Published in final edited form as:

J Chem Inf Model. 2017 August 28; 57(8): 1990–1998. doi:10.1021/acs.jcim.7b00118.

Modeling Sequence Dependent Peptide Fluctuations in Immunologic Recognition

Cory M. Ayres, Timothy P. Riley, Steven A. Corcelli*, and Brian M. Baker*

Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana 46556, United States

Abstract

In cellular immunity, T cells recognize peptide antigens bound and presented by major histocompatibility complex (MHC) proteins. The motions of peptides bound to MHC play a significant role in determining immunogenicity. However, existing approaches for investigating peptide/MHC motional dynamics are challenging or of low throughput, hindering the development of algorithms for predicting immunogenicity from large databases, such as those of tumor or genetically unstable viral genomes. We addressed this by performing exhaustive molecular dynamics simulations on a large structural database of peptides bound to the most commonly expressed human class I MHC protein, HLA-A*0201. The simulations reproduced experimental indicators of motion and were used to generate simple models for predicting site-specific, rapid motions of bound peptides through differences in their sequence and chemical composition alone. The models can easily be applied on their own or incorporated into immunogenicity prediction algorithms. Beyond their predictive power, the models provided insight into how amino acid substitutions can influence peptide and protein motions and how dynamic information is communicated across peptides. They also indicate a link between peptide rigidity and hydrophobicity, two features known to be important in influencing cellular immune responses.

Graphical Abstract

Corresponding Authors: brian-baker@nd.edu, scorcelli@nd.edu.

Notes

The authors declare no competing financial interests.

Supporting Information

PDF file: Structure accession codes, peptide sequences, raw RMS fluctuation data, terms and weights of models, and additional figures. Spreadsheet: List of amino acid specific terms and associated values considered during the construction of the predictive models of peptide fluctuations. This information is available free of charge via the Internet at <http://pubs.acs.org>.

Introduction

Recognition and elimination of cells expressing aberrant or non-self protein is a primary function of the adaptive immune response. The process is facilitated through the presentation of peptides derived from intracellular protein by class I major histocompatibility complex (MHC) proteins on cell surfaces. T cells can then sample peptide-MHC complexes via their T cell receptor (TCR), with sufficiently strong TCR binding triggering peptide-dependent immune responses. Crystallographic structures have been crucial in understanding peptide-MHC complexes and their interactions with TCRs.¹ Recent findings, though, have emphasized the role of molecular motion in influencing T cell immunity, with motional properties such as high frequency fluctuations of peptides in MHC binding grooves influencing peptide binding, TCR recognition, and ultimately immunogenicity.²⁻⁵

While there are many techniques available to investigate peptide-MHC dynamics, molecular dynamics simulations have been most commonly used owing to their ability to provide atomic-level detail across the entire peptide. For example, we recently compared the dynamics of complexes with wild-type and tumor-derived peptides and could correlate differences in the magnitude of C α fluctuations with tumor antigen immunogenicity.⁵ In another study, molecular dynamics simulations were used to show that a peptide-based vaccine candidate's poor immunogenicity was associated with its more rapid mobility in the MHC binding groove.⁴

Given this progress, there is growing interest in broadly implementing molecular dynamics simulations to help screen peptide immunogenicity. Results from such screening could facilitate the discovery of immunodominant epitopes in pathogens and cancers and facilitate vaccine design. However, such broad implementation could be challenging, as pathogen genomes can be very large and cancers can harbor up to 400 mutations per megabase.⁶ Thus, comprehensive screening could require hundreds if not thousands of separate molecular dynamics simulations, each requiring lengthy, computationally expensive simulations.

To help bypass these limitations and facilitate the incorporation of high frequency motion into peptide immunogenicity screening, we sought to generate simple mathematical models for predicting rapid peptide fluctuations based on sequence and chemical composition alone. This was done by generating fully solvated atomistic 1 μ s simulations of 73 nonredundant complexes of nonamer peptides presented by HLA-A*0201 whose crystal structures are available in the protein data bank. We focused on complexes with HLA-A*0201 as it is the most common class I MHC protein in humans,⁷ and nonameric peptides as they are most frequently presented.⁸ From these simulations we generated models which can predict site-specific fluctuations across the entire length of any nonameric peptide. This was achieved through a series of multiple linear regressions which accounted for peptide sequence and chemical composition. The final models accurately predict the fluctuations of nonameric peptides bound to HLA-A*0201 and can be easily used on their own or incorporated into algorithms for predicting peptide immunogenicity. The models also indicate a link between peptide rigidity and hydrophobicity, two features known to be important in influencing cellular immune responses.

Methods

Molecular Dynamics Simulations

Simulations were performed with the GPU accelerated version of the AMBER 14 molecular dynamics suite and utilized the ff14SB force field.^{9–11} All starting coordinates were obtained from the Protein Data Bank. When multiple molecules were present in the asymmetric unit, coordinates from the first were used. Terminal residues of the peptide were modeled in their charged states. Missing side chains or residues, almost always localized to the $\alpha 3$ domain, were modeled in using the crystal structure of HLA-A*0201 presenting the Tax peptide (PDB accession code 1DUZ) as a template (1DUZ was chosen as a template owing to the greater level of characterization of the Tax/HLA-A2 complex in our laboratory compared to other complexes).¹² Prior to preproduction, all systems were charge neutralized with sodium ions and solvated with SPC/E water. Systems were then energy minimized to 300K using a Langevin thermostat and solute restraints. After minimization and heating, solute restraints were gradually relaxed from 25 kcal/mol/Å² to 0 kcal/mol/Å² in the NPT ensemble. Following removal of solute restraints, volume was fixed to the average volume of a 100 ps NPT simulation with no restraints. After a brief 50 ps simulation in the NVT ensemble, production trajectories were calculated with a 2 fs time step and a 10 Å cutoff for non-covalent interactions. Bonds involving hydrogen were restrained using the SHAKE algorithm.¹³ Simulations were calculated for a total time of 1 μ s. Trajectories were output every picosecond and utilized a random seed based on the current date and time for Langevin dynamics. The C α RMS fluctuations were calculated following the generation of production trajectories. Each trajectory was subjected to a global superimposition of all peptide/MHC C α atoms to their respective initial coordinates through use of the “rmsd” implementation in cpptraj.¹⁴ This was performed to remove the global translation and tumbling of the peptide-MHC. Following this, C α RMS fluctuations for each residue were calculated from the superimposed trajectory through use of the “atomicfluct” implementation in cpptraj. All simulations were performed using Titan X GPUs with an average wall time of 680 hours per simulation.

In earlier work, we noted that unstable peptides showed a tendency for large conformational changes and eventual dissociation from the MHC peptide binding groove.⁵ To restrict our study to high frequency motions of stably-bound peptides, we eliminated trajectories in which the peptide was not stably presented by the HLA-A*0201 molecule for the duration of the simulation, indicated by a RMS fluctuation of 2.5 Å or greater at the C or N terminus. In the case that instability was observed in preproduction, the preproduction process was repeated as described above but with an additional 10 ns NPT preproduction simulation. If the peptide was again unstably presented, that structure was omitted from further study. Simulations were also excluded if there was low frequency conformational sampling by the peptide, defined as peptide C α RMS deviations falling outside of a 1 Å range for at least 70% of the simulation. In total, 11 out of 73 structures with unique peptide sequences were omitted from the study, with nine removed due to peptide instability and two removed due to low frequency conformational sampling (Table S1). Of the nine simulations which were removed due to instability, four had poor anchor residues and two had missing electron

density in the central region of the peptide. There were no other commonalities among the unstable peptides.

Generation of Predictive Models

Prior to constructing the models, an array of potential terms was generated, with each term addressing differences in chemical composition of the training set peptides. Terms considered were:

- Molecular weight of the residue
- Volume of the residue
- Number of hydrogen bond acceptors in the side chain
- Number of hydrogen bond donors in side chain
- Sum of hydrogen bond donors and acceptors in side chain
- ff14sb parametrized charge for Ca
- Kyte Doolittle, Whimley White, Hessa, and Moon Fleming hydrophathy values for the residue
- Net charge of the residue
- Binary Terms – is the residue or does the residue possess a:
 - Acidic
 - Basic
 - Aliphatic
 - Aromatic
 - Hydrophilic
 - Hydrophobic
 - Nonpolar
 - Neutral
 - Uncharged
 - Polar
 - Positive charge
 - Negative charge
 - Hydroxyl group
 - Sulfur atom

The final array consisted of a total of 1665 terms. These 1665 terms described the chemical and other properties of each position of the peptide, or were a summation of each term across every permutable range (1–2, 2–3, ... 8–9, 1–3, 1–4, ..., 1–9). Using this array, a total of seven models were constructed which predicted the RMS fluctuations of the 52 training

set peptides at positions 1,4,5,6, or 7 as well as the averaged RMS fluctuations for positions 2–3 or positions 8–9. To construct the models, multiple rounds of linear regression were performed in a stepwise selection process through the use of the `fitlm` function in Matlab. Initially, for each model, a linear regression was performed for that position's set of RMS fluctuations against every entry in the 1665 term array. After this, correlation coefficients for each term were calculated. After calculating the correlation coefficients, a term was retained if it had both a high correlation coefficient and described the position being investigated or the positions immediately adjacent. Following this, a linear regression was performed against the array of 1665 terms once again, including the previously identified term. As in the initial regression, a second term was retained if it further improved the correlation coefficient and described the position being modeled or positions immediately adjacent. This process was repeated until each model had up to five terms. If during the construction of the model a term's p-value suggested it was no longer significant, indicated by a p-value >0.25 ,¹⁵ it was removed and the forward stepwise selection process performed again. This methodology was repeated until no significant improvement could be made to the model or the model met its limit of five terms. During the construction of the seven models, several models found significant improvement with the inclusion of terms describing distal positions. As these terms were repeatedly identified in several models, they were included in the final models.

Data Sets

In total, 92 separate 1 μ s molecular dynamics simulations were performed for HLA-A*0201 presenting different peptides. After applying the above rejection criteria, the number of simulations used in the study was reduced to 81. These 81 simulations were partitioned into the following data sets:

- Training set: 52 simulations were included in the training data set. The RMS fluctuations from these simulations were used to train the seven predictive models. These simulations were chosen to include a wide range of sequence variability and include a large range of RMS fluctuations. The frequency plot for amino acids in the training set (Figure S1) was generated using Weblogo.¹⁶
- Validation set: 10 simulations distinct from those in the training set were included in the validation data set. The results from these simulations were excluded when constructing predictive models. The validation data set was approximately 20% of the training set, and was used to verify the models.
- Repeat set: 19 simulations distinct from those in the training and validation sets were included in the repeat data set. These simulations were used to verify reproducibility and were used in generating rejection criteria. Of the 19 simulations in the repeat set, 14 were calculated from the same initial coordinates as the original simulations, and five were calculated using initial coordinates of a different structure of HLA-A*0201 presenting the same peptide (e.g., PDB entries 1DUZ and 1HHK).

Validation of linear models

Monte Carlo Cross Validation (MCCV) and Y-shuffling were used to validate the models. For MCCV, one random point was removed from the training dataset and a multiple linear regression was performed incorporating the predictor variables of the original model. We then calculated the RMS error of the left out point to its true value using the newly generated model. This process was repeated 10,000 times and an average RMS error was calculated from all 10,000 repeats. We repeated this process but leaving out between two to five points in order to validate the robustness and overfitting of the original models to the training set data.

For Y-shuffling, three different protocols were used to verify the statistical significance of each model as well as its overall performance.¹⁷ The first protocol was to shuffle the RMS fluctuations at of each model and to perform new multiple linear regressions with the unshuffled predictor variables. The second protocol was to perform new multiple linear regressions of the unshuffled RMS fluctuations with randomly chosen predictor variables from the 1665 term matrix. The third protocol was to perform new multiple linear regressions of the unshuffled RMS fluctuations with random integers as predictor variables, chosen from a range between 0 and 32767. This process was repeated 10,000 times for each protocol, and statistical significance of the original models was determined by comparing the correlation coefficient of the original model to the correlation coefficient of the 1,000 best performing validation models.

Results

Differences and reproducibility of RMS fluctuations

Focusing on individual residues of the peptide, the training set simulations showed a large range of alpha-carbon (C α) RMS fluctuations as indicated in Table S2 and visualized in Figure 1. The minimum and maximum RMS fluctuation for any individual position was 0.57 Å and 2.39 Å, respectively. The range of RMS fluctuations for any position ranged from 1.12 Å and 1.62 Å, with an average of 1.39 Å across all nine positions. Several peptides were rigid across their length when comparing per-residue fluctuations to positional averages, and some were flexible across their length. Other peptides showed no clear trends. Finally, some peptides seemed to have partitioned fluctuations, with one half of the peptide comparatively flexible and the other comparatively rigid.

We next verified the reproducibility of our simulations by generating 19 replicate 1 μ s simulations. Out of these 19, 14 were repeats generated utilizing the same starting coordinates from the initial simulations (structural repeats), and five were repeats generated from a different crystallographic structure of HLA-A2 presenting the same peptide (sequence repeats) (Table S1). There was reasonable agreement between all pairs of simulations, with a correlation coefficient of 0.68 between pairs of observed RMS fluctuations. This agreement persisted when separately comparing fluctuations for the 14 structural repeats and 5 sequence repeats, which yielded correlation coefficients of 0.65 and 0.82, respectively (Figure S3). While it may appear that that the sequence repeats perform

better, this is merely a consequence of the smaller number of simulations performed and should not be considered a significant improvement.

Comparison of simulation data with experimental observations

After confirming that the simulations generated a wide range of reproducible RMS fluctuations, we sought to determine how the simulations recapitulated experimental observations. As expected based on previous work,¹⁸ we observed no clear correlations between C α fluctuations and raw or normalized temperature factors (correlation coefficients of 0.15 and 0.24, respectively). This is expected as temperature factors generally report poorly on fast atomic motions in protein structures.¹⁹ However, the simulations did reproduce direct experimental evidence of peptide conformational heterogeneity. For the MART-1 (A(A/L)GIGILTV) and HER-2/neu (I(I/L/L)SA(V/L/L)VGI(L/V)) peptides, modifications of “primary anchor” positions (denoted by bold or italicized sequence) resulted in altered flexibility or stability of the peptide, as evaluated by weak or missing electron density (MART-1, HER-2/neu) and the presence of multiple spectroscopic signals (MART-1).^{3–4, 20} For the MART-1 peptide, in agreement with NMR data we observed two conformations centered at position 5 for the modified peptide, whereas no conformational change was observed for the wild type peptide (Figure 2). For the HER-2/neu peptides, we observed that the two peptides with weak or missing electron density in crystal structures began to dissociate from the HLA-A*0201 molecule, whereas the peptide with clear density in the central region remained stably bound. Collectively, the data indicate the simulations reproduce experimental inferences of the dynamical properties of various peptides, and are thus suitable for use in the generation of predictive models for peptide fluctuations.

Partitioned correlations in peptide RMS fluctuations

As noted above, modification of peptide primary anchors can alter motional and structural properties across the peptide.^{3–4, 20} To explore such non-local influences, we compared the magnitude of C α fluctuations at each position with those at every other position by generating a cross-correlation matrix. We observed strong correlations with immediately adjacent residues, and more moderate correlations for sites up to three residues away (Figure 3). This observation is consistent with experimental findings that peptide binding and conformational properties can be influenced non-locally.^{3–4, 20} The more distal inter-residue correlations were stronger in the N-terminal portion of the peptide (positions 1–5) than the C-terminal (positions 6–9). Thus, the correlation analysis suggests bound peptides can be considered as two separate sections, split at position 5. The mechanistic details of this observation are discussed further below.

Predictive models of peptide RMS fluctuations

Prior relationships between peptide motion and immunogenicity were drawn from differences in fluctuations at individual amino acid positions. Thus, we sought to use the simulation data to generate independent, site-specific models, with each model predicting the RMS fluctuation at specific positions along a nonameric peptide, accounting for non-local influences when appropriate. To achieve this, we performed multi-linear regression analyses of site specific C α RMS fluctuations versus differences in peptide and amino acid

properties within the simulations. Since positions 2 and 9 have limited sequence variability due to their roles as primary anchors (Figure S1), we combined position 2 with 3 and position 8 with 9 and treated the 2/3 and 8/9 combinations as single sites. This was done by averaging the Ca RMS fluctuations for positions 2–3 and 8–9 (we chose to average positions 2–3 instead of 1–2 as the fluctuations of positions 2 and 3 were more strongly correlated and of a smaller range than those for positions 1 and 2). Overall, seven models were generated for predicting site specific Ca fluctuations on the nanosecond timescale.

Terms included in the models described the chemical and physical properties of each amino acid, such as hydropathy, volume, and charge character.^{21–24} We also considered amino preferences for HLA-A2.²⁵ We considered terms for single sites as well as ranges. Given the total sample size of 52 simulations, each model was permitted up to five predictors to avoid overfitting. Models were constructed via sequential rounds of multiple linear regressions. Preference was given to nonbinary terms to avoid overfitting, although some highly significant binary terms were retained (for example, whether the position 1 side chain was polar or apolar). Preference was also given to terms that described the position being modeled as well as those immediately adjacent. Predictors that described distal positions were permitted if they occurred in multiple models. Overall, a total of 13 unique terms were used to construct the seven final models (Table 1). Notably, amino acid preferences for HLA-A2 were not among the significant terms, consistent with the view that these preferences are dictated by chemical properties. The final models have correlation coefficients ranging from 0.57 to 0.67. When all simulated and predicted fluctuations were compiled, the correlation coefficient was 0.71 with an RMS error of 0.22. The intercept of the best fit line when comparing all data points was 0 and the slope 1, indicating an unbiased fit between predicted and observed Ca fluctuations. (Figure 4).

Model validation

To test for overfitting, 10,000 rounds of Monte Carlo cross validation (MCCV) were performed on each of the seven models. The process was performed five times, with each repeat leaving out one additional random point (one to five points; between 2 and 10% of the data). The average RMS error for the omitted points with the leave-one-out MCCV was 0.19 Å. The average RMS error of the omitted points for the leave-five-out MCCV was 0.23 Å. These errors are comparable to the average RMS Error of 0.22 Å in the aggregate fit of all predicted RMS fluctuations.

To further validate the models, ten additional simulations were performed for nonamer peptide-HLA-A*0201 complexes whose structures are in the Protein Data Bank but were not included in the original training set (Table S1). Ca RMS fluctuations computed from the seven models reproduced the simulation fluctuations well, recapitulating the observed linear trend and with 86% of the data points falling within the 95% confidence bands (Figure 4, green points).

Finally, we employed three Y-shuffling validation protocols to investigate the predictive ability of our models.¹⁷ Each protocol was performed 10,000 times and the statistical significance was drawn from the best 1,000 models. For all three protocols, the seven original models outperformed the validation models at the 95% confidence level ($p < 0.05$),

and for two of the three protocols, the original models outperformed at the 99% confidence level ($p < 0.01$). The Y-shuffling results confirm that our models have a high level of predictive power, and that they are among the best performing given the molecular dynamics data and chemical information data used in generating them.

Recurrent incorporation of nonlocal terms

As mentioned previously, models were permitted terms for distal positions if the term was deemed significant in multiple models. For example, the volume of the peptide C-terminus appears in the models for sites 1, 2/3, 4, and 5. Additionally, the size of the P5 side chain appeared in several models. To elucidate the mechanisms of the non-local influences, we investigated select recurrent terms and how they impacted distal fluctuations. We first addressed the recurrence of the volume at site 5 impacting the fluctuations of sites one through six. In the models for sites 1, 2/3, 4, 5, and 6, the weight for the volume of site 5 is negative, indicating greater volume diminishes distal fluctuations. The weights become smaller the more removed a site is from position 5: for the P5 model, the volume term translates into reductions in RMS fluctuations from 0.27 to 0.85 Å; for the position 1 model the range is 0.11 to 0.34 Å. Non-local influences therefore taper the more removed a site is from position 5.

We hypothesized that the volume of position five accounted for the number of interactions between the side chain with the MHC, as well as the side chain solvent accessible surface area. We studied this by replacing P5 volume with terms describing the side chain solvent accessible surface area and the number of interatomic interactions between the side chain and the MHC protein with over 20% occupancy (Figure 5A). Performing a new multiple linear regression with the two new terms generated a model with a correlation coefficient of 0.65 and RMS error of 0.25, compared to the original model's correlation coefficient of 0.63 and RMS error of 0.26, indicating a similar predictive power and supporting the hypothesis. As to why P5 volume impacts N-terminal and not C-terminal fluctuations, we speculate that propagation to the C-terminus is hindered by the structural heterogeneity of P6: in 40% of the structures the P6 side chains are buried, and in 60% the P6 side chains are exposed.²⁶ As such, while P5 fluctuations may be impacting the dynamics of P6, the impact is not consistent across simulations. Position 6 therefore seems to act as a “checkpoint” on the propagation of peptide dynamics due to its varied structural properties. In contrast, positions 2–4 are repeatedly found in the same conformation and positions, leading to consistent P2–P4 motional propagation across the simulations. Structural heterogeneity at P6 could explain the partitioning of correlated motions shown in Figure 3.

We next investigated the impact of the ff14SB parameterized P7 Ca charge on the fluctuations of positions 6 through 9. All three models have positive weights for the P7 Ca charge; however, the charge itself can be positive or negative. Further simulations artificially varying the P7 Ca charge did not, however, show a clear connection between the charge and fluctuations. We thus sought to identify if the P7 Ca charge was correlated with other properties. New linear regressions of the Ca charge vs. the full set of terms showed a correlation between hydrophathy, volume, aromaticity, and polar character, with an overall correlation coefficient of 0.83 for all amino acids except lysine and arginine, whose Ca

charges are a magnitude smaller than the charges for the other amino acids, and are not represented in any structures of peptides bound to HLA-A2. Thus, the P7 Ca charge is accounting for a diverse set of molecular properties that can influence peptide fluctuations.

We also investigated the role of volume at P7 - P9, and how changes in volume can impact changes in fluctuations all the way to position 1. Due to the lack of motional correlations between the peptide termini, we hypothesized the C-terminal residues could modify the geometry of the MHC and impact peptide N-terminal fluctuations indirectly via a “through protein” effect. As positions 7 – 9 lie adjacent to the short arm of the MHC $\alpha 2$ helix, we asked if increased volume at P7 - P9 altered the geometry of the MHC and served to “open” the peptide binding groove. We measured average distances between each pair of residues across the $\alpha 1$ and $\alpha 2$ helices for each simulation, and additional linear regressions were calculated to determine if a correlation existed between the P7 - P9 volumes and the inter-helix distances. A positive relationship was observed between volumes and an overall broadening of the binding groove between pairs of residues adjacent to peptide positions 5 through 9 (Figure 5B). Indeed, replacing the P7 – P9 volume with the average distance between Ala69 and Leu166, which lie next to position 5, generated similar or somewhat improved models for models P1 through P7. The correlation coefficients for models for positions 1 and 2 through 3 increased by approximately 0.09, and the correlation coefficients for models 4 through 7 decreased by approximately 0.04. Thus, changes in side chain volume at P7 - P9 impact the width of the binding groove at P5 and P6. This increase permits greater fluctuations at these positions, which for P5 can propagate across the N-terminal half of the peptide. These observations can explain subtle variations in HLA-A*0201 structures seen with different peptides bound.²⁷

Additional trends in incorporated terms

Lastly, we looked for any trends in the terms incorporated within the seven models. One notable observation was that multiple models predict a decrease in peptide fluctuations as hydrophobic residues are incorporated. Models for sites 4–7 have negative weights for residue hydrophobicity using the Kyte-Doolittle scale. The site 8/9 model has a negative weight for the number of hydrophobic residues in positions 6–9. Additionally, the model for site 1 has a positive weight for polar residues at this position. Furthermore, the position 7 Ca charge present in three models suggests a reduction in fluctuation as the charge becomes more negative, and the regression models performed above indicate a decrease in charge the more hydrophobic the residue is as indicated by the Kyte-Doolittle scale. Collectively, these translate into reduced motions with increasing hydrophobicity. The connection between mobility and hydrophobicity is significant, as it links two prior observations of the determinants of peptide immunogenicity: reduced flexibility as we have previously reported,⁵ and greater hydrophobic character as reported by others.²⁸ Previous work suggested greater hydrophobicity may be associated with favoring TCR binding through burial of apolar surface area, but an alternative (or additional) influence could be a reduced entropic penalty for receptor binding resulting from smaller peptide fluctuations.

Discussion

Advances in immunology and immunotherapy have led to considerable interest in the development of algorithms for predicting the immunogenicity of peptides bound and presented by class I MHC proteins. The majority of such tools are based on predicting peptide binding affinity,²⁹ although other physical properties are known to influence immunogenicity. One such property is the rapid motion of peptides within MHC binding grooves. For example, enhanced conformational sampling of a modified tumor antigen led to the loss of immunogenicity and the failure of a vaccine candidate,⁴ and the immunogenicity of cancer neo-antigens has been positively correlated with peptide rigidity.⁵ Although useful in isolated cases or with small sets of peptides, molecular dynamics simulations are prohibitive for large-scale prediction efforts, such as those needed for work with cancer neo-antigens or pathogens with large or unstable viruses.

Here we aimed to bypass this limitation by developing a set of models for predicting site-specific fluctuations of nonameric peptides bound to the most common class I MHC protein HLA-A*0201. We achieved this by first generating an exhaustive library of lengthy molecular dynamics simulations for HLA-A*0201 presenting nonameric peptides, collecting almost 100 μ s of simulation data that reproduced available experimental indicators of motion. From these simulations, we computed RMS fluctuations and performed a series of multiple linear regressions against terms reflecting peptide sequence and chemical composition. As the resulting models bypass the need for lengthy molecular dynamics simulations, they are easily incorporated into immunogenicity algorithms and can be readily deployed against large databases.

In addition to their predictive utility, our results revealed non-local influences on peptide motion. These findings are of interest given the common approach of modifying select positions to enhance peptide binding to MHC proteins or to alter recognition by T cell receptors.^{30–31} Such modifications have in some cases unexpectedly altered peptide motion.^{2,4} peptide conformation,²⁰ and T cell recognition,³² and our findings reveal these effects can occur via through-peptide and through-protein mechanisms. Finally, our results suggest a link between peptide rigidity and hydrophobicity, two previously established determinants of peptide immunogenicity. Areas for additional work include extension to peptides of different lengths, lengthier simulations which can capture slower, low-frequency peptide motions, such as those often associated with T cell receptor binding,^{33–34} and the inclusion of MHC proteins other than HLA-A*0201. On this last note, an interesting area for future studies will be the exploration of MHC polymorphisms in influencing peptide motional properties.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

Supported by NIGMS grant R35118166 to BMB.

We thank the National Institute of General Medical Sciences (NIGMS R35118166) for financial support of this research.

References

1. Rossjohn J, Gras S, Miles JJ, Turner SJ, Godfrey DI, McCluskey J. T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules. *Ann Rev Immunol*. 2015; 33(1):169–200. [PubMed: 25493333]
2. Borbulevych OY, Insaiddoo FK, Baxter TK, Powell DJ Jr, Johnson LA, Restifo NP, Baker BM. Structures of Mart-1(26/27-35) Peptide/Hla-A2 Complexes Reveal a Remarkable Disconnect between Antigen Structural Homology and T Cell Recognition. *J Mol Biol*. 2007; 372(5):1123–1136. [PubMed: 17719062]
3. Sharma AK, Kuhns JJ, Yan S, Friedline RH, Long B, Tisch R, Collins EJ. Class I Major Histocompatibility Complex Anchor Substitutions Alter the Conformation of T Cell Receptor Contacts. *J Biol Chem*. 2001; 276(24):21443–21449. [PubMed: 11287414]
4. Insaiddoo FK, Borbulevych OY, Hossain M, Santhanagopalan SM, Baxter TK, Baker BM. Loss of T Cell Antigen Recognition Arising from Changes in Peptide and Major Histocompatibility Complex Protein Flexibility: Implications for Vaccine Design. *J Biol Chem*. 2011; 286(46):40163–40173. [PubMed: 21937447]
5. Duan F, Duitama J, Al Seesi S, Ayres CM, Corcelli SA, Pawashe AP, Blanchard T, McMahon D, Sidney J, Sette A, Baker BM, Mandoiu II, Srivastava PK. Genomic and Bioinformatic Profiling of Mutational Neopeptides Reveals New Rules to Predict Anticancer Immunogenicity. *J Exp Med*. 2014; 211(11):2231–2248. [PubMed: 25245761]
6. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale A-L, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjord JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Illicic T, Imbeaud S, Imielinski M, Jager N, Jones DTW, Jones D, Knappskog S, Kool M, Lakhani SR, Lopez-Otin C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt ANJ, Valdes-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR, PedBrain I, Zucman-Rossi J, Andrew Futreal P, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR. Signatures of Mutational Processes in Human Cancer. *Nature*. 2013; 500(7463):415–421. [PubMed: 23945592]
7. Chen K, Liu J, Ren E. Structural and Functional Distinctiveness of Hla-A2 Allelic Variants. *Imm Res*. 2012; 53(1–3):182–190.
8. Trolle T, McMurtrey CP, Sidney J, Bardet W, Osborn SC, Kaever T, Sette A, Hildebrand WH, Nielsen M, Peters B. The Length Distribution of Class I–Restricted T Cell Epitopes Is Determined by Both Peptide Supply and Mhc Allele–Specific Binding Preference. *J Immunol*. 2016; 196(4):1480–1487. [PubMed: 26783342]
9. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC. Routine Microsecond Molecular Dynamics Simulations with Amber on Gpus. 2. Explicit Solvent Particle Mesh Ewald. *J Chem Theory Comput*. 2013; 9(9):3878–3888. [PubMed: 26592383]
10. Salomon-Ferrer R, Case DA, Walker RC. An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. 2013; 3(2):198–210.
11. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. Ff14sb: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99sb. *J Chem Theory Comput*. 2015; 11(8):3696–3713. [PubMed: 26574453]
12. Khan AR, Baker BM, Ghosh P, Biddison WE, Wiley DC. The Structure and Stability of an Hla-a*0201/Octameric Tax Peptide Complex with an Empty Conserved Peptide-N-Terminal Binding Site. *J Immunol*. 2000; 164(12):6398–6405. [PubMed: 10843695]

13. Ryckaert J-P, Ciccotti G, Berendsen HJC. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J Comput Phy.* 1977; 23(3):327–341.
14. Roe DR, Cheatham TE. Ptraaj and Cpptraaj: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput.* 2013; 9(7):3084–3095. [PubMed: 26583988]
15. Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful Selection of Variables in Logistic Regression. *Source Code Biol Med.* 2008; 3(1):17. [PubMed: 19087314]
16. Crooks GE, Hon G, Chandonia J-M, Brenner SE. Weblogo: A Sequence Logo Generator. *Gen Res.* 2004; 14(6):1188–1190.
17. Rücker C, Rücker G, Meringer M. Y-Randomization and Its Variants in Qspr/Qsar. *J Chem Inf Mod.* 2007; 47(6):2345–2357.
18. Ayres CM, Scott DR, Corcelli SA, Baker BM. Australian Pancreatic Cancer Genome, I., Consortium, I. B. C., Consortium, I. M.-S. Differential Utilization of Binding Loop Flexibility in T Cell Receptor Ligand Selection and Cross-Reactivity. *Sci Rep.* 2016; 6:25070. [PubMed: 27118724]
19. Kuzmanic A, Pannu NS, Zagrovic B. X-Ray Refinement Significantly Underestimates the Level of Microscopic Heterogeneity in Biomolecular Crystals. *Nat Commun.* 2014; 5:3220. [PubMed: 24504120]
20. Kuhns JJ, Batalia MA, Yan S, Collins EJ. Poor Binding of a Her-2/Neu Epitope (Gp2) to Hla-A2.1 Is Due to a Lack of Interactions with the Center of the Peptide. *J Biol Chem.* 1999; 274(51):36422–36427. [PubMed: 10593938]
21. Robinson J, Mistry K, McWilliam H, Lopez R, Parham P, Marsh SG. The Imgt/Hla Database. *Nuc Acid Res.* 2011; 39(Database issue):D1171–1176.
22. Kyte J, Doolittle RF. A Simple Method for Displaying the Hydropathic Character of a Protein. *J Mol Biol.* 1982; 157(1):105–132. [PubMed: 7108955]
23. Wimley WC, White SH. Experimentally Determined Hydrophobicity Scale for Proteins at Membrane Interfaces. *Nat Struct Biol.* 1996; 3(10):842–848. [PubMed: 8836100]
24. Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, Nilsson I, White SH, von Heijne G. Recognition of Transmembrane Helices by the Endoplasmic Reticulum Translocon. *Nature.* 2005; 433(7024):377–381. [PubMed: 15674282]
25. Falk K, Rotzschke O, Stevanovic S, Jung G, Rammensee HG. Allele-Specific Motifs Revealed by Sequencing of Self-Peptides Eluted from Mhc Molecules. *Nature.* 1991; 351(6324):290–296. [PubMed: 1709722]
26. Chen H, Zhou H-X. Prediction of Solvent Accessibility and Sites of Deleterious Mutations from Protein Sequence. *Nuc Acid Res.* 2005; 33(10):3193–3199.
27. Hawse WF, Gloor BE, Ayres CM, Kho K, Nuter E, Baker BM. Peptide Modulation of Class I Major Histocompatibility Complex Protein Molecular Flexibility and the Implications for Immune Recognition. *J Biol Chem.* 2013; 288(34):24372–24381. [PubMed: 23836912]
28. Chowell D, Krishna S, Becker PD, Cocita C, Shu J, Tan X, Greenberg PD, Klavinskis LS, Blattman JN, Anderson KS. Tcr Contact Residue Hydrophobicity Is a Hallmark of Immunogenic Cd8+ T Cell Epitopes. *Proc Natl Acad Sci.* 2015; 112(14):E1754–E1762. [PubMed: 25831525]
29. Lundegaard C, Lund O, Buus S, Nielsen M. Major Histocompatibility Complex Class I Binding Predictions as a Tool in Epitope Discovery. *Immunology.* 2010; 130(3):309–318. [PubMed: 20518827]
30. Piepenbrink KH, Borbulevych OY, Sommese RF, Clemens J, Armstrong KM, Desmond C, Do P, Baker BM. Fluorine Substitutions in an Antigenic Peptide Selectively Modulate T-Cell Receptor Binding in a Minimally Perturbing Manner. *Biochem J.* 2009; 423(3):353–361. [PubMed: 19698083]
31. Borbulevych OY, Baxter TK, Yu Z, Restifo NP, Baker BM. Increased Immunogenicity of an Anchor-Modified Tumor-Associated Antigen Is Due to the Enhanced Stability of the Peptide/Mhc Complex: Implications for Vaccine Design. *J Immunol.* 2005; 174(8):4812–4820. [PubMed: 15814707]
32. Cole DK, Edwards ES, Wynn KK, Clement M, Miles JJ, Ladell K, Ekeruche J, Gostick E, Adams KJ, Skowera A, Peakman M, Wooldridge L, Price DA, Sewell AK. Modification of Mhc Anchor

Residues Generates Heteroclitic Peptides That Alter Tcr Binding and T Cell Recognition. *J Immunol.* 2010; 185(4):2600–2610. [PubMed: 20639478]

33. Baker BM, Gagnon SJ, Biddison WE, Wiley DC. Conversion of a T Cell Antagonist into an Agonist by Repairing a Defect in the Tcr/Peptide/Mhc Interface. Implications for Tcr Signaling. *Immunity.* 2000; 13(4):475–484. [PubMed: 11070166]
34. Tynan FE, Reid HH, Kjer-Nielsen L, Miles JJ, Wilce MCJ, Kostenko L, Borg NA, Williamson NA, Beddoe T, Purcell AW, Burrows SR, McCluskey J, Rossjohn J. A T Cell Receptor Flattens a Bulged Antigenic Peptide Presented by a Major Histocompatibility Complex Class I Molecule. *Nat Immunol.* 2007; 8(3):268–276. [PubMed: 17259989]

Figure 1.

RMS fluctuations observed in the simulations. (A) Visualization of peptide Ca RMS fluctuations. Red represents the minimum value observed for that position and cyan represents the maximum. (B) Box plot of the Ca RMS fluctuations at each position. Red points are the raw values, blue boxes are the interquartile range (25% to 75%), whiskers are the greater of either 1.5 times the interquartile range or maximal/minimal values, black lines are the mean for each position and the white circles are the median for each position.

Figure 2.

Simulations replicate experimental observations. (A) Two general conformations were sampled in the simulation of the MART-1 anchor modified peptide, centered around a backbone flip at position 5 (left). These structural observations mimic what is seen crystallographically for the same peptide (right). (B) RMS deviations for the wild type and anchor modified MART-1 peptides relative to their initial coordinates, calculated for the C α atoms of positions 4 through 6. (C) For the HER-2/neu peptides, the peptides which lacked density in the central region of the peptide were not stably presented by HLA-A*0201. The position 9 maximal displacements (translucent, values indicated in inset) from the initial

coordinates (solid) were significantly larger for 1EEY and 1QR1 when compared to the stably presented 1EEZ. (D) Position 9 displacement from the initial coordinates versus time for the HER-2/neu peptides.

Figure 3.

Matrix of correlation coefficients for pairs of observed RMS fluctuations between each peptide position. A value of 1 indicates perfect correlation and 0 no correlation. Values below 0.5 are omitted to emphasize pairs of residues which are moderately-to-strongly correlated.

Figure 4.

Aggregate fit of simulation versus predicted RMSFs for all seven models. The best-fit line is indicated in red, and 95% confidence bands are indicated in blue. The training set is indicated in black, and the validation set is indicated in green.

Figure 5.

Recurrent nonlocal terms show the influence of the MHC protein in peptide dynamics. (A) Final models for position five when including the volume of position 5 (black), or the average side chain solvent accessible surface area and count of interactions (red). The two models generate indistinguishable results, with the best fit line for the volume model in black and the best fit line for the SASA and interaction model in transparent red. (B) New multiple linear regressions were performed to look for correlations between the volumes at peptide positions 7–9 and the average pairwise distances between the $\alpha 1$ and $\alpha 2$ helices throughout the simulations. In total, 137 pairs of distances were found which correlated with the volumes of peptide positions 7, 8 and 9 with a correlation coefficient of 0.5 or greater. To verify that the modulation of groove width correlated with the RMS fluctuations, the average distance of A69 and L166, indicated in red, was used as a replacement for all models which incorporated the P7_9 volume. The incorporation of this term generated similar or slightly improved models for positions 1 through 7, with correlation coefficients for the models at positions 1 through 3 improving by approximately 0.09, and correlation coefficients for positions 4 through seven decreasing by approximately 0.04.

Table 1

List of terms and weights for each of the seven finalized models. In total, 13 unique terms were used to construct all seven models. While most terms describe differences in chemical composition for that model's residue or the residue adjacent, several models include terms describing the chemical composition of distal residues, notably the volume at residue position 5 and the summed volume of residue positions 7 through 9. M.Constant refers to the constant for each individual model, volume refers to the volume for that position or summed volume for those positions, wwHydro is residue hydropathy on the White-Wimley scale, kdHydro is residue hydropathy on the Kyte-Doolittle scale, P1 polar is a binary term for whether the side chain is polar or not, P7C α charge is the ff14SB parametrized charge for the C α carbon for that residue, and P6_9HPhbc is the summed total of hydrophobic residues from positions 6 through 9.

Model Position 1 (cc = 0.67)		
Term	Weight	P-value
P2wwHydro	-5.23E-01	3.28E-04
M.Constant	9.50E-01	5.40E-04
P1Polar	2.01E-01	3.50E-03
P7_9Volume	1.02E-03	1.04E-01
P5Volume	-1.48E-03	1.20E-01
Model Positions 2 – 3 (cc = 0.67)		
Term	Weight	P-value
M.Constant	8.85E-01	6.00E-04
P2wwHydro	-3.38E-01	6.10E-03
P5Volume	-1.93E-03	1.66E-02
P4kdHydro	2.59E-02	3.14E-02
P4Volume	1.13E-03	1.21E-01
P7_9Volume	6.80E-04	1.89E-01
Model Position 4 (cc = 0.59)		
Term	Weight	P-value
P4Volume	2.93E-03	1.98E-03
P5Volume	-2.37E-03	4.41E-03
P7_9Volume	1.58E-03	7.46E-03
M.Constant	5.30E-01	5.38E-02
P5kdHydro	-1.63E-02	1.86E-01
Model Position 5 (cc = 0.62)		
Term	Weight	P-value
P5kdHydro	-5.98E-02	2.34E-04
P5Volume	-3.74E-03	6.60E-04
P7_9Volume	2.12E-03	3.90E-03
M.Constant	8.39E-01	1.25E-02
P4Volume	1.52E-03	1.45E-01
Model Position 6 (cc = 0.58)		
Term	Weight	P-value
P5kdHydro	-5.02E-02	2.72E-03

M.Constant	9.47E-01	3.55E-03
P7CaCharge	3.12E+00	6.92E-03
P9wwHydro	3.41E-01	1.13E-02
P5Volume	-1.95E-03	6.56E-02
P7_9Volume	1.45E-03	7.43E-02

Model Position 7 (cc = 0.65)

Term	Weight	P-value
P9wwHydro	4.32E-01	3.40E-04
P7CaCharge	3.01E+00	2.51E-03
P5kdHydro	-3.70E-02	7.91E-03
M.Constant	7.06E-01	9.18E-03
P7_9Volume	1.50E-03	2.29E-02
P6kdHydro	-2.55E-02	1.02E-01

Model Positions 8 – 9 (cc = 0.57)

Term	Weight	P-value
M.Constant	3.55E+00	3.54E-06
P9wwHydro	8.71E-01	8.44E-05
P9Volume	-1.62E-02	1.31E-03
P7CaCharge	3.44E+00	2.11E-03
PP6_9HPhbc	-1.18E-01	1.45E-02
P7Volume	1.49E-03	1.43E-01