Molecular Basis of pH-dependent HIV gp120 Differences Revealed Using BESI

Scott P. Morton[1], Julie B. Phillips[2] and Joshua L. Phillips[1]

[1]Middle Tennessee State University, Murfreesboro, Tn. And [2]Cumberland University, Lebanon, Tn.

Middle Tennessee State University
College of Basic and Applied Sciences

Department of Computer Science &
The Computational Sciences Program

Cumberland University
Department of Biology

Abstract

Research in the field of HIV transmission has yet to provide a vaccine for the imponderable virus. Though progress has been made to extend the life of those chronically infected, a solution to the transmission of the disease outside of abstinence is still no where to be found. Previous laboratory studies involving electrostatic surface charge analysis revealed the sensitivity of gp120 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 laboratory results. We previously refined the process and utilized additional methods to determine a system capable of classifying Env structures through machine learning techniques. We have expounded upon the analytical procedure to encompass the residue level and expanded the process to include minimization steps to ensure the integrity of the protein structures. 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 continue to validate previous work in several studies as well as increase the returned knowledge through a new technique that reveals what we hypothesize to be the mechanistic residues responsible for the binding process.

Introduction

Boeras et al provided a unique set of data with more than nine hundred HIV RNA sequences drawn from twenty individuals from Rwanda and Zambia from both clades A1 and C [1]. The samples were classified and reduced to provided 252 gp120 protein assemblies in total. More importantly, subjects were grouped into donor/recipient pairs (10 pairs total). Each pair consisted of two individuals of which one was known to be infected (donors) and the other was expected to acquire infection at some point (recipients). Samples were taken prior to communication of the disease and after infection of the recipient occurred.

This research provided unique insight to the transmission process of HIV and served as the foundation for the inception of Bio-molecular Electro-Static Indexing.

Background

Bio-molecular Electro-Static Indexing (BESI)

Previously, we developed a machine learning method of classification to determine what we hypothesize to be subspecies of HIV that possess characteristics that increase the likelihood of transmission [2]. More precisely, the ability of a particular variant of gp120 that would be more successful at bind CD4. The method applies a score to each structure based on a cosine similarity analysis of the first two principal components produced by a principal component analysis. This method is based on latent semantic indexing. The scores are then presented as an overlay via color gradient applied to a phylogeny tree to provide a visual representation of the process compared to an evolutionary depiction of a set of gp120 envelopes.

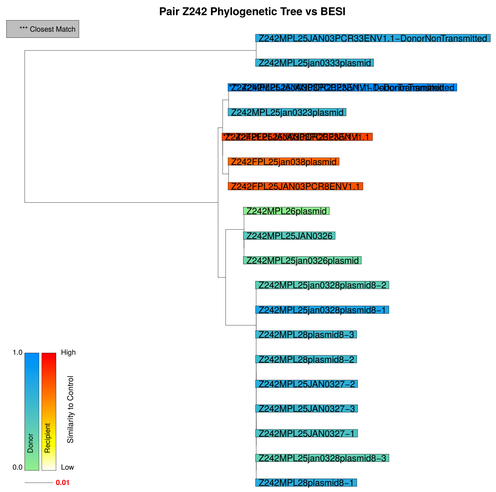

Figure 1 Boeras et al [1] provided the control sequence used for BESI. This figure depicts the phylogeny tree with BESI score overlays showing the selection accuracy of the process. The closest match sequences differ by a single amino acid from donor to recipient.

Figure 1 has two gradients applied, one for donor and one recipient, with darker shades indicating an increase in BESI score. For clarity, the closest matching sequences are marked with three asterisks.

Purpose

Research has shown that the highest populations of HIV subspecies are not the variants that transmit from host to host [1]. The overall goal of this research is to predict which Env variants are most likely to bind to a given CD4 receptor. This information allows researchers to focus on the transmission of the virus in a preventative manner on a focused set of subspecies where previous research validated the concept of unique structures crossing the transmission boundary [1].

Electrostatic Variance Masking

The process developed by Stieh [3] was extended to the residue level by Howton et al [4] to attempt to describe differences in amino acids in terms of total binding energy. Although the results were inconclusive, the method paved the way to discover a mathematical presence in the data referencing total variance. By this we are describing the variance of each amino acid across the pH range associated with homo-sapiens. Applying this concept across a processed set of aligned sequences and combining the data we present a sorted view in Figure 2.

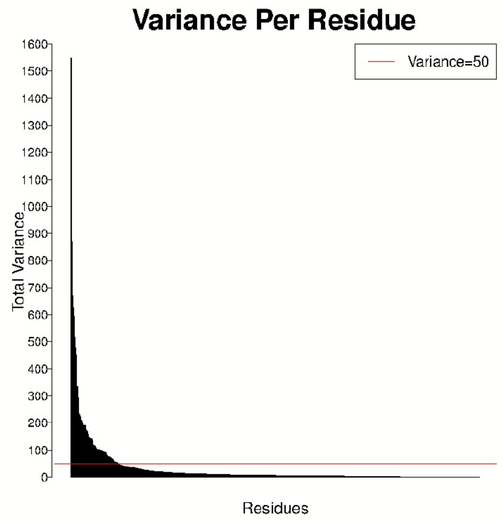

Figure 2 Showing the variance of the sequences processed. With a determined cutoff point based on the least possible value that does not select an invalid residue across the chosen sequences used for imagery.

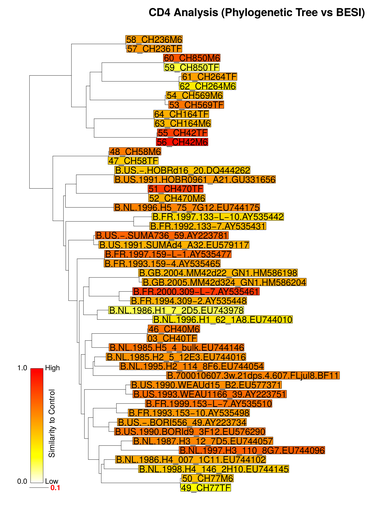

Figure 3 BESI overlay of sequence set provided by Howton. With no donor - recipient information, a single gradient represents the scoring.

Sequence Selection

For the sequences provided by Howton [4] , we applied BESI using the control provided by Morton [2] to produce Figure 3. The sequences from Howton were across Clades B and C consisting of 24 pairs of proteins of which one is a transmitted founder (TF) and one a chronic control (CC) strain. The control sequence from Morton is a member of Clade C. We applied the following conditions to produce a sample set of structures and imagery that displays the power of the system. We selected a set of sequences based on BESI scores choosing the highest BESI score ’0.955’ to acquire ’56_CH42M6’, a mid score of ’0.849’ presents ’B.FR.2000.309-L-7.AY535461’ and a low score of ’0.340’ selects ’B.NL.1996.H1_62_1A8.EU744010.’ The last selection also contains the longest sequence chain in the set. No other considerations were made for the selection of sequences used for imagery.

Conservation of Residues

Mapping the high variance residues back into sequence terms and applying the selections to a weblogo [5,6] representation produces Figure 4 displaying the conservation of amino acids through the EVM process.

Figure 4 Logos representation of the conservation of amino acids through the EVM selection process.

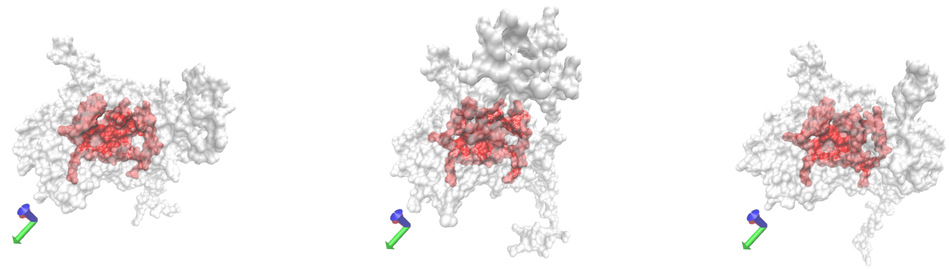

Figure 5 Binding site view of EVM imagery for sequences 56_CH42M6, BFR2000309-L-7AY535461, and BNL1996H1621A8EU744010 respectively.

Methods

Selection of Mechanistic Residues

Selection of residues which showed mechanistic response to pH shifts involves calculating the electrostatic charge variance of each residue across all aligned sequences vertically. Where gaps are encountered in the alignment a value of zero (0) is assigned. For each residue, the median value of individual residues for each model at a specific pH is taken to create a 1 x 61 vector for the pH range of 3.0 to 9.0 in 0.1 increments. The vectors are stacked row-by-row to create an array of dimension M x 61, where M is the number of sequences involved in the study. The mean value of each column is then calculated to produce a vector for which the variance is determined and stored. This is repeated for each alignment position. 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 [7] 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).

Generating Imagery

For each of the selected assemblies we loaded the first model of the unbound conformations related to each selected sequence into VMD [7]. We produced an additional representation of the molecule and selected the residues selected by EVM. We present the primary representation as a surface colored by charge and set the material to transparent for a uniform appearance. For the EVM residue selections we set the surface as opaque with red as the chosen color id producing the images in Figure 5.

Phylogeny Trees

Phylogeny trees were constructed as follows. Sequences were separated by subject, and aligned with MAFFT v7.222 using the L-INS-i strategy [8]. A maximum likelihood (ML) phylogenetic tree was constructed using the RAxML software, version 8.2.11 [9] with the HIVW amino acid model of substitution [10] and 100 bootstrap replicates. Trees were midpoint-rooted using the phylogenetic visualization software FigTree, version 1.4.3 [11]. Phylogeny tree rendered and gradient applied using APE via R [12].

Conclusion

EVM displays a powerful means of investigation of differences across sequences of various Clades and classes of HIV in conjunction with BESI. The data suggests that the highly conserved and localized high variance residues are not the source of selectivity for the proteins ability to bind to CD4. This data implicates the structural mutations surrounding the binding site as the distinctive variation in performance.

Acknowledgments

I would like to personally thank Dr. Phillips for his guidance and open-mined approach to alternate ideas and ways of thinking. Furthermore, I would like to acknowledge the entire Computer Sciences Department and all the professors I have had the privilege of studying under at MTSU for providing a great opportunity to learn and grow as a human being. I would also like to thank Dr. Julie B. Phillips for the fantastic phylogeny trees and the knowledge to reproduce the structures on my own.

References

1. Boeras, D. I., Hraber, P. T., Hurlston, M., Evans-Strickfaden, T., Bhattacharya, T., Giorgi, E. E., … Hunter, E. (2011). Role of donor genital tract HIV-1 diversity in the transmission bottleneck. Proceedings of the National Academy of Sciences, 108(46), E1156–E1163.
2. Morton, S.P.; Phillips, J.B.; Phillips, J.L. High-Throughput Structural Modeling of the HIV Transmission Bottleneck. Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine - BIBM-HPCB ’17; IEEE Press: Kansas City, MO, USA, 2017.
3. Stieh, D.J.; Phillips, J.L.; Rogers, P.M.; King, D.F.; Cianci, G.C.; Jeffs, S.A.; Gnanakaran, S.; Shattock, R.J. Dynamic electrophoretic fingerprinting of the HIV-1 envelope glycoprotein. Retrovirology 2013, 10, 33.
4. Howton, J.; Phillips, J.L. Computational Modeling of pH-dependent gp120-CD4 Interactions in Founder and Chronic HIV Strains. Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics - ACM-BCB ’17; ACM Press: Boston, MA, USA, 2017; pp. 644–649.
5. Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, Genome Research, 14:1188-1190, (2004) Full Text
6. Schneider TD, Stephens RM. 1990. Sequence Logos: A New Way to Display Consensus Sequences. Nucleic Acids Res. 18:6097-6100
7. Acad. Sci. USA 98, 10037-10041 2001. Humphrey W, Dalke A, Schulten K: VMD: Visual molecular dynamics. J Mol Graph Model 1996, 14(1):3338.
8. Taylor & Francis. pp. 6187. ISBN 978-0-8058-6237-9.] Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution, 30(4), 772780. http://doi.org/10.1093/molbev/mst010
9. Stamatakis, A. (2014)., RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England), 30(9), 13121313. http://doi.org/10.1093/bioinformatics/btu033
10. Nickle, D. C., Heath, L., Jensen, M. A., Gilbert, P. B., Mullins, J. I., & Kosakovsky Pond, S. L. (2007)., HIV-specific probabilistic models of protein evolution. PloS One, 2(6), e503. http://doi.org/10.1371/journal.pone.0000503
11. Rambaut, A., FigTree v.1.43. http://tree.bio.ed.ac.uk/software/figtree/
12. Paradis E., Claude J. & Strimmer K. 2004., APE: analyses of phylo-genetics and evolution in R language. Version 4.1. Bioinformatics 20:289-290.