High-Throughput Structural Modeling of the HIV Transmission Bottleneck

Scott P. Morton
spm3c@mtmail.mtsu.edu
Center for Computational Science
College of Basic and Applied Sciences
Middle Tennessee State University
Murfreesboro, TN 37130, USA

Julie B. Phillips
jphillips@cumberland.edu
Department of Biology
Cumberland University
Lebanon, TN 37087, USA

Joshua L. Phillips*
Joshua.Phillips@mtsu.edu
Department of Computer Science
College of Basic and Applied Sciences
Middle Tennessee State University
Murfreesboro, TN 37130, USA

***Abstract*—After three decades of research on human immunodeficiency virus (HIV), the causative agent of acquired immunodeficiency syndrome (AIDS), a vaccine has yet to be discovered. Most theoretical and experimental work on HIV vaccines has focused on the relevant molecular interactions at systemic pH levels, but HIV is typically transmitted sexually at mucosal pH levels. We previously developed a computational approach for calculating pH-sensitivity which predicted optimal transmission at mucosal pH levels, and was validated by experimental electrophoretic measurements and envelope protein binding assays. We have recently augmented this approach using a unique combination of protein dynamical modeling, parallel computation, and data compression tools which enable high-throughput calculations. The resulting fully-automated pipeline was capable of predicting pH sensitivity for a recent study involving more than 250 unique HIV envelope proteins utilizing approximately 1 million individual electrostatic surface calculations. We provide strong evidence that supports the previous hypothesis of a computational approach to determining the pH sensitivity of HIV envelopes. Furthermore, a PCA-based indexing method is proposed that allows for a comparison of biomolecular structures in terms of electrostatic pH sensitivity. We utilize the results to predict highly transmissible HIV variants with implications for vaccine design and efficacy.**

I. Introduction

Over thirty years has passed since the discovery of Acquired Immune Deficiency Syndrome (AIDS) and a vaccine has yet to be developed for the Human Immunodeficiency Virus (HIV) that causes the disease. The challenge that researchers face is the overwhelming mutation rate of the virus due to host immune system pressure once introduced to the body.

HIV is typically transmitted during sexual intercourse in an acidic mucosa pool. Since protein assemblies and their ability to interact with other proteins are affected by pH, we focus our attention on this principal component. HIV transmission occurs when the gp120 portion of the viral envelope protein (Env), attached to the outer surface of the virus, makes contact with CD4 protein receptors at the target host cell periphery. The interaction between the two structures initiates binding and subsequent cellular infection.

Boeras et al. concluded that the highest populations of HIV subspecies are not the variants that transmit from host to host [2]. Their determinations were backed by statistical analysis of population subspecies and transmission data through direct investigation of human volunteer donors. With the large pool of subspecies extracted, and the capture of variants at the time of transmission, this data set presents a potential to determine differences in protein structure that may explain the transmission bottleneck.

II. Background

A. Dynamic Electrophoretic Fingerprinting

Electrophoretic mobility (EM) is an experimental measure of protein surface charge used to characterize and separate micro-organisms [3,4]. Stieh et al. hypothesized the method could be applied across saline and pH ranges relevant to mucosal environments where transmission is common and results in systemic infection. The study was performed on trimeric gp120/gp41 Env from clade B HIV-1 strain BX08 [1]. The results described surface charge variations across the titration indicating decreased Env surface charge in mucosal environments, complementing the postive surface charge of the CD4 receptor surface. This potentially could be caused by variations of the gp120 protein structure and the interactions of the surrounding solvent where blood plasma and mucous vary in pH and saline levels.

B. A Computational Approach for Calculating pH-Sensitivity

Stieh, et al. hypothesized that HIV binding rates are influenced by pH and are greater in the acidic conditions present in genital mucous [1]. A method was produced to calculate the pH sensitivity of gp120 envelope crystal structures computationally by iterating through a range of pH values while converting from the protein data bank (PDB) format to the protein charge radii (PQR) format via PDB2PQR [5,6]. All titration states (pKa values, also referred to as protonation or acidic strength) were determined using PROPKA 3.0 [7] during this process and the AMBER 99 force-field [8] was used to produce the atomic radii and partial charges. Grid dimensions were determined using the psize.py script available with APBS. The Adaptive Poisson-Boltzmann Solver (APBS) [10] was then invoked for the nolinear solver using temperatures of 310K with default parameters. To determine the moleculear solvent accessible surface (SAS), the measure function of VMD [11] is employed at 0.14nm radius. To

TABLE I

LIST OF DONORS. SUBJECT INDICATES COUNTRY OF ORIGIN, COUPLE IDENTIFIER AND GENDER RESPECTIVELY. D/R INDICATES THE SUBJECTS STATUS AS THE DONOR AND COMMUNICATION RECIPIENT, RESPECTIVELY. TOTAL IS THE NUMBER OF VARIANTS PROVIDED. * INDICATES THE SUBJECT PAIR IS NOT MENTIONED IN THE BOERAS ET AL. STUDY.

Subject	D/R	Total	Subject	D/R	Total
R56F	R	4	R56M	D	13
Z153F*	D	11	Z153M*	R	10
Z185F*	R	10	Z185M*	D	10
Z201F	D	42	Z201M	R	14
Z205F*	R	5	Z205M*	D	7
Z216F	D	24	Z216M	R	1
Z221F	D	26	Z221M	R	10
Z238F	D	20	Z238M	R	2
Z242F	R	3	Z242M	D	16
Z292F	D	18	Z292M	R	6

calculate the mean electrostatic surface potential (ESP) a 3-dimensional convolution process is executed across the SAS, summed and divided by the total surface area to produce the final result.

The resulting ESP data agreed with assayed binding rates and total bound protein measurements, suggesting CD4-complementary EM in physiological environments at mucosal pH levels strongly impacted Env-CD4 binding [1]. So called, Dynamic Electrophoretic Fingerprinting (DEF) of HIV envelopes is an unique application of EM for characterizing HIV Env proteins and whole virions [1]. However, this research was performed on a limited set of Env subspecies and needs to be expanded upon with a larger set of HIV Env proteins to further test the hypothesis.

C. Target Data

From a pool of more than nine hundred HIV RNA sequences, Boeras et al. provided 252 gp120 protein assemblies drawn from twenty individuals from Rwanda and Zambia. The structures are in the A1 and C clade domains of HIV to provide a broad range of comparison opportunities. The donors consisted of couples of which one was known to be infected and the other was expected to acquire infection at some point. Samples were taken prior to communication of the disease and after infection of the recipient occurred. The naming conventions used for the sequences indicate the country of origin, the gender, a subject pair identifier, and a donor (D) / recipient (R) indicator as shown in Table I.

III. METHODS

The process employed by Steih et al. is enhanced to perform larger studies in a high-throughput, totally automated, and phased approach. Utilizing basic system calls to execute required third party software, the pipeline process executes across any number of compute nodes in a producer-consumer model. Here we provide an overview of the process enhancements and the software employed.

A. Structure Modeling

The methods used by Steih et al. are based on crystal structures to perform the analysis. This method has been extended to include full structure modeling using Modeller [12]. Modeller constructs a gp120 monomer based on a set of gp120 core and gp120 fragment proteins from the Protein Data Bank that are employed as templates. The template sequence codes used are 1G9M, 1RZK, 2B4C, 2BF1, 2NY7, 3JWD, 3JWO, and 3LQA. Complete sequence data is consumed by Modeller in a plain text file that is similar to the FASTA file layout [13]. The results are returned in PDB [14] format for the next phase of the operation.

B. Stereo-Chemically Acceptable Conformations

The protein models are then shifted into bound and unbound conformations via FrodaN [15] to maintain a stereo-chemically acceptable state. FrodaN was configured to perform targeted geometrical simulations toward target conformations while keeping all stereo chemical constraints fixed.

Target states are represented by 1RZK in respective conformations. The gp120 structure is only available bound to a CD4 protein (1RZK) or antibody structure (2NY7). 2B1F is the only available putative unbound gp120, at the time of this writing, and is from the Simian Immunodeficiency Virus (SIV) gp120 core [16]. By utilizing 2B1F as a target for 1RZK to be manipulated from the bound state to the unbound state using FrodaN, we are able to provide a consistent bound and unbound target set for all models in the study. Figure 1 shows examples of a model structure (A) shifted into the bound (B) and unbound (C) conformations.

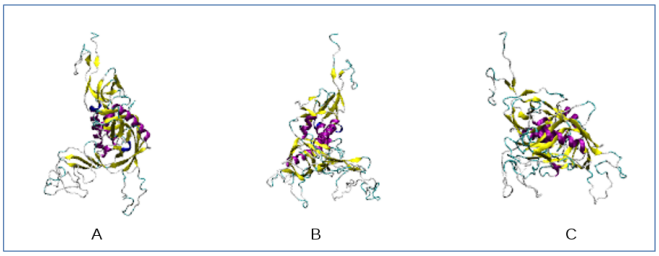

Fig. 1. Model representations of a single gp120 envelope in various conformations. Modeller creates a base structure (A) that is guided into stereo-chemically feasible conformations representing bound (B) and unbound (C) states.

C. Energy Minimization

To ensure a stable structure pre- and post-manipulation by FrodaN, energy minimization was performed using Gromacs [17,18]. This process relaxes the structure and helps to ensure a stable assembly throughout the process pipeline. We selected the conjugate gradient algorithm as the integrator and limited the procedure to 100,000 steps using the Amber99SB-ILDN force field [9]. Other force fields are not considered since only minimization is performed.

D. Electro-Static Surface Charge

To calculate the surface potential of each structure, we first convert from PDB format by invoking PDB2PQR [6]. The standard AMBER forcefield provided by APBS 1.4 was used, PROPKA was the pH calculation method, and each

value of the pH range is iterated to create 61 PQR files for each sequence and conformation. The pipeline then executes psize.py [6] against each PQR to determine grid points, center of mass, fine and coarse mesh lengths. Using the measure function of VMD, the solvent accessible surface (SAS) is determined for the polymer. Screened environmental charges for the molecule are calculated using APBS. At this juncture of the pipeline, a convolution process is performed to determine the surface potential of the gp120 envelope. For each point on the surface of the system, the sum of the surrounding points are added to the charge and averaged.

E. Data Compression

The initial estimates of the total data to be produced during the study of 252 sequences was estimated at approximately 130TB. This is an enormous amount of information to store and handle just to extract 75MB of results for analysis. The largest producer of data is APBS, from which all charge data is stored in DX format which is textual based consisting of descriptive and numeric content. Basic methods of encapsulation (eg. GZIP) typically achieve 2:1 compression ratios. This level of compaction is easily achievable by storing binary array data directly versus native methods provided by APBS. At 64bit precision, no information is lost up to machine epsilon. In either case, ratios of 2:1 are entirely inadequate for large scale analysis of structures in the manner presented.

To overcome this limitation we utilize recently developed methods for compressing floating point data at impressive ratios. ZFP [19] works exclusively with radix based exponential data by ingesting binary arrays and compressing them through signal processing methods. A typical operation in our study produced compression ratios of 75:1, a maximum error of 0.016 kT/e with a peak signal to noise ratio of 113:1. This compression method reduced our overall data storage requirements down to an easily manageable size that preserves the work for future analysis.

F. Parallel Processing

The original process was developed using bash shell scripts and pseudo-multiprocessing techniques, suitable for prototyping the process. In order to evaluate larger sets of sequences, the process was translated into Python [20] to create a completely automated system. Once the process was validated as functional and complete, limitations were evaluated for alternate execution methods. FrodaN and Gromacs presented specific challenges in regards to threading and/or multiprocess execution in single user space. MPI overcomes this issue and allows for the simultaneous execution of processes across multiple systems. An MPI driver for the pipeline was created using Python and MPI4PY [21] software. The method also employs a unique recovery model where each process is handed an 64bit integer as an index of the work to be performed. The driver then extracts a work unit, w, from the index through mathematical techniques to ensure that all processes for a particular sequence structure are completed in order. For example, if we have s sequences, n models per sequence, 2 conformations per model, and p pH levels per state, then the total number of events is simply the product of all terms, $s * n * 2 * p$. Keeping in mind the process that is executed, simply taking the modulus of the total by the desired task is a valid solution. However, the work is performed in a scattered fashion across the set in the later stages as the sequence, model, state and pH solution has to be determined at the time of execution. To achieve an ordered process for APBS, where sequence 0 is operated on until completion, the following calculation is employed:

$$sequenceID = (w/(n * p * 2))\%s$$

Similar calculations are invoked for determining other factors of the work unit. Take note that all terms of the calculation are integers and therefore result in an integer operation. Upon any unrecoverable failure, the work unit is written out to log files. By using the method described, a failure on a sequence would create a series of indexes allowing for that range to be fed back into a slightly modified version of the driver. This establishes a unique and simple recovery model in the event that a subsection of the study needs to be reprocessed due to programming errors, or restarted because of hardware failures.

G. Resulting Data

Sequences from ten transmission pairs from the Boeras et al. data set were provided for this study. The envelopes consist of HIV clade domains A1 and C. Samples were taken from blood plasma, peripheral blood mononuclear cells, genital tracts, swab-associated, cell-associated and cell-free methods for a genetically diverse set [2]. The pipeline resulted in the production of $7,560$ protein structures using Modeller, with conformations of those assemblies in bound and unbound states being produced by FrodaN consisting of $15,120$ new structures. Each envelope is then prepared in 61 different pH solvents ranging from $3 - 9$ in 0.1 increments using PDB2PQR to produce $922,320$ solutions. All systems are then calculated for electrostatic surface charges using APBS. The final size of data is approximately 6.3TB utilizing ZFP and other compression techniques previously described. The entire process was completed in approximately 60 days utilizing a computing cluster resource composed of 4 rack mounted DELL R815 servers. Each server houses 4 x 16-core 2.3 GHz AMD Opteron processors (64 cores per machine, 256 total), 512MB of RAM (2TB total), 7.2 TB of workspace hard disk (28.8 TB total).

H. Electrostatic Fingerprint Indexing

We utilize component data of Principal Component Analysis (PCA) and Cosine Similarity (CS) to establish a manner of identification for functionally similar envelopes. To our knowledge the methods described here have never been performed on protein substructures of a virion periphery.

We utilize the rotation data (eigenvectors) produced by PCA, a common method of dimensionality reduction [23,24] used in a wide range of fields. Methods of use include

exploratory analysis and predictive modeling where high dimensional multivariate datasets can be presented using reduced dimensionality better suited for visualization.

The method utilizes CS analysis as a means of comparing vectors on a Cartesian plane, where the cosine of the angle between the two vectors is an indicator of the similarity between them, i.e. $cos(0) = 1$ indicates the vectors are on the same line. This holds true for $cos(180) = -1$ where the direction of the ray is reversed: the line on which the vectors exist is still identical. The calculation is:

$$cos(\theta) = \frac{a \cdot b}{||a||_2||b||_2} = \frac{\sum_{i=1}^{x} a_i b_i}{\sqrt{\sum_{i=1}^{x} a_i^2}\sqrt{\sum_{i=1}^{x} b_i^2}}$$

where a_i and b_i are vector components of the n^{th} PC of the target and the control sequences respectively.

The combination of the two methods of analysis is the basis of Latent Semantic Indexing (LSI) [25]. LSI is a method of retrieval that uses PCA to identify patterns between terms and concepts in unstructured textual data. This process involves scoring paragraphs of unordered text based on word content using principal component analysis to generate eigenvectors representing each paragraph. The method then compares a query target to the eigenvectors of the unstructured text to identify similarities of the query versus the text by means of CS. The application of LSI in this study initially utilizes the first PC of each representative PCA object as the target query component and the first PC of the control object is then the source query term. The analysis of the data utilizes LSI in a unique manner to predict the likely Env to transgress the transmission barrier from donor to recipient.

We term the combination of these three approaches, for this specific application, Biomolecular Electro-Static Indexing (BESI) for simplicity. We hypothesize that BESI can be used to produce a clear indication of similarities and compare to phylogenetic trees to assess the value of the method in a comparative analysis.

I. Phylogenetic Trees

Phyologenetic trees were constructed as follows. Sequences were separated by subject, and aligned with MAFFT v7.222 using the L-INS-i strategy [28]. A maximum likelihood (ML) phylogenetic tree was constructed using the RAxML software, version 8.2.11 [29] with the HIVW amino acid model of substitution [30] and 100 bootstrap replicates. Trees were midpoint-rooted using the phylogenetic visualization software FigTree, version 1.4.3 [31].

IV. Results

The initial analysis confirmed the process produced acceptable results in a comparative view of the original work performed by Stieh et al. [1]. Figure 2 displays a typical view of a single sequence after processing that expresses the surface charge of a bound (top) and unbound (middle) structure and the difference between the structures, bound - unbound (bottom). Looking at the lower graph in Figure 2, in the pH range of 4 to 6, one can observe a dip in the

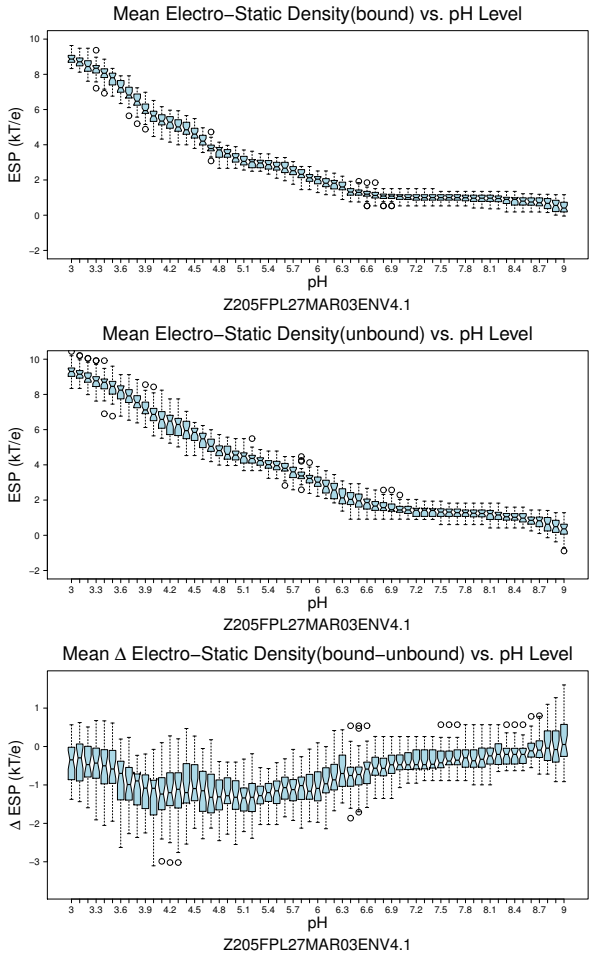

Fig. 2. Electrostatic surface charges. Figure shows bound (top) and unbound (middle) surface charges and the difference, bound - unbound (bottom), illustrating the sensitivity fingerprint.

remaining electrostatic density. This low value range is due to the higher charge value of the unbound structure and is analogous to the fingerprint observed in [1]. The phenomenon is determined across the entire Env set and the additional data supports the general hypothesis that Env preferentially binds CD4 at mucosal pH. Figure 3 provides a representative sample of the more than 250 fingerprints produced.

The raw data produced by the BESI pipeline is dependent upon the number of models of each sequence produced and the list of pH values in the titration. For this study we modeled 30 structures per sequence to ensure a wide range of randomized ligand conformations were represented by Modeller. We processed each model in a range of pH levels from 3 to 9 in increments of 0.1. We processed each sequence with the same parameters to produce consistent and uniform details across the study. Results for each sequence/conformation combination of 61 pH values is then compiled into a single table (m x n) where 'm' represents the number of models and 'n' exhibits the pH solutions. This data is then processed to determine the principal components for each sequence.

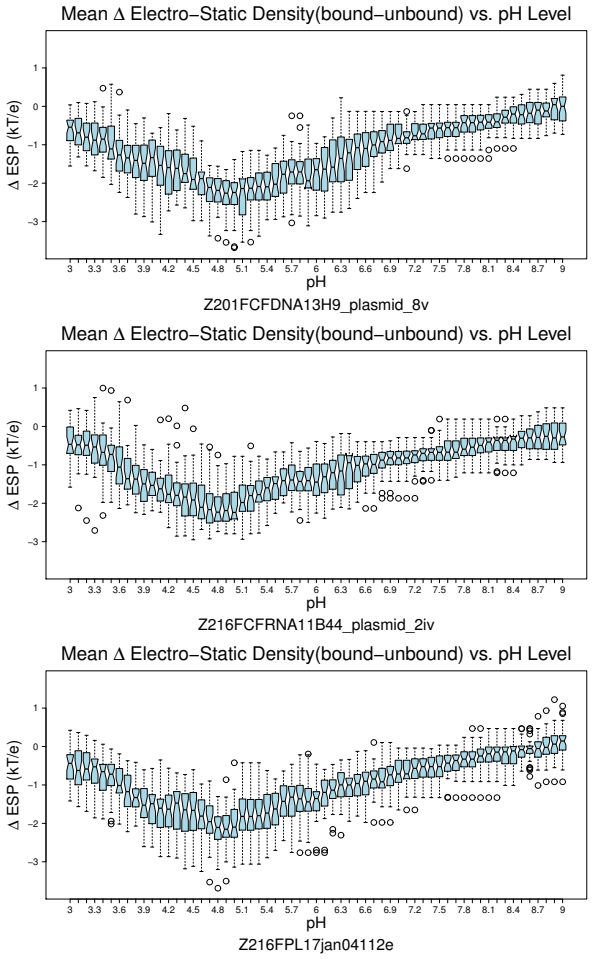

Fig. 3. Electrostatic fingerprinting. Figure provides representative fingerprints displaying the wide range of variation across gp120 structures.The area of significance is at pH levels between 4 and 6 where the surface charge difference between bound and unbound conformations is largest. This fingerprint is present for every gp120 structure evaluated using the current method.

We initially compared the first PC of the unbound named variant 'Z242MPL25JAN03PCR23ENV1.1DonorTransmitted' against the first PC of all other subspecies in unbound conformations as provided by Boeras et al. and retrieved a signal for several sequences. We then sorted the results by donor and score in Figure 4 to provide a clear representation of significance. Subsequent runs of the process were used to extract the list of sequences that were within a 20% threshold of the control. Table II provides data extracted from the initial cosine similarity analysis and the list was observed to show signs of significance. *Pairs* represent Donor/Recipient couples from [2] and indicate a potentially transmitted sequence from one host to another. *Matches* are significant in that they represent duplicate samples from a single host. *Singles* are from hosts with undetermined p24, sero, and genital fluid sample data as provided to this study.

Additionally, this could be expanded to include a number of principal components comprising a representation of some numbered order of variance. This is a standard practice and is

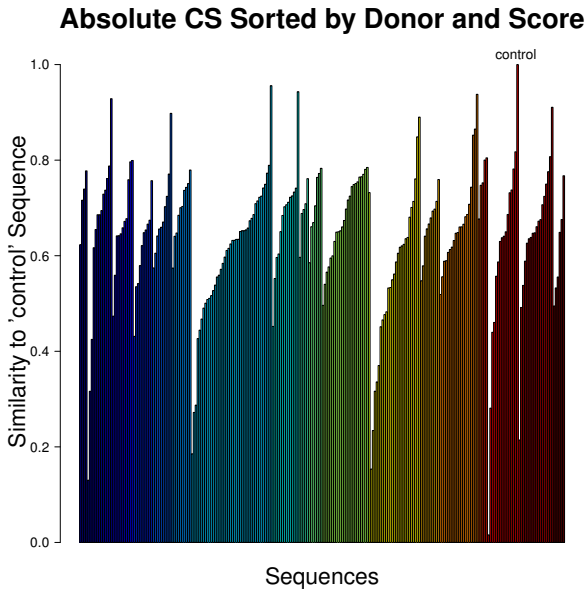

Fig. 4. Ordered CS signal. Figure represents the cosine similarity of the first principal component. The x-axis represents each gp120 variant grouped and colored by donor and sorted by score.

TABLE II
INITIAL FINDINGS. LIST OF SIGNIFICANCE FROM CSA. PAIRS REPRESENT DONOR/RECIPIENT TRANSMISSION VARIANTS, MATCHES INDICATE SIMILAR MUTANTS WITHIN A SINGLE HOST, AND SINGLES SIGNIFY A MATCH TO THE CONTROL ONLY.

Pairs	**Score**
Z201FPL7FEB03ENV2.1	0.956
Z201MPL7FEB03ENV2.1	0.943
Z242FPL25JAN03PCR8ENV1.1	0.800
Z242MPL25JAN03PCR23ENV1.1-DonorTransmitted	1.0
Matches	**Score**
Z221FPL7MAR03ENV2.3	0.890
Z221FPL7MAR03ENV3.3	0.849
Z238FCF15A39_plasmid_9ii	0.865
Z238FCF29oct0215A39	0.938
Z238FPL29nov024	0.852
Z242FPL25JAN03PCR23ENV1.1	0.805
Z292FCA24may0512D10_plasmid_5iii	0.807
Z292FCF24may0512E26_plasmid_10iv	0.911
Singles	**Score**
R56MCF21aug0511_plasmid_1v	0.929
Z185FPL17AUG02ENV3.1	0.898

referred to as dimensionality reduction with several stopping rules for the selection of significant components such as: Kaiser's rule, Scree plots, number of non-trivial factors, *a priori*, and percent total variance [26]. Kaiser's rule states that all PC's having a standard deviation of ≥ 1.00 be included in the set for evaluation [27]. Scree plots are a visual examination, that could be coded against, referencing an almost subjective determination of the angle of the graph relative to the x-axis and implies Kaiser's rule via a measurement of the standard deviation. Options 3 and 4 are not applicable in this reference space leaving option 5 as a suitable alternative to

Kaiser. We dismissed all basic rules in this case on the grounds that data reconstruction is not taking place. We therefore present the following explanation for the selection of Principal Components for this analysis.

We submit that the usable PC's for our purposes are the first two PC elements where the mean percentage of variance across the sample set is above fifty percent. We note that the addition of principal components above two presents an attenuation of the data and render a challenge in presentation. Additionally, we note these inclusions do not make a significant change in the selection of the closest match across the entire study. We compare the attenuation effect to that of a photograph, the most common form of dimensionality reduction in use to date, the image displays evidence of a third dimension, however, it is impossible to measure the depth of any one component as the projection of that object onto a two dimensional space distorts the representation through attenuation of the representative rays. Supporting information can be referenced at https://github.com/jlphillipsphd/besi/blob/master/Morton-Phillips_Computational_Advances-SI.pdf under section I 'Selection of Principal Components.'

The selection of additional components also requires the understanding that each eigenvector is orthogonal to the next. This fact imposes a requirement of abstraction to avoid cancellation. We circumvent this limitation by calculating the CS for the second PC and use the mean of the absolute values of the two cosine similarity scores for each sequence. For example, the second PC can influence the total similarity of the structure in comparison to the control by averaging out a strong hit in the first PC with a weaker hit from the second, ie: the CS of $PC_1 = 0.8$, and the CS of $PC_2 = 0.2$, the mean of the two values is given by $mean(c(0.8, 0.5)) = 0.5$, and the overall similarity of the selected sequence is reduced. The remaining piece to be determined is the means by which the results are compared.

Phylogenetic trees are useful in representing evolutionary data in a graphical format and are an established and mature means of inferring the relationships of genetic evolution for biological species. All images were rendered using R [22] and the APE version 4.1 [32] package by employing a gradient color scheme to each tip of the trees as a representation of BESI scores for each sub-species envelope.

Finally, we make the supposition that 'Z242MPL25JAN03PCR23ENV1.1' is the result of infection by 'Z242FPL25JAN03PCR23ENV1.1DonorTransmitted' based on the following information: the two sequences vary by a single residue and based on the naming conventions used by Boeras et al. [2], the polymerase chain reaction (PCR) designation, and Env identifier of the donor and the recipient subspecies match. We conclude the accuracy of the indexing process based on Figure 5. This conclusion is established by the phylotree arrangement as compared to the closest match provided by BESI for pair Z242. the graph is interpreted by familiarizing ones self with the gradient color schemes for both donor and recipient where the likeness of the subspecies to the control increases towards the top of the gradient. The phylogeny tree follows standard methods of representation in that branches to the left indicate an evolutionary clade and the length of the branch implies the significant difference between two clades. The closest score for the donor and recipient are noted by three(3) asterisk (***) before the subspecies name. Because this is the control set, sequence Z242MPL25JAN03PCR23ENV1.1DonorTransmitted has an automatic score of 1.0. The score for sequence Z242FPL25JAN03PCR23ENV1.1 can be obtained from Table II, and is 0.800. The color gradient is based on the value of the score and can be the same for closely valued sequences. This illustrates the coloring method used on the tips and also exposes how a single difference in amino acid can effect the electrostatic characteristics of a protein structure.

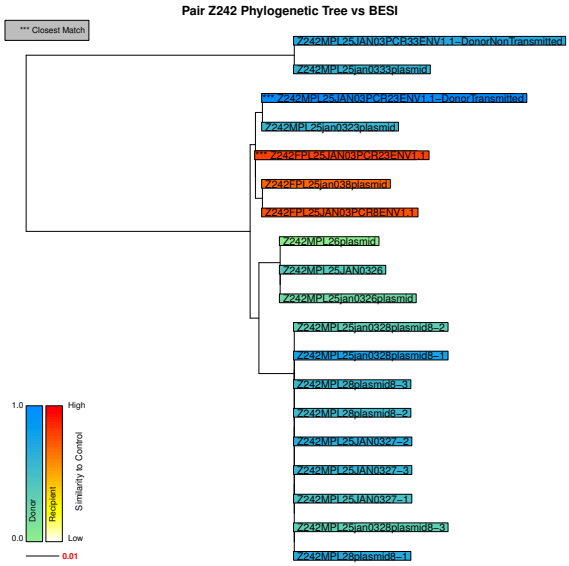

Fig. 5. BESI applied to the named variant displaying agreement with a phylogeny tree. Branch length and scale are both observed to indicate proper selection through evolutionary means and by BESI. This conclusion is based on the proposition that 'Z242MPL25JAN03PCR23ENV1.1' is the result of infection by 'Z242FPL25JAN03PCR23ENV1.1_Donor_Transmitted.'

Expanding the application of this method to the other donor/recipient pairs produces the following list of additional closest matches to the control:

R56FPL21apr05E7_plasmid_b
R56MCF21aug0511_plasmid_1v
Z153FPB13MAR02ENV1.1
Z153MPL13MAR02ENV1.1
Z185FPL17AUG02ENV3.1
Z185MPB17AUG02ENV1.2
Z201FPL7FEB03ENV2.1
Z201MPL7FEB03ENV2.1
Z205FPL27MAR03ENV4.1
Z205MPB27MAR03ENV9.1
Z216FPL17jan0485f
Z216MPL133_plasmid

Z221FPL7MAR03ENV2.3
Z221MPB7MAR03ENV5.4
Z238FCF29oct0215A39
Z238MPL17_plasmid_a
Z292FCF24may0512E26_plasmid_10iv
Z292MPL113_plasmid_e

Reviewing the BESI/Phylogenetic trees we assert that pair Z242 is the known transmission set that provides the control for this research, see Figure 5. Both of the top BESI hits for the Z242 pair are found in the same clade of the tree, indicating that the surface charge fingerprint extracted by BESI is predictive of sequence transmission potential. To express the accuracy of BESI for determination of transmission subspecies we present pair R56 (Figure 6) and associated scores (Table III). Considering that upon infection a mutation takes place, it should be no surprise that in some cases the mutation drifts enough to shift the score to the point of a lost signal. This appears to be the case when referencing Figure 6 and the associated scores, noting that the donor has a solid hit.

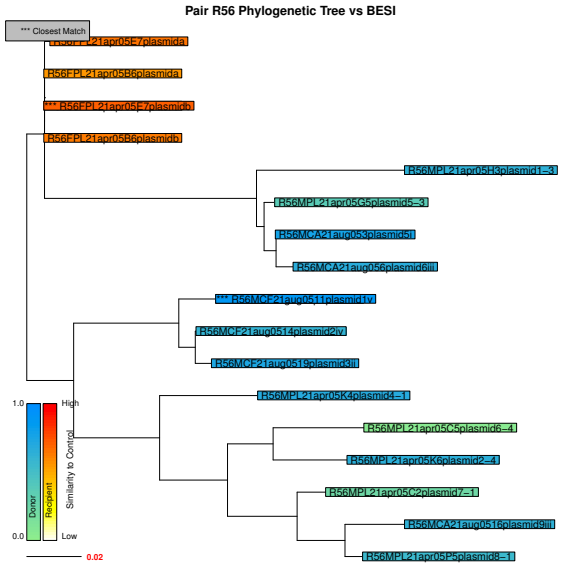

Fig. 6. Phylogeny versus BESI for pair R56. Displays disparate selections by BESI indicated by distinct branches of the phylogeny tree.

The recipient is effectively irrelevant in terms of a solution to transmission in that the donor supplies the subspecies that infects and is reflected in the phylogenetic trees by the recipient sequences segregating into a single clade. Out of the ten donors, six scored above 0.85 with the remaining four having scores in excess of 0.75. We present all graphs of the donor/recipient pairs for examination and consideration in the supporting information document hosted at https://github.com/jlphillipsphd/besi/blob/master/Morton-Phillips_Computational_Advances-SI.pdf under section II 'BESI vs. Phylogeny Trees.' The document presents all scoring information beside each graph for easy comparison. The table data for each pair tree is sorted alphabetically, then by score and the tables are clearly marked for

TABLE III
BESI SCORES FOR PAIR R56. THE LOW SCORES FOR THE RECIPIENT (R56F) ARE IN AGREEMENT WITH THE PHYLOGENY TREE OF R56 (FIGURE 6) DISPLAYING THE ACCURACY OF THE SELECTION METHOD PRESENTED BY BESI. THE DONOR HAS A HIGH MATCHING SCORE, INDICATING THAT THE SUBSPECIES OF TRANSMISSION IS PRESENT.

Role	Sequence	Score
Recipient	R56FPL21apr05B6_plasmid_a	0.563919054126656
	R56FPL21apr05B6_plasmid_b	0.64914025990008
	R56FPL21apr05E7_plasmid_a	0.668113200596071
	R56FPL21apr05E7_plasmid_b	0.726523796985736
Donor	R56MPL21apr05C5_plasmid_6-4	0.069038035927937
	R56MPL21apr05C2_plasmid_7-1	0.240246676350975
	R56MPL21apr05G5_plasmid_5-3	0.385319672126639
	R56MPL21apr05P5_plasmid_8-1	0.562805416203866
	R56MCF21aug0514_plasmid_2iv	0.624100461824737
	R56MPL21apr05H3_plasmid_1-3	0.641810107144584
	R56MCA21aug0516_plasmid_9iii	0.6510877349576
	R56MPL21apr05K6_plasmid_2-4	0.661026070503639
	R56MCA21aug056_plasmid_6iii	0.687419348128837
	R56MPL21apr05K4_plasmid_4-1	0.709966645467017
	R56MCF21aug0519_plasmid_3ii	0.752070920633736
	R56MCA21aug053_plasmid_5i	0.784921204312612
	R56MCF21aug0511_plasmid_1v	0.914998937284965

'Donor' and 'Recipient' roles. The graphs are vector based graphic presentations to allow close examination as some representations are tightly nit. Each 'closest hit' is marked with three (3) asterisk (***) to allow one to easily distinguish the BESI selections. The color gradients can be used to quickly determine the relative score of each sequence in comparison to the evolutionary data. Keep in mind that the BESI score is relative to a single control sequence and is not intended to match the phylogeny tree exactly.

V. DISCUSSION

Envelope binding rates are influenced by pH and are greater in the acidic conditions present in genital mucous [1]. Stieh et al. additionally provided computational evidence to support their findings by computation of the mean surface potential of the gp120 envelope and comparing the difference between bound and unbound gp120 conformations across a wide pH range. The difference in pH has been shown to alter the residual surface charge of the Env and shift the binding characteristics in laboratory tests [1]. The overall goal of this research is to predict which Env variants are most likely to bind to CD4 receptors while taking into account the role of environmental pH in the transmission process. This information would allow researchers to focus on the communication of the virus in a preventative manner, on a focused set of subspecies.

BESI represents a high throughput method for processing sequences in an automated fashion. The program is easily configured and the pipeline is efficient in terms of execution time. The driver developed for this study utilized a simple work unit delivery method and recovery scheme.

The inclusion of full structure models enhances the original process and establishes a more realistic approach to studying protein sequence data. Additionally we developed a PCA-based method of structure evaluation, in terms of electrostatic

fingerprinting, similar to LSI, to provide a means of visual qualitative comparison between phylogeny and electrostatic surface charge developed by Steih et al.

In addition to the analysis presented here, we suspect BESI would be a useful process for a variety of fields such as protein/enzyme engineering for optimal performance in different pH conditions, developing pH-specific functionality, or evolutionary studies of pH-dependent protein function acquisition. For detailed implementations at an engineering level, BESI can be used to pre-evaluate structures before physical experiments take place that have the goal of specific pH functionality.

Despite the increased number of gp120 assemblies used in this study compared to Stieh et al., further examination across larger numbers of subspecies would be beneficial. The study by Boeras et al. sampled in excess of 900 different RNA sequences across sixteen individuals. Performing BESI across the provided population by Boeras et al. has exposed unique, identifiable electrostatic characteristics which enable their identification (useful for vaccine development) and verified the hypothesized mechanism. However, the assumptions of transmission and the accuracy of the provided data may play a role in the relative statement of positive results. For example, the current process is generalized by the mean electrostatic charge of the gp120 structure and the possibility of focus at the periphery of the Env, specifically at the CD4 binding site, presents an opportunity to investigate.

VI. Acknowledgments

We would like to thank Cynthia Derdeyn (Emory University) for providing the sequence data and Peter Hraber (Los Alamos National Laboratory) for helpful discussions about the sequence data and analyses used in the study by Boeras et al.

References

[1] Stieh, D. J., Phillips, J. L., Rogers, P. M., King, D. F., Cianci, G. C., Jeffs, S. A., Shattock, R. J. (2013): *Dynamic electrophoretic fingerprinting of the HIV-1 envelope glycoprotein.* Retrovirology, 10(1), 33. https://doi.org/10.1186/1742-4690-10-33

[2] 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), E1156E1163. https://doi.org/10.1073/pnas.1103764108

[3] Mehrishi JN, Bauer J: *Electrophoresis of cells and the biological relevance of surface charge.* Electrophoresis 2002, 23(13):19841994.

[4] Richmond DV, Fisher DJ: *The electrophoretic mobility of microorganisms.* Adv Microb Physiol 1973, 9:129.

[5] Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, Baker NA: *PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations.* Nucleic Acids Res, 35, W522-5, 2007.

[6] Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA: *PDB2PQR: an automated pipeline for the setup, execution, and analysis of Poisson-Boltzmann electrostatics calculations.* Nucleic Acids Res, 32, W665-W667, 2004.

[7] Olsson MHM, Sondergaard CR, Rostkowski M, Jensen JH: *PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK(a) Predictions.* J Chem Theory Comput 2011, 7(2):525537.

[8] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C: *Comparison of multiple amber force fields and development of improved protein backbone parameters.* Proteins-Structure Function and Bioinformatics 2006, 65(3):712725.

[9] Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., & Shaw, D. E. (2010). *Improved side-chain torsion potentials for the Amber ff99SB protein force field.* Proteins: Structure, Function, and Bioinformatics, 78(8), 1950-1958. https://doi.org/10.1002/prot.22711

[10] Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA: *Electrostatics of nanosystems: application to microtubules and the ribosome.* Proc. Natl. Acad. Sci. USA 98, 10037-10041 2001.

[11] Humphrey W, Dalke A, Schulten K: *VMD: Visual molecular dynamics.* J Mol Graph Model 1996, 14(1):3338.

[12] Sali A, Blundell TL (December 1993): *Comparative protein modelling by satisfaction of spatial restraints.* J. Mol. Biol. 234 (3): 779815. doi:10.1006/jmbi.1993.1626. PMID 8254673

[13] Pearson, WR; Lipman, DJ (1988): *Improved tools for biological sequence comparison.* Proceedings of the National Academy of Sciences of the United States of America. 85 (8): 24448. PMC 280013Freely accessible. PMID 3162770. doi:10.1073/pnas.85.8.2444

[14] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000): *The Protein Data Bank Nucleic Acids Research.* 28: 235-242. http://www.rcsb.org

[15] Farrell DW, Speranskiy K, Thorpe MF: *Generating stereochemically acceptable protein pathways. Proteins.* 2010;78: 29082921. doi:10.1002/prot.22810. Pmid:20715289

[16] Bing Chen, Erik M Vogan, Haiyun Gong, John J Skehel, Don C Wiley, and Stephen C Harrison. 2005: *Structure of an unliganded simian immunodeficiency virus gp120 core.* (2005). DOI:https://doi.org/10.1038/nature03327

[17] Berendsen, H.J.C., van der Spoel, D. and van Drunen, R: *GROMACS: A message-passing parallel molecular dynamics implementation.*, Comp. Phys. Comm. 91 (1995), 43-56.

[18] Lindahl, E., Hess, B. and van der Spoel, D: *GROMACS 3.0: A package for molecular simulation and trajectory analysis.*, J. Mol. Mod. 7 (2001) 306-317.

[19] Lindstrom, P. (2014): *Fixed-Rate Compressed Floating-Point Arrays.* IEEE Transactions on Visualization and Computer Graphics, 20(12), 26742683. https://doi.org/10.1109/TVCG.2014.2346458

[20] *Python* https://www.python.org/

[21] *MPI4PY* http://pythonhosted.org/mpi4py/

[22] R Core Team (2016). R: *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[23] Pearson, K. (1901): *On Lines and Planes of Closest Fit to Systems of Points in Space (PDF).* Philosophical Magazine. 2 (11): 559572. doi:10.1080/14786440109462720

[24] Hotelling, H. (1933): *Analysis of a complex of statistical variables into principal components.* Journal of Educational Psychology, 24, 417441, and 498520. Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 28, 321377

[25] Singhal, Amit (2001): *Modern Information Retrieval: A Brief Overview.* Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 3543.

[26] Brown, J. D. (2009)., *Choosing the Right Number of Components or Factors in PCA and EFA.* Shiken: JALT Testing & Evaluation SIG Newsletter, 13(2), 1923. Retrieved from http://hosted.jalt.org/test/PDF/Brown30.pdf

[27] Bandalos, D.L.; Boehm-Kaufman, M.R. (2008): *Four common misconceptions in exploratory factor analysis.* In Lance, Charles E.; Vandenberg, Robert J. Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences. Taylor & Francis. pp. 6187. ISBN 978-0-8058-6237-9.

[28] 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

[29] 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

[30] 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

[31] Rambaut, A., *FigTree v.1.43.* http://tree.bio.ed.ac.uk/software/figtree/

[32] Paradis E., Claude J. & Strimmer K. 2004., *APE: analyses of phylogenetics and evolution in R language. Version 4.1.* Bioinformatics 20: 289-290.