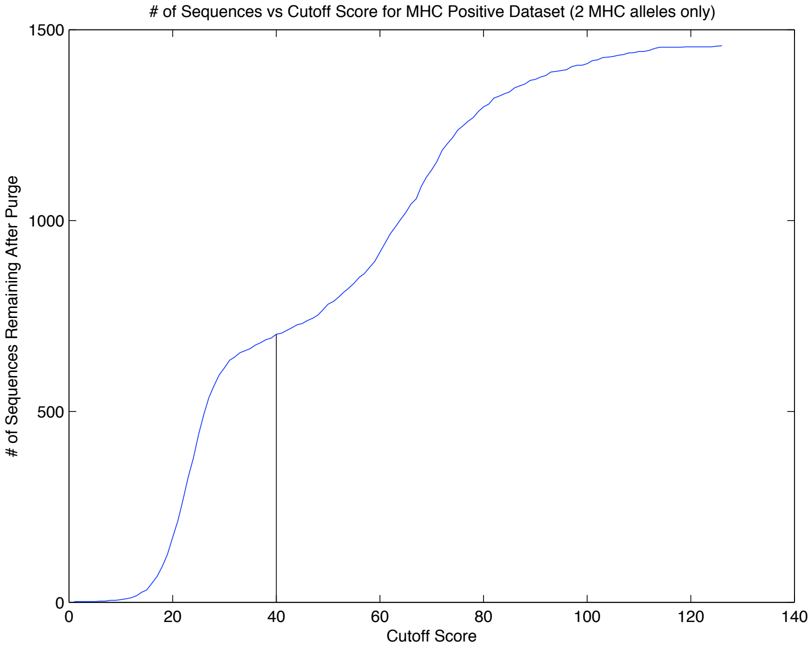
[13] Sturniolo, T., Bono, E., Ding, J., Raddrizzani, L., Tuereci, O., Sahin, U., Braxenthaler, M., Gallazzi, F., Protti, M. P., Sinigaglia, F., and Hammer, J. (1999) *Nat Biotech* **17**(6), 555-561

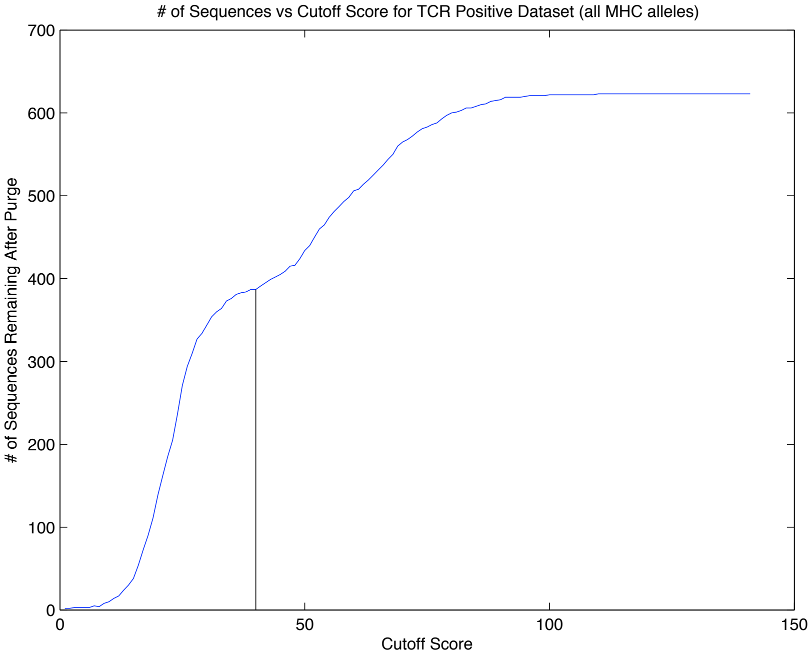
[14] Wang, F., Bade, E., Kuniyoshi, C., Spears, L., Jeffery, G., Marty, V., Groshen, S., and Weber, J. (1999) *Clin Cancer Res* **5**(10), 2756-2765

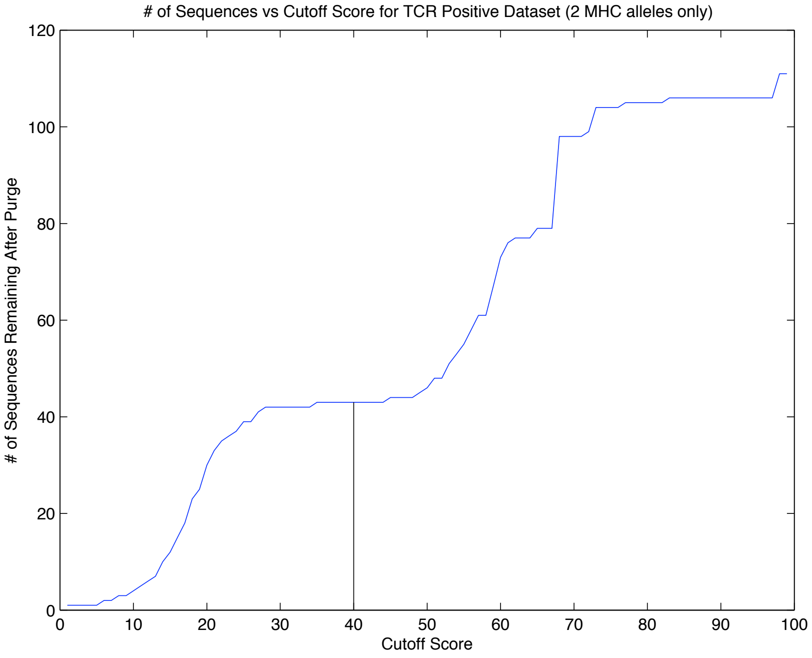
A Purge Graphs

Number of Sequences Remaining after Purge versus Cutoff Score for each positive dataset

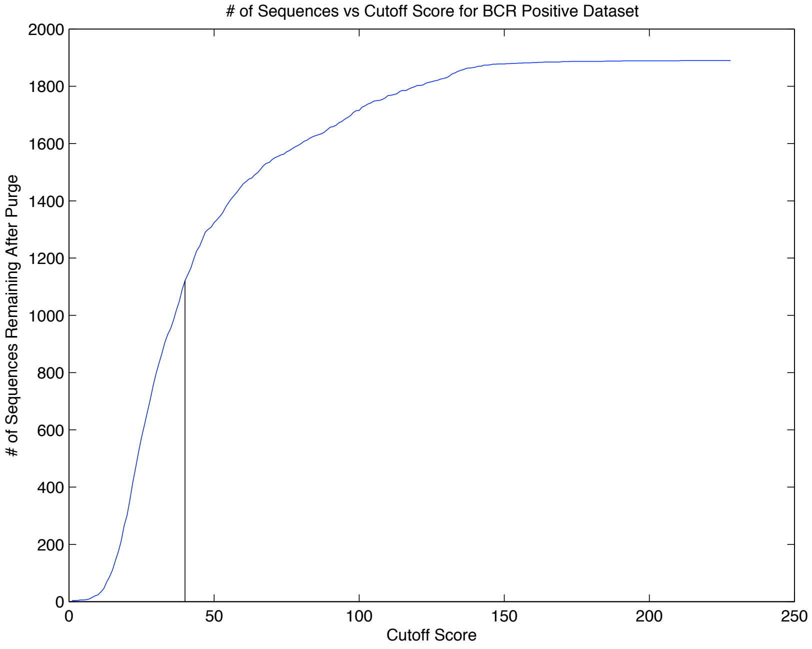



# of Sequences vs Cutoff Score for MHC Positive Dataset (2 MHC alleles only)
# of Sequences Remaining After Purge
Cutoff Score
1500
1000
500
0
0
20
40
60
80
100
120
140


# of Sequences vs Cutoff Score for TCR Positive Dataset (all MHC alleles)
# of Sequences Remaining After Purge
Cutoff Score
700
600
500
400
300
200
100
0
0
50
100
150



# of Sequences vs Cutoff Score for TCR Positive Dataset (2 MHC alleles only)
# of Sequences Remaining After Purge
Cutoff Score
0
20
40
60
80
100
120
0
10
20
30
40
50
60
70
80
90
100



# of Sequences vs Cutoff Score for BCR Positive Dataset
# of Sequences Remaining After Purge
Cutoff Score
0
200
400
600
800
1000
1200
1400
1600
1800
2000
0
50
100
150
200
250

B Logos

The logos for our PSSMs from each dataset. testing data.

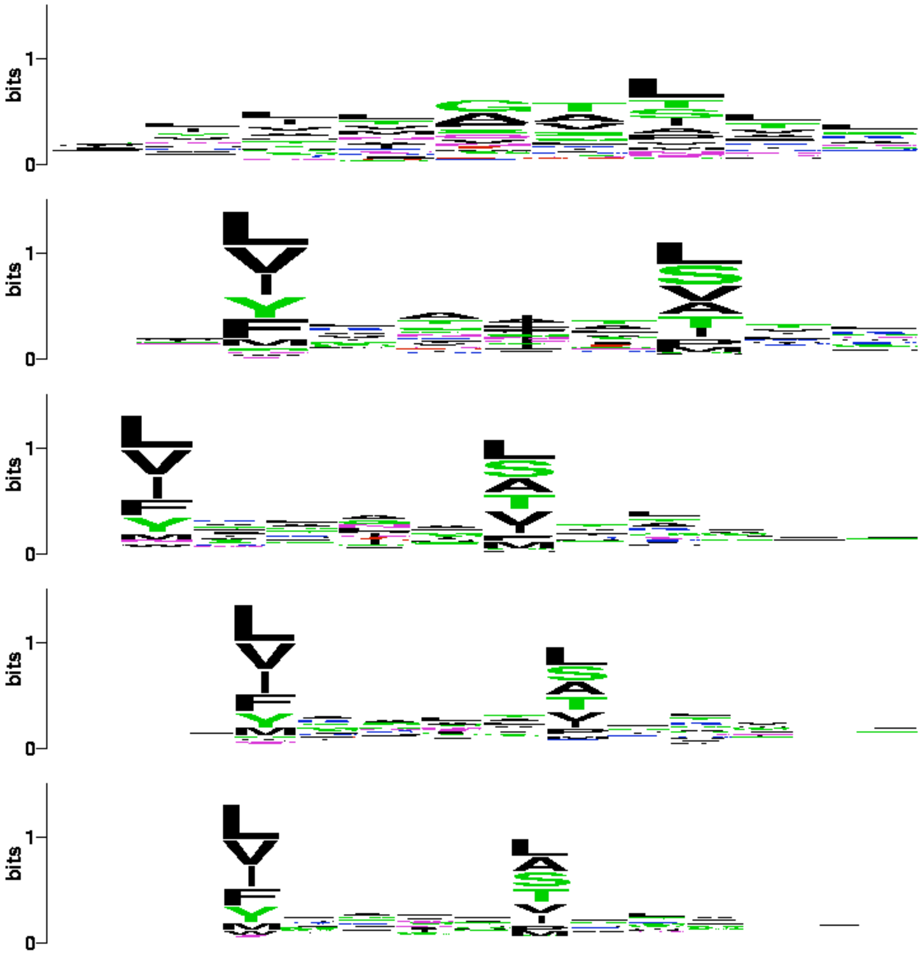


Figure 4: PSSM Logos for MHC (2 alleles), lengths (9, 10, 12, 14, and 15)

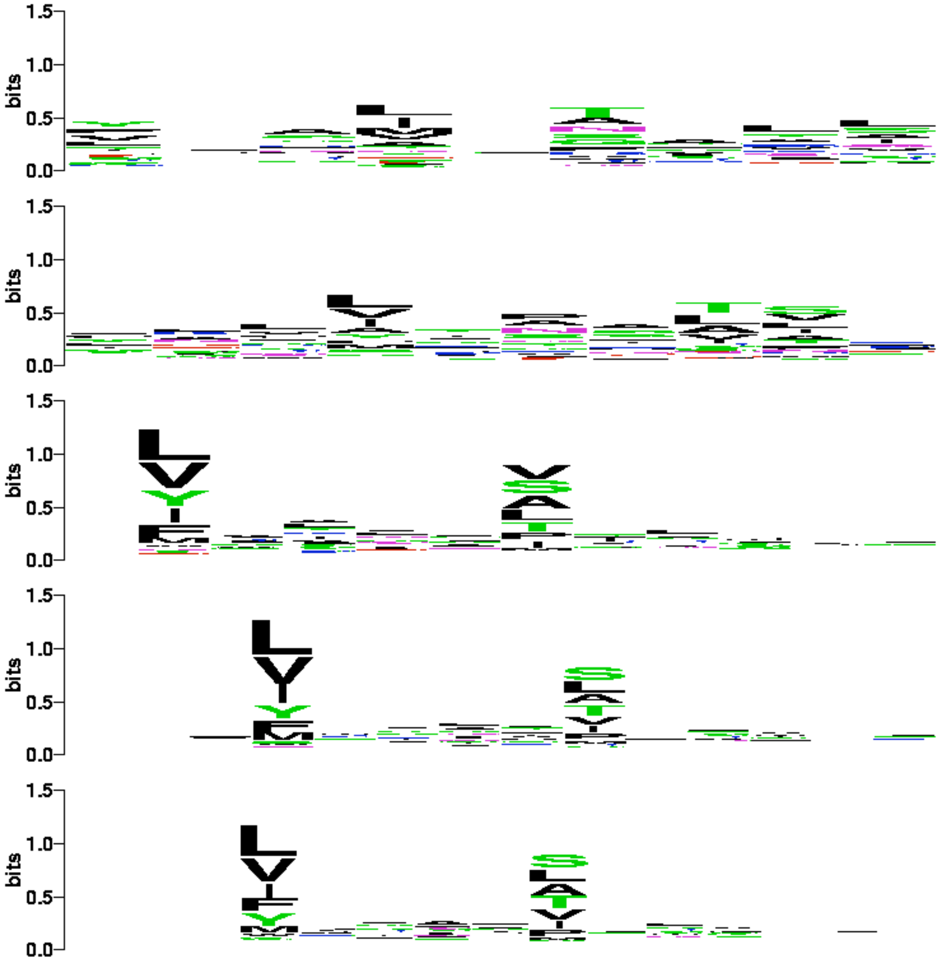


Figure 5: PSSM Logos for MHC (all alleles), lengths (9, 10, 12, 14, and 15)

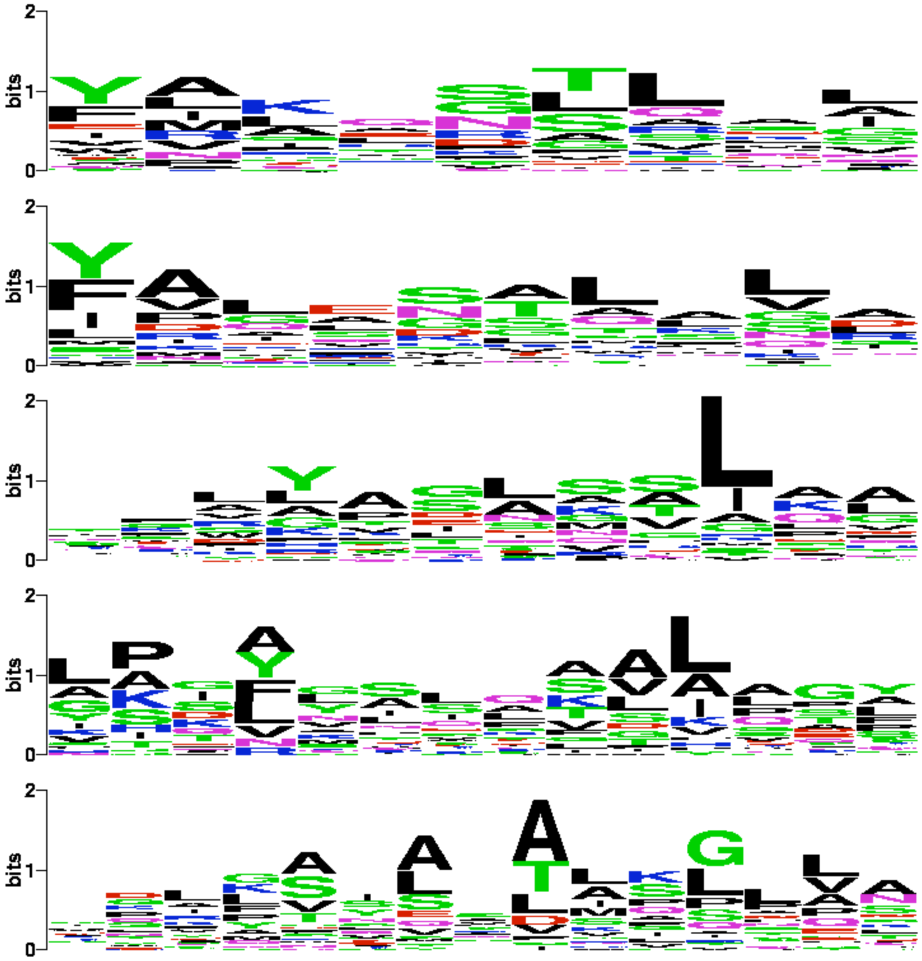


Figure 6: PSSM Logos for TCR (2 alleles), lengths (9, 10, 12, 14, and 15)

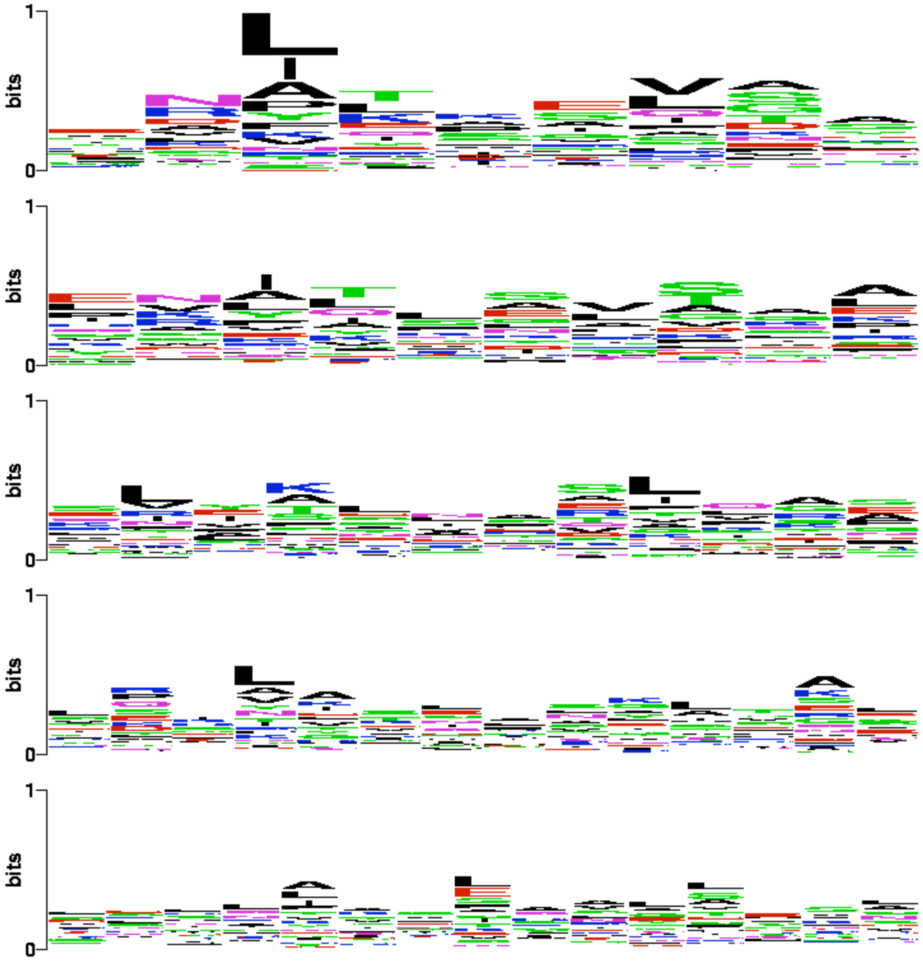


Figure 7: PSSM Logos for TCR (all alleles), lengths (9, 10, 12, 14, and 15)

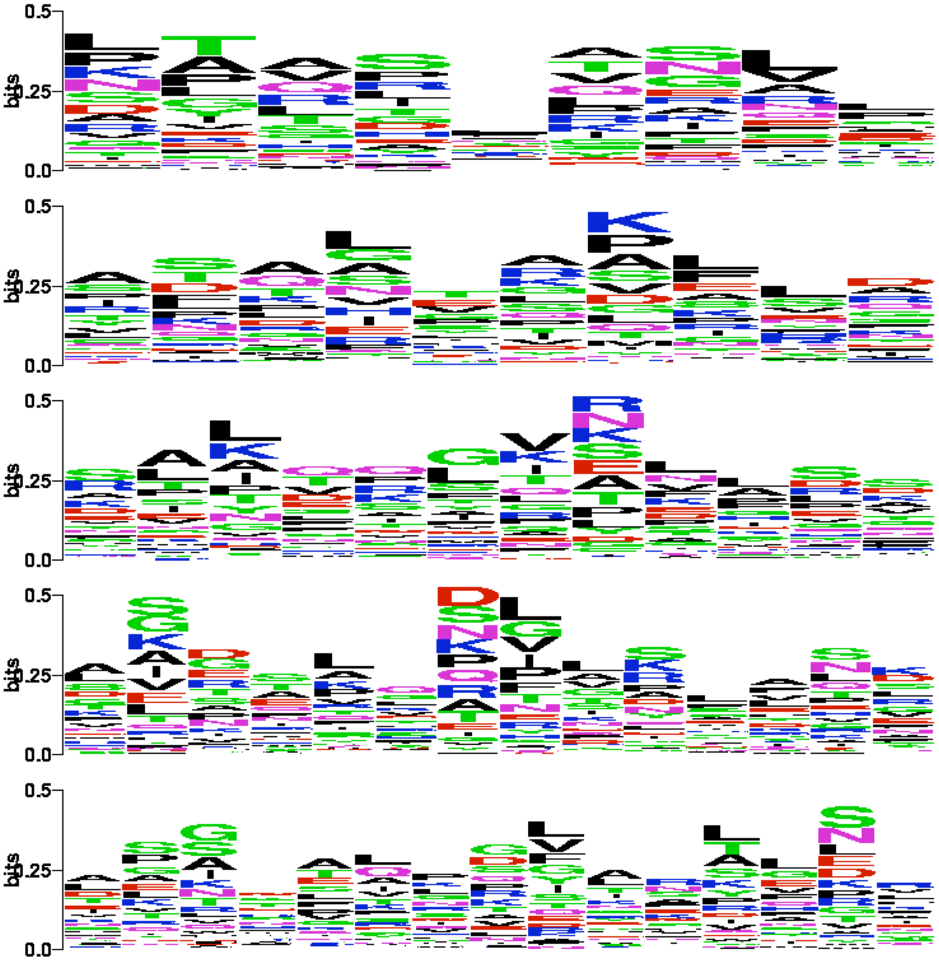


Figure 8: PSSM Logos for BCR, lengths (9, 10, 12, 14, and 15)

C ROC Curves

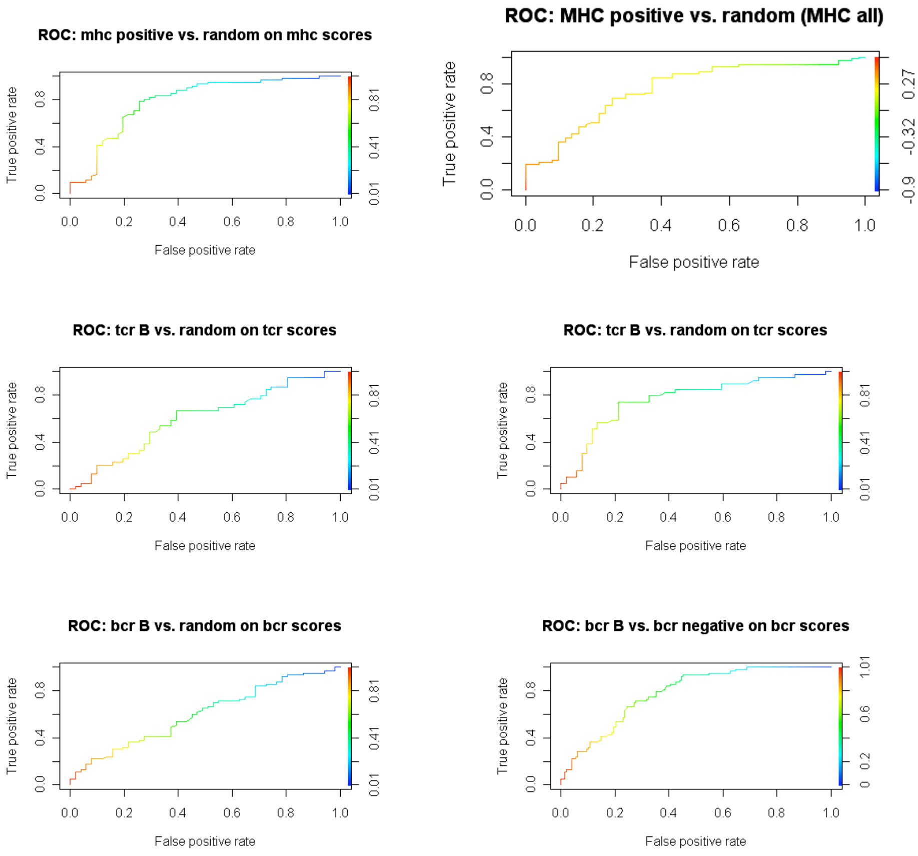


Figure 9: ROC Curves