Sub-Class Differences of PH-Dependent HIV GP120-CD4 Interactions

Scott P. Morton[*]
Middle Tennessee State University
Murfreesboro, Tennessee
spm3c@mtmail.mtsu.edu

Jonathan Howton[†]
Middle Tennessee State University
Murfreesboro, Tennessee
jh6w@mtmail.mtsu.edu

Joshua L. Phillips[*][†]
Middle Tennessee State University
Murfreesboro, Tennessee
Joshua.Phillips@mtsu.edu

ABSTRACT

Research in the field of HIV transmission has yet to provide a vaccine for this imponderable virus. Though progress has been made to extend the life of those chronically infected, a solution to the transmission of the disease remains elusive. Previous studies involving electrostatic surface charge analysis revealed the sensitivity of gp120 envelope (Env) protein function to changes in pH across levels consistent with those found in the human body. A prototype computational approach was developed and found to agree with these results. A refined process was developed capable of classifying Env sequences/structures through machine learning techniques. We expound this analytical procedure to encompass residue-level analysis and include minimization steps to ensure the integrity of the protein models. Additionally, the process has been enhanced with advanced data compression techniques to allow for more in-depth analysis of the systems. In this research we explore a new technique termed electrostatic variance masking (EVM), that reveals what we hypothesize to be the mechanistic residues responsible for the pH sensitivity of Env binding site. The data implies that a conserved set of core residues may be responsible for modulation of the binding process in varying environmental conditions mainly involving pH.

CCS CONCEPTS

• **Computing methodologies → Principal component analysis**; **Molecular simulation**; **Distributed simulation**; • **Applied computing → Computational biology**; **Bioinformatics**;

KEYWORDS

HIV; Env; gp120; CD4; electrostatics; binding; pH; mucosa

[*]Center for Computational Science
[†]Department of Computer Science

1 INTRODUCTION

Despite three decades of study on Acquired Immune Deficiency Syndrome (AIDS), a vaccine against the Human Immunodeficiency Virus (HIV) is still in development [22]. HIV's high rate of mutation allows antigenic regions targeted by host immune responses to vary greatly across HIV virions. Most research has focused on inducing so-called broadly neutralizing antibodies (bnAbs) to combat the infection. BnAbs target protein antigenic regions conserved due to functional requirements of the binding process [8]. The gp120 extracellular subunit of the viral envelope protein (Env) is responsible for binding CD4 on the surface of host T-cells to begin infection; this Env subunit is a common target for bnAbs [41]. Env fragments selected via computational optimization to potentially invoke the production of bnAbs are often employed in current work on vaccine production [14]. Studies using these Envs have varied from successful [4] to unsuccessful [26]. One potential explanation is that environmental impacts on gp120-CD4 interactions are not considered during Env selection. In particular, isolating bnAbs from a blood/plasma environment (slightly basic pH) might obfuscate the impact of mucosal environments (often acidic pH) on transmission. Therefore, it is reasonable to assume that both Env structure and binding affinity with CD4 and/or bnAbs will be altered under physiological conditions consistent with sexual transmission.

Recent experimental and computational studies have shown that pH impacts both Env conformation and CD4 binding. In 2013, Stieh et al. hypothesized that electrophoresis could be applied at a protein level and performed direct experimentation to reveal a pattern of change in surface charge across a physiological pH range. The findings produced a fingerprint of trimeric gp120 indicating a change in electrophoretic mobility from negative to positive as pH increased [38]. A computational pipeline was developed which produced results consistent with those of laboratory experiments: a difference was seen from negative to positive as pH levels increased in the physiological range consistent with observed increase in Env-CD4 binding affinity at acidic pH.

Morton et al. enhanced and refined Stieh et al. by incorporating protein modeling, parallel processing, structure energy minimization, and advanced floating point data compression that allowed for larger studies to be performed and a greater depth of analysis to take place [31]. The procedure also developed a classification method called Biomolecular Electro-Static Indexing (BESI) based on principal component analysis (PCA), cosine similarity analysis (CSA) and is loosely based on latent semantic indexing (LSI) [10].

Howton et al. independently introduced an extension to the prototype computation from Stieh et al. to the protein residue level [17]. Howton approached the problem with the hypothesis that strains in chronic infection, so called chronic control (CC) strains, will likely have adapted to systemic pH, and will be less efficient at binding CD4 under acidic conditions when compared to transmitted founder (TF) strains. Using computational modeling, some differences between TF and CC strains as well as between B and C clades were discovered using a more extensive set of 28 Env proteins [17]. However, the specific molecular mechanism (eg. surface residues and mutations) responsible for the pH sensitivity of the gp120-CD4 interaction could not be determined using the resulting data. Again, the simulations performed by Howton et al. were limited due to the amount of data being produced using the original methods developed by Stieh et al.

In this study we have incorporated the prototype developed by Howton et al. into the BESI pipeline to utilize state-of-the-art compression libraries and our machine learning technique for determining transmitted variants [31]. We expand the research across a broader solvation range and extend BESI to perform residue level analysis more thoroughly than was possible in Howton et al. Additionally, we introduce an imaging technique based on the variance of pH observed across all amino acids of the protein structure.

2 BACKGROUND

For each of 24 individuals, Howton et al. analyzed a TF/CC sequence pair. 18 pairs were drawn from clade B, and 6 pair from clade C. TF sequences were defined as those subspecies collected within the first 6 months of infection, and CC sequences were collected after the initial period prescribed for TF as designated by Parrish et al. [34]. B clade sequences were obtained from [3, 7, 9, 20, 25, 36, 39, 40] and C clade sequences were obtained from [1, 24, 29]. Additional sequence details including accession numbers and sequence alignments can be found in [16].

Howton et al. found that the pH sensitivity of gp120 charge density was mildly more sensitive in CC strains when compared to TF strains. One possible explanation for these results might be that the typical rules used to distinguish TF/CC are not sufficiently stringent since sequence diversity within the first 6 months of infection may be broad enough to dilute the responsible signal [6].

Given the progress in data compression and scoring of transmissible variants developed by Morton et al., revisiting the more diverse data set used in Howton et al. through the BESI framework would be advantageous. The refinements in BESI would produce clearer results and allow for a broader range of analysis to be performed in the simulations. In particular, BESI scores would now provide an objective, structure-based metric for determining highly transmissible variants in the data set. Such information would be informative for Env acquisition during vaccine development where selection of highly transmissible variants could potentially increase efficacy. Additionally, during this simulation an imaging process was developed that displays the pH active residues in a conserved and highly localized region of the gp120 envelope.

3 RESULTS

3.1 Electrophoretic Finger Printing

As the standard for validation of calculating the electrostatic charges on the surface of the assemblies we present this sample electrophoretic fingerprint as proof of the process and data retrieval methods. Lower values indicate more negative surface exposure for the bound conformation of gp120 relative to the unbound. The prevalence of more negative surface charge on the bound conformation of gp120 at low pH suggests that gp120 favors the bound conformation at these pH levels. Hence, gp120 is also primed for binding the positively charged CD4 target at low pH, consistent with previous findings. While some variation among the Env subunits exists, this example indicates the low pH trigger is a general property of the gp120 subunit.
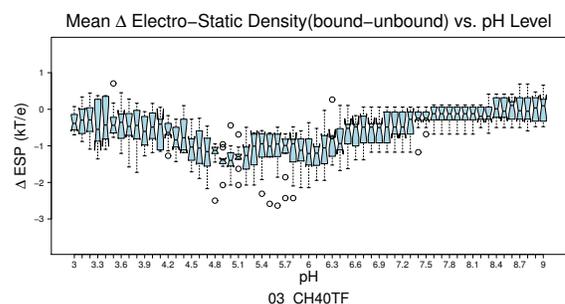
Figure 1: Fingerprint of sequence 0 3_CH40TF indicates all processing and data retrieval processes were executed correctly. Negative values indicate environmental pH where the Env prefers the bound conformation.

3.2 Full Structure Binding Energies

Howton et al. concluded that the differences in binding energies between clades and classes were minimal and therefore inconclusive. We provide this section as a confirmation of the process introduced by Howton et al. and to confirm their findings even through a more refined process. We express the similarities of findings using Figure 2. Observing the specified graph, one can determine variation between the clade and class combinations exists at the lower spectrum of pH. While differences in amplitude do exist, the graphs follow similar trajectories along the x-axis. The variation in amplitude may be explained by the number of variants observed from the two clades, where clade B is the dominant factor in the study by approximately 3:1. Classes are distributed evenly across the study and these facts invoke the need to observe a larger set of sequences to determine if differences exist at lower pH.

3.3 Biomolecular Electro-Static Indexing

The BESI analysis of the data is used to create a principal component subspace to compare and classify Env charge data presented by APBS [2]. With the exception of donor/recipient comparisons being made, a functional comparison was performed against the control subspecies of Morton et al. [31]. BESI made a limited selection of gp120 structures closely related to the control in terms of the
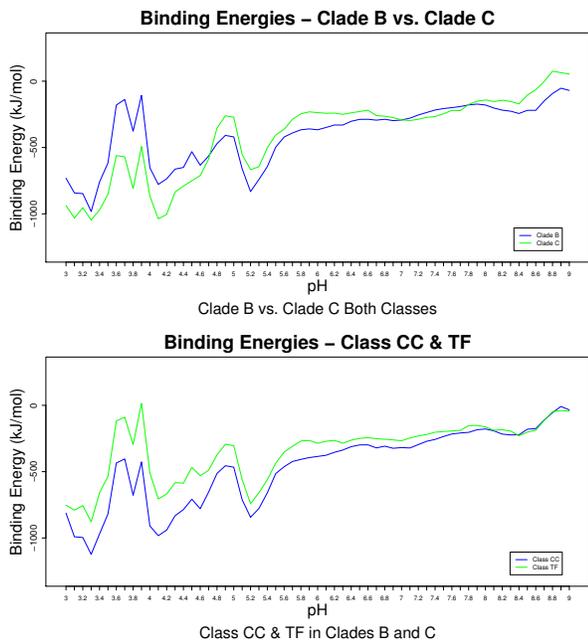
Figure 2: Binding energies of clades and classes. Clade B vs. Clade C (top) and class CC vs. class TF (bottom)

functional capabilities of the assemblies. We hypothesize that BESI is a unique method of Env classification that can be used to focus research efforts on a limited number of predicted active sequences versus blind trials. Referencing Figure 3, the overlay of scores in a color gradient indicate the scale of separation amongst subspecies functionally. The scores range from zero to one and are least to most similar to the control sequence respectively. Morton et al. noted that while the selected transmitted founder Env and the control sequence were different by a single amino acid, the TF score is 0.807, displaying the range of difference a single residue can introduce. The methods used to produce Figure 3 can be found under "Methods and Materials."

3.4 Electrostatic Variance Masking

The selection of amino acid assemblies across functional sections of a protein provides a unique view of the Env of what we hypothesize to be the mechanism involved in the binding process. The method described focuses on the surface charge variance of a residue position across all unbound gp120 assemblies. Performing the selection process prescribed by EVM (see Materials and Methods) produced a uniform selection of amino acids for each of the chosen sequences. Statistical information returned from this data set is as follows: Standard Deviation = 101.9, 1/2 Standard Deviation = 51.0, Variance cutoff selected = 51.0, % of variance selected = 75.7, Number of Selected Residues = 64.0, % of residues selected = 11.6.

Figure 4 displays the variance across all gp120 assemblies in this study. The reader should note that EVM involves the entire residues set in all Env subunits processed.

By completing the methods in 'Selection of Mechanistic Residues' as described, the sequence logo graph of Figure 5 is presented to
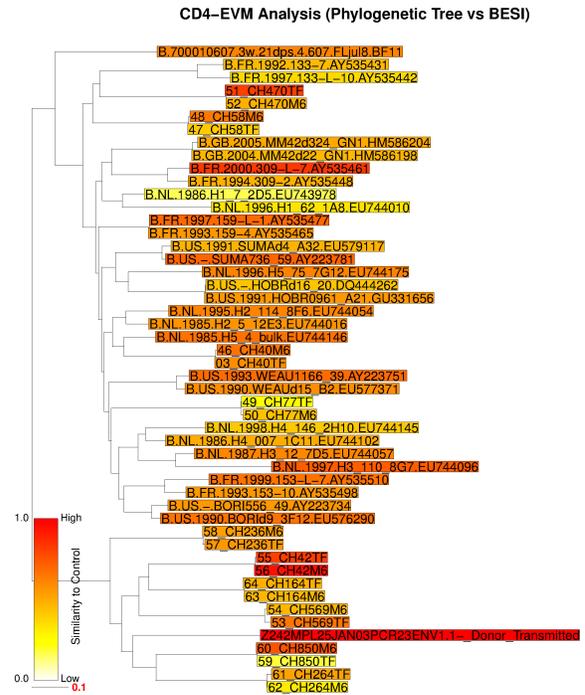
Figure 3: BESI score overlay via gradient onto phylogeny tree of sequences
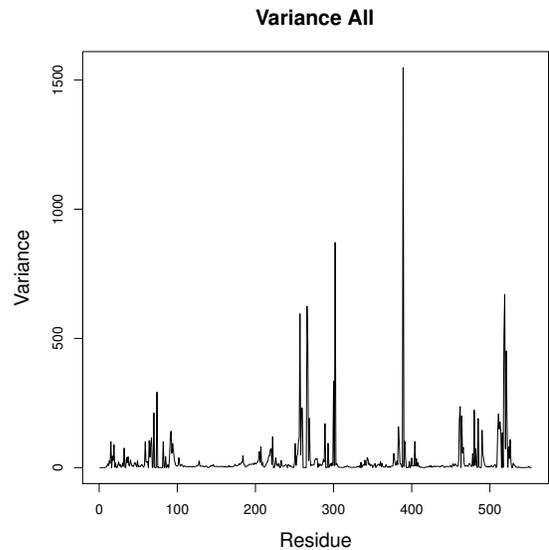
Figure 4: Graph of variances for each residue from all sequences.

display the conservation of amino acids EVM selects across the sequence set.

Additionally, we applied the following conditions to produce a sample set of structures and imagery that displays the power of the

Figure 5: Logos representation of the conservation of amino acids through EVM selection method

approach. We selected a set of sequences based on BESI score. We selected the highest BESI score '0.955' to acquire '56_CH42M6' and a low score of '0.340' to select 'B.NL.1996.H1_62_1A8.EU744010' which also contains the longest sequence chain in the set.

We then applied the following additional methods to the prescribed guidelines (see Materials and Methods) to invoke a percent variance selection process similar to those used in PCA. For the sequences to be imaged, the minimum possible variance value is determined to be greater than 51. This selects approximately 75% of the total variance and presents 64 residues to image. The results show that approximately 11% of the amino acids (based on all aligned residues) contribute just over 75% of the total variance across the entire sequence set. We present a graph of the total variance in Figure 6.
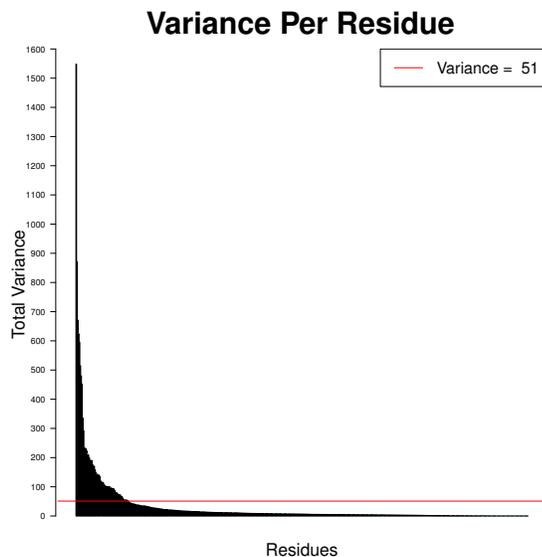
Figure 6: Graph of all residues in the set indicating the majority of observed variance exists in the least number of residues. Red horizontal line is the selected cutoff value.

The method produces the following actual residue selection lists (not HXB2):

- 56_CH42M6: length - 64
  14 18 31 58 63 65 66 69 73 81 90 91 92 93 94 170 172 184 185 187 216 219 220 221 222 224 225 231 232 234 254 258 265 267 333 339 340 345 347 360 394 395 397 398 399 411 413 415 418 423 424 440 441 442 443 444 446 448 449 450 451 452 454 456
- B.NL.1996.H1_62_1A8.EU744010: length - 64
  14 18 31 58 63 65 66 69 73 81 90 91 92 93 94 176 178 190 191

Figure 7: Binding site view of Env 56_CH42M6. The $\alpha2$ helix is clearly marked to confirm the binding site location.
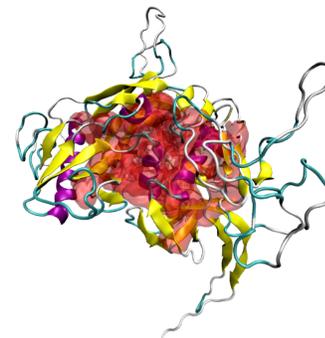
Figure 8: Binding site view of Env B.NL.1996.H1_62_1A8.EU744010. This protein is oriented identically as Figure 7

193 222 225 226 227 228 230 231 237 238 240 260 264 271 273 339 345 346 351 353 366 414 415 417 418 419 431 433 435 438 443 444 461 462 463 464 465 467 469 470 471 472 473 475 477

For each of the selected assemblies we loaded the first model of the unbound conformations into VMD. We produced an additional representation of the molecule and selected the residue set provided. We present the primary representation as a cartoon colored by secondary structure. For the residue selection a red, transparent surface is selected. The imagery of Figures 7 and 8 display the similarities across gp120 structures and expounds upon the conservation of residues at or near the CD4 binding site. These data suggest that mutation of residues outside the CD4 binding site is the primary mechanism for modulating the potential transmission rate of the virus since residues outside of the conserved binding site are primarily where sequence differences arise. Allosteric interactions must drive much of the process in this case, but BESI cannot determine this definitively since it only works on the static endpoints of the binding process. It might be possible to model these transitions using molecular dynamics or FRODAN, but the increase in APBS calculations needed to sample the intermediate states is still computationally prohibitive with current software. However, we hope to utilize MD simulations on a subset of these sequences in future work.

We then invoked MAFFT v7.273 using E-INS-i with an open gap penalty of 2.0 as described by Foley et al. in [15] and align each structure individually to HXB2CG. We convert the selections previously mentioned to the HXB2CG alignment as describe by Korber et al. in [23] to produce identical alignments for the two selected sequences. The HXB2CG residues selected are:

- 47 51 64 91 96 98 99 102 106 114 123 124 125 126 127 199 201 213 214 216 245 248 249 250 251 253 254 260 261 263 283 287 294 296 364 370 371 376 378 391 426 427 429 430 431 443 445 447 450 455 456 470 471 472 473 474 476 478 479 480 481 482 484 486

For all forty-eight (48) structures in this simulation, forty-one(41) presented identical HXB2CG selections while the remaining seven(7) structures varied by a single identical selection. The alternate list of selected residues with the difference in red bold-faced font are:

- 47 51 64 91 96 98 99 102 106 114 123 124 125 126 127 199 201 213 214 216 245 248 249 250 251 253 254 260 261 263 283 287 294 296 364 370 371 376 378 **396** 426 427 429 430 431 443 445 447 450 455 456 470 471 472 473 474 476 478 479 480 481 482 484 486

The seven(7) structures with the alternate selection were evenly distributed, to the extent possible, across TF/CC classes. Five(5) of these variants were of clade B the dominant subspecies of this study.

Most notably, EVM selected amino-acids 124-127, 283, 364, 370, 371, 426-431, 455, 456, 470-474, 476 which are CD4 contact residues. Other pertinent selections aligned to the following: Residues 64,and 91 are adjacent to interface contacts with gp41. Residue 123 is a co-receptor binding site outside of V3. Residues 199, 201,251 are co-receptor specific R5/X4 sites. Residues 261 and 263 are adjacent to glycosite 262. Residue 294 is adjacent to glycosite 295. Residue 296 is the start of V3 loop. Residue 391 is adjacent to glycosite 392. Residue 396 is at the V4 hyper-variable hot spot. Residue 447 is adjacent to glycosite 448. Descriptions per the 'HXB2 Annotated Spreadsheet' [21].

4 DISCUSSION

These results suggest that the TF/CC distinction is too non-specific to elucidate transmitted variants in a reliable way due to the rapid evolution of the viral pool shortly after transmission. This would also explain the mild differences between TF/CC classes and the minimal observation of mechanistically-informative signals in the previous study. Nevertheless, the pH-sensitivity of gp120 under acidic conditions was still observed across all TF/CC pairs.

The processes involved with BESI and EVM show the effectiveness of the pipeline for analyzing pH sensitivity for protein-protein interactions. The methods build upon previous approaches for computing electrostatic potential across a range of environmental conditions [30, 38] by integrating high-throughput structural modeling, conformational search, and targeted docking for a large set of sequences [11, 13, 35]. Electrostatic Variance Masking was able to correctly identify the gp120-CD4 binding interface even in the absence of CD4 surface data which has never been performed before to our knowledge. These results suggest that the technique might be useful in the future for identifying pH-sensitive binding sites in other protein-protein interactions.

Additionally, this work shows the importance of the pH sensitive binding mechanism and its potential role in the gp120-CD4 interaction. The binding site residues were correctly identified using the residue-specific EVM process suggesting that mutations in non-binding site residues drive binding sensitivity allosterically. This is particularly important for HIV vaccine research because the CD4 binding site is an important target for vaccine conception, and pH has been shown to affect antibody binding at pH levels associated with mucosa [12]. Additionally, bnAbs typically target the CD4 binding site of gp120 [41]. Investigations into the pH sensitivity of gp120-bnAb interactions using the BESI and EVM are currently ongoing.

5 MATERIALS AND METHODS

5.1 Env Sequence Data

The sequence set constitutes twenty-four (24) pairs of Env consisting of one (1) TF and one (1) CC gp120 structures from Clades B and C with eighteen (18) and six (6) pairs respectively. Accession numbers with clade/class designations can be found in [16]. B clade Env were acquired from [3, 7, 9, 20, 25, 36, 39, 40] and C clade sequences were obtained from [1, 24, 29].

5.2 Binding Energies

Full structure and residue binding energies (BE) were calculated using the procedures described in [16] with the following exceptions: 1) additional energy minimization steps via Gromacs [5, 27] and 2) compression techniques via ZFP [28] as described in [31] to allow for full solvation analysis of the protein structures.

5.3 BESI

Processes developed in [31] were executed for this study to generate the surface charge data and BESI scores. The phyologenetic tree is constructed as follows. Sequences were aligned with MAFFT v7.273 using the L-INS-i strategy [19]. A maximum likelihood (ML) phylogenetic tree was inferred using the RAxML software, version 8.2.11 [37] with the HIVW amino acid model of substitution [32] and 100 bootstrap replicates. Trees were midpoint-rooted and rendered using APE version 5.0 [33]. Expression of the phylogenetic tree involves minor differences where Donor/Recipient information is unavailable for the sequence data used for this study.

5.4 Residue Electrostatics

Residue electrostatic data was derived in the same manner as described in [17] with the following exceptions: 1) additional energy minimization steps via Gromacs [5, 27] and 2) compression techniques via ZFP [28] as described in [31] to allow for full solvation analysis of the protein structures.

5.5 Selection of Mechanistic Residues

Selection of residues that respond to pH shifts involves calculating the electrostatic charge variance of each residue across all aligned sequences. Where gaps are encountered in the alignment a value of zero (0) is assigned. For each residue, the median value of individual residues is calculated for each model in the pH range of 3.0 to 9.0 with 0.1 increments. The mean and variance of each position in

the global sequence alignment is then calculated and stored. This method allows us to effectively filter out residues which showed small variations in mean surface charge across the pH shift concurrent with relatively little impact on electrostatic binding energy.

For each sequence alignment a reverse mapping is created to align selections with correct residue numbers on the individual proteins. Where a gap exists in the alignment a value of negative one (-1) is assigned. This allows the determination of a cutoff value for the variance where a selection of a gap will show -1 and is considered an invalid selection. The selected residues are then applied to a VMD representation [18] to display the substructures involved. For this method of imaging residue structures participating in the mechanistic functions of the binding function we have termed the process Electrostatic Variance Masking (EVM).

REFERENCES

[1] M-R Abrahams and et al. 2009. Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. *Journal of Virology* 83, 8 (Apr 2009), 3556–67. https://doi.org/10.1128/JVI.02132-08

[2] N A Baker and et al. 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America* 98, 18 (2001), 10037–41. https://doi.org/10.1073/pnas.181342398

[3] Katharine J Bar and et al. 2012. Early low-titer neutralizing antibodies impede HIV-1 replication and select for virus escape. *PLoS Pathogens* 8, 5 (2012), e1002721. https://doi.org/10.1371/journal.ppat.1002721

[4] Dan H Barouch and et al. 2013. Protective efficacy of a global HIV-1 mosaic vaccine against heterologous SHIV challenges in rhesus monkeys. *Cell* 155, 3 (Oct 2013), 531–9. https://doi.org/10.1016/j.cell.2013.09.061

[5] H.J.C. Berendsen, D. van der Spoel, and R. van Drunen. 1995. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* 91, 1-3 (sep 1995), 43–56. https://doi.org/10.1016/0010-4655(95)00042-E

[6] D. I. Boeras and et al. 2011. Role of donor genital tract HIV-1 diversity in the transmission bottleneck. *Proceedings of the National Academy of Sciences* 108, 46 (2011), E1156–E1163. https://doi.org/10.1073/pnas.1103764108

[7] Evelien M Bunnik and et al. 2008. Autologous neutralizing humoral immunity and evolution of the viral envelope in the course of subtype B human immunodeficiency virus type 1 infection. *Journal of Virology* 82, 16 (Aug 2008), 7932–41. https://doi.org/10.1128/JVI.00757-08

[8] Dennis R Burton and et al. 2012. Broadly neutralizing antibodies present new prospects to counter highly antigenically diverse viruses. *Science (New York, N.Y.)* 337, 6091 (Jul 2012), 183–6. https://doi.org/10.1126/science.1225416

[9] Laurent Dacheux and et al. 2004. Evolutionary dynamics of the glycan shield of the human immunodeficiency virus envelope during natural infection and implications for exposure of the 2G12 epitope. *Journal of Virology* 78, 22 (Nov 2004), 12625–37. https://doi.org/10.1128/JVI.78.22.12625-12637.2004

[10] Scott Deerwester, Susan T. Dumais, and Richard Harshman. 1990. Indexing by Latent Symantic Analysis. *Journal of the American Society for Information Science and Technology* 41, 6 (1990), 391–407. http://www.psychology.uwo.ca/faculty/harshman/latentsa.pdf

[11] Narayanan Eswar and et al. 2002. *Comparative protein structure modeling using modeller.* John Wiley & Sons, Inc. https://doi.org/10.1002/0471250953.bi0506s15

[12] Kelly M. Fahrbach, Olga Malykhina, Daniel J. Stieh, and Thomas J. Hope. 2013. Differential Binding of IgG and IgA to Mucus of the Female Reproductive Tract. *PLOS ONE* 8, 10 (10 2013), 1–11. https://doi.org/10.1371/journal.pone.0076176

[13] Daniel W. Farrell, Kirill Speranskiy, and M. F. Thorpe. 2010. Generating stereochemically acceptable protein pathways. *Proteins: Structure, Function and Bioinformatics* 78, 14 (2010), 2908–2921. https://doi.org/10.1002/prot.22810

[14] Will Fischer and et al. 2007. Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nature Medicine* 13, 1 (Jan 2007), 100–6. https://doi.org/10.1038/nm1461

[15] Brian Thomas Foley and et al. 2015. HIV Sequence Compendium 2015. (10 2015). https://doi.org/10.2172/1222684

[16] Jonathan Howton. 2017. *A Computational Electrostratic Modeling Pipeline for Comparing pH-dependent gp120-CD4 Interactions in Founder and Chronic HIV Strains.* Master's thesis. Middle Tennessee State University, Murfreesboro, TN. http://jewlscholar.mtsu.edu/xmlui/handle/mtsu/5324

[17] Jonathan Howton and Joshua L. Phillips. 2017. Computational Modeling of pH-dependent gp120-CD4 Interactions in Founder and Chronic HIV Strains. In

[18] *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics - ACM-BCB '17.* ACM Press, Boston, MA, USA, 644–649. https://doi.org/10.1145/3107411.3107506

[18] William Humphrey, Andrew Dalke, and Klaus Schulten. 1996. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* 14 (1996), 33–38.

[19] Kazutaka Katoh and Daron M. Standley. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30, 4 (2013), 772–780. https://doi.org/10.1093/molbev/mst010

[20] Brandon F Keele and et al. 2008. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences of the United States of America* 105, 21 (2008), 7552–7. https://doi.org/10.1073/pnas.0802203105

[21] Korber and et al. 2017. HXB2 Annotated Spreadsheet. Retrieved June 24,2018 from https://www.hiv.lanl.gov/content/sequence/HIV/MAP/hxb2.xls

[22] Bette Korber and S. Gnanakaran. 2011. Converging on an HIV Vaccine. *Science* 333, 6049 (2011), 1589–1590. https://doi.org/10.1126/science.1211919

[23] B Korber-Irrgang and et al. 1998. Numbering positions in HIV relative to HXB2CG. https://www.scienceopen.com/document?vid=2661d7cb-fb50-4a20-9044-71a8501579f3

[24] Denise L Kothe and et al. 2006. Ancestral and consensus envelope immunogens for HIV-1 subtype C. *Virology* 352, 2 (Sep 2006), 438–49. https://doi.org/10.1016/j.virol.2006.05.011

[25] Bing Li and et al. 2006. Evidence for potent autologous neutralizing antibody titers and compact envelopes in early infection with subtype C human immunodeficiency virus type 1. *Journal of Virology* 80, 11 (Jun 2006), 5211–8. https://doi.org/10.1128/JVI.00201-06

[26] Hua-Xin Liao and et al. 2013. Vaccine induction of antibodies against a structurally heterogeneous site of immune pressure within HIV-1 envelope protein variable regions 1 and 2. *Immunity* 38, 1 (Jan 2013), 176–86. https://doi.org/10.1016/j.immuni.2012.11.011

[27] Erik Lindahl, Berk Hess, and David van der Spoel. 2001. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Journal of Molecular Modeling* 7, 8 (Aug 2001), 306âĂŞ317. https://doi.org/10.1007/s008940100045

[28] P. Lindstrom. 2014. Fixed-Rate Compressed Floating-Point Arrays. *IEEE Trans. on Visualzation and Computer Graphics* 20, 12 (2014), 2674–2683. https://doi.org/10.1109/TVCG.2014.2346458

[29] Michael K P Liu and et al. 2013. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *The Journal of clinical investigation* 123, 1 (jan 2013), 380–93. https://doi.org/10.1172/JCI65330

[30] Aaron C. Mason and Jan H. Jensen. 2008. Protein-protein binding is often associated with changes in protonation state. *Proteins: Structure, Function and Genetics* 71, 1 (2008), 81–91. https://doi.org/10.1002/prot.21657

[31] Scott P. Morton, Julie B. Phillips, and Joshua L. Phillips. 2017. High-Throughput Structural Modeling of the HIV Transmission Bottleneck. In *Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine - BIBM-HPCB '17.* IEEE Press, Kansas City, MO, USA.

[32] David C. Nickle and et al. 2007. HIV-Specific Probabilistic Models of Protein Evolution. *PLoS ONE* 2, 6 (2007). https://doi.org/10.1371/journal.pone.0000503

[33] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 2 (2004), 289–290. https://doi.org/10.1093/bioinformatics/btg412

[34] Nicholas F Parrish and et. al. 2013. Phenotypic properties of transmitted founder HIV-1. *Proceedings of the National Academy of Sciences of the United States of America* 110, 17 (2013), 6626–33. https://doi.org/10.1073/pnas.1304288110

[35] Joshua L. Phillips and S. Gnanakaran. 2015. A data-driven approach to modeling the tripartite structure of multidrug resistance efflux pumps. *Proteins: Structure, Function and Bioinformatics* 83, 1 (2015), 46–65. https://doi.org/10.1002/prot.24632

[36] Jesus F Salazar-Gonzalez and et al. 2009. Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *The Journal of Experimental Medicine* 206, 6 (Jun 2009), 1273–89. https://doi.org/10.1084/jem.20090378

[37] Alexandros Stamatakis. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 9 (2014), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033 arXiv:10.1093/bioinformatics/btu033

[38] Daniel J. Stieh and et al. 2013. Dynamic electrophoretic fingerprinting of the HIV-1 envelope glycoprotein. *Retrovirology* 10, 1 (2013), 33. https://doi.org/10.1186/1742-4690-10-33

[39] Emma L Turnbull and et al. 2009. Kinetics of expansion of epitope-specific T cell responses during primary HIV-1 infection. *Journal of Immunology (Baltimore, Md. : 1950)* 182, 11 (Jun 2009), 7131–45. https://doi.org/10.4049/jimmunol.0803658

[40] Xiping Wei and et al. 2003. Antibody neutralization and escape by HIV-1. *Nature* 422, 6929 (Mar 2003), 307–12. https://doi.org/10.1038/nature01470

[41] R Wyatt and et al. 1998. The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature* 393, 6686 (Jun 1998), 705–11. https://doi.org/10.1038/31514