A Computational Electrostatic Modeling Pipeline for Comparing pH-dependent gp120-CD4 Interactions in Founder and Chronic HIV Strains

By

Jonathan Howton

A thesis submitted in partial fulfillment

of the requirements for the degree of

MASTER OF SCIENCE

in

Computer Science

Middle Tennessee State University

May 2017

Thesis Committee:

Dr. Joshua L. Phillips

Dr. Sal Barbosa

Dr. Stephen Wright

I would like to thank my amazing thesis advisor, Dr. Joshua L. Phillips, for all of his time and patience throughout this work. He was an excellent mentor who was always available to help me through any problems I encountered, and he always ensured that I left our meetings focused on the next goal with the tools and insight needed to accomplish it.

I would also like to thank my thesis committee members, Dr. Sal Barbosa and Dr. Stephen Wright, for their insightful comments and suggestions.

Additionally, I would like to thank the Computer Science Department. I greatly appreciate the opportunity that it gave to me. I applied to the program without a background in Computer Science, and the decision to admit me and allow me time to remedy these deficiencies enabled me to pursue a career that I otherwise would not have been possible. This decision truly changed my life. On top of this, I have had the honor of being taught by some of the absolute best professors I have ever encountered.

Thank you Kristen and Riley. Getting to come home to such a loving and supportive family made it easier to overcome the more stressful obstacles I encountered. I appreciate you both being so understanding of the time I needed to complete this work and for giving me additional motivation to focus and push through the difficult parts. I would also like to thank the rest of my family for their unwavering support and encouragement. I love you all.

ABSTRACT

Though Human Immunodeficiency Virus has been studied for several decades, a consistently effective vaccine has not yet been produced. While most experimental and computational work in this area has been performed under slightly basic conditions (eg. blood/-plasma), the viral transmission event generally occurs at the highly acidic mucosa. Since pH can greatly affect protein structure, it likely affects epitope exposure to either inhibit or facilitate transmission. In this thesis, a pipeline for analyzing the pH sensitivity of protein-protein interactions is applied to the transmission critical interaction between the HIV gp120 and host CD4 proteins. The interaction between gp120 and CD4 is shown to be stronger at low pH for all strains tested, which is consistent with previous work and supports the accuracy of the introduced pipeline. Also, early transmitted founder (TF) strains generally bind CD4 better at low pH and are more pH sensitive than systemically circulating chronic control (CC) strains.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I.

INTRODUCTION

More than thirty years after the discovery of Acquired Immune Deficiency Syndrome (AIDS), there is still no vaccine against the Human Immunodeficiency Virus (HIV) that causes the disease. While antiretroviral therapies are quite effective at reducing the transmission rate of HIV [40, 1], economic and social challenges [41], as well as a need for extremely high adherence to treatment [11] prevent this from being a universally viable option; vaccines would provide a much simpler and direct means of preventing the spread of HIV [29].

HIV has a very high mutation rate, so antigenic regions which are targeted by antibodies vary greatly across HIV virions within a single host. Most vaccine research has focused on inducing broadly neutralizing antibodies (bnAbs). However, bnAbs are only produced by a small fraction of individuals infected with HIV, and the production of these antibodies only occurs after chronic infection [36]. The bnAbs are able to target regions of the virus that must be conserved due to functional requirements [10], most of which are found on the gp120 extracellular subunit of the envelope protein (Env) that is responsible for binding CD4 on the surface of T-Cells to begin infection [62]. This indicates that the CD4 binding region of Env is very important for vaccine production, since the virus must conserve this region to maintain its ability to infect [10].

Vaccines have been produced from Env fragments that have been computationally optimized to invoke the production of bnAbs [22]; results from these vaccines have varied from successful [7] to unsuccessful [35]. A possible explanation for this inconsistency is that the bnAbs are isolated from the blood, which has a slightly basic pH, while HIV is transmitted at the mucosa, which is highly acidic. Since protein structure and protein-protein interactions are typically affected by pH, it is likely that the structure of Env and its affinity for other proteins, such as CD4 and bnAbs, are altered. It has been shown that gp120,

the subunit of Env that interacts with CD4, binds CD4 better under acidic conditions [55]. This indicates that HIV is better able to bind its target and begin infection under lower pH conditions. Additionally, since HIV mutates rapidly within the host, the strains in a chronic infection, so called chronic control (CC) strains, will likely have adapted to the systemic pH, and will be less efficient at binding CD4 under acidic conditions when compared to transmitted founder (TF) strains. Consequently, the bnAbs produced in chronically infected individuals are less likely to neutralize HIV transmission at the mucosa. Therefore, it is important to study gp120-CD4 binding under mucosal pH because this conserved interaction is an important target for vaccine production.

The large variation in gp120 sequence across HIV strains makes experimental studies prohibitive, but computational modeling can aid in filling this gap in a predictive capacity. In order to model the large number of sequences in this dataset, a pipeline must be able to create fast, accurate models of Env and the Env-CD4 interaction. It also must be able to incorporate environmental permutations, such as pH and salinity, so that their effects can be evaluated. Lastly, the pipeline must robustly incorporate potential structural rearrangements because entropic factors contribute to the stability of particular conformations, which affects the stability of an interaction.

In this thesis, a dataset of TF and CC pairs, which spans HIV clades B and C, was used to test several hypotheses. It was predicted that the Env-CD4 interaction would be strongest at low, mucosal level pH. It was also predicted that the Env protein from TF strains would bind CD4 better under low, mucosal level pH when compared to CC strains, and that this interaction would be more pH sensitive in TF strains. To test these hypotheses, a new pipeline was constructed that has all of the aforementioned necessary components.

The data gained from the pipeline was used to elucidate potential mechanisms responsible for differences between HIV classes. Key structural motifs were identified which have implications for future study of Env-CD4 and Env-antibody interactions.

CHAPTER II.

BACKGROUND

HIV Env Protein Structure

The HIV envelope protein (Env) (Figure 1A) is a non-covalently linked homotrimer of heterodimers [38]. Each heterodimer consists of the transmembrane glycoprotein gp41 and the surface glycoprotein gp120 [24]. The gp120 subunit contains 5 conserved (C1-C5) and 5 hypervariable (V1-V5) regions [32]. The gp120 surface that composes the interface between gp120 and its target receptor CD4 is large, 800Å^2, and it contains a highly variable 280Å^3 cavity that does not interact with CD4 as well as a highly conserved 150Å^3 cavity that is required for CD4 binding.

HIV Env Protein and Infection

Env is responsible for target cell recognition and initiating fusion. It is first involved in target cell adhesion; this can be a specific [52] or non-specific [24, 32] interaction with cell attachment factors on the surface of the target cell. While these attachment factors are not essential, they likely bring Env close to CD4 for the required Env CD4 binding event [60].

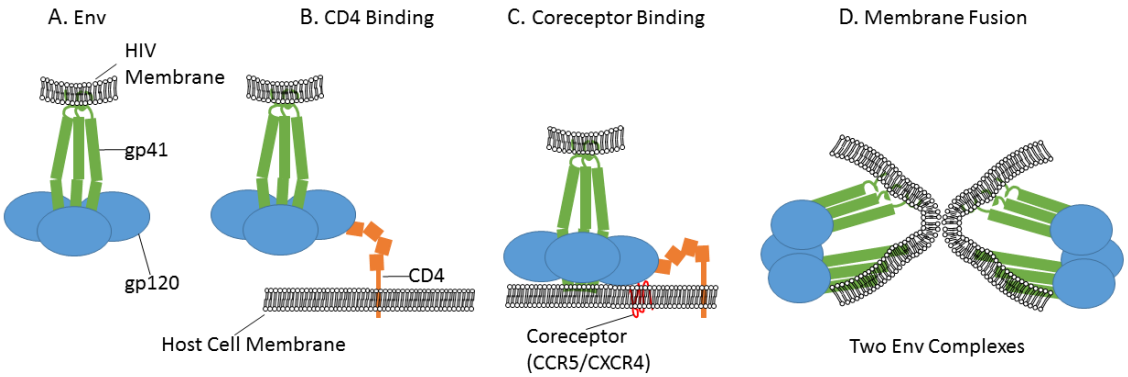


Figure 1: HIV Binding and Entry. A) HIV Env consists of gp120 and gp41. B) gp120 binds CD4 on the host cell. C) This triggers a conformational change that allows for coreceptor binding. D) This initiates membrane fusion.

When Env binds CD4 (Figure 1B) through the conserved CD4 binding site on gp120, Phe43 of CD4 blocks the entrance of the 150Å^3 cavity of the CD4 binding site; this is critical for this interaction to occur [47]. Env binding CD4 leads to rearrangements in the V1/V2 loop as well as V3, along with a shift in a 4 -strand bridging sheet [14, 60].

These shifts in gp120 structure induced by binding to CD4 enable gp120 to bind a coreceptor (Figure 1C), which appears to nearly always be CCR5 or CXCR4 [4, 47]. The coreceptor interaction induces the exposure of the hydrophobic gp41 fusion peptide, which inserts into the host membrane and forms a fusion pore (Figure 1D), which facilitates the delivery of the viral contents into the host cell cytoplasm [60].

HIV Vaccine

Binding to CD4 and the coreceptor is necessary for infection, so both the CD4 binding site and the coreceptor binding site of gp120 must maintain conserved regions. This makes these regions strong candidates for epitopes to include in a vaccine. However, the coreceptor binding site does not fully form until CD4 binds gp120, so the coreceptor binding site is protected from neutralization [24]. This makes the CD4 binding site a stronger candidate for effective vaccine production. To further support this, the most abundant bnAbs target the gp120 CD4 binding site to inhibit the interaction between Env and CD4 [62].

While production of bnAbs is a primary goal of an HIV vaccine, bnAbs are typically specific to the founder strain and occur in only approximately 20% of infected individuals after chronic infection [36]. Additionally, the maturation process of the bnAbs seems much more complex than typical antibody production; gp120-bnAbs targeting the CD4 binding site accumulate 40 to 46% changes in the variable domain-amino acid sequence during affinity maturation, which is much larger than the typical 5 to 15% mutation rate in this region [61]. This is likely due to the highly variable exposed region accompanied by the generally shielded conserved regions of the gp120 peptide [44].

To overcome this problem, composite peptides computationally optimized to contain the maximum number of epitopes from a set of viral proteins were produced to create a vaccine to effectively induce the production of bnAbs [22]. This approach had success in protecting rhesus monkeys from difficult-to-neutralize simian-human immunodeficiency virus SHIV-SF162P3 [7]; however, this approach was generally unsuccessful in the RV144 Thai HIV-1 vaccine efficacy trial [35].

A possible explanation for this inconsistency is that the isolated bnAbs come from the slightly basic blood, while HIV transmission occurs at the very acidic mucosa. This difference in environment would likely alter the surface chemistry and the epitope exposure of gp120. Indeed, it has been shown that the surface conformation of gp120 is affected by pH, and that gp120 binds CD4 better under acidic environments [55]. Consequently, a better understanding of the Env-CD4 interaction under mucosal acidity would likely be useful in determining effective epitopes for vaccine production.

Transmitted Founders and Chronic Controls

In most clinical HIV infections, a single TF virion is responsible for the transmission event [28]. TF viruses share common traits that distinguish them from chronic control (CC) strains, and these traits likely enhance TF virus fitness for crossing the mucosal barrier and promoting productive initial infection [45]. TF strains have a higher ENV content, which likely contributes to their increased virulence; while a particular mechanism was not determined, TF virions bind dendritic cells more efficiently and are more resistant to IFN-α [45]. Since TF virions are typically transmitted at acidic mucosa, it is expected that TF strains are better adapted for transmission at low pH relative to CC strains.

B Clade and C Clade

Clades B and C are subgroups of the HIV-1 group major [56]. Clade C is the most prevalent clade of HIV globally, and is the most common in China, India, and Africa, while clade B is most prevalent in America and Europe [3]. Clade C is less virulent [3],

which causes a slower progression and a longer asymptomatic period; this increases the opportunities for transmission [49].

Previous gp120-CD4 Modeling

The gp120 protein and the gp120-CD4 interaction were previously modeled using several solved gp120 structures [55]. In that study, partial atomic charges and protonation states were calculated across pH and salinity ranges using the PDB2PQR framework [17] and the PROPKA3.0 [42] program, respectively; the APBS [5] tool was used to determine surface charges. It was found that the surface potentials of the CD4 protein and the CD4 binding site of gp120 complemented one another at low pH.

The use of crystal structures allowed for calculating the effect of pH on the surface of an actual solved structure; it also eliminated the need for computationally predicting structures. However, this greatly limited the number of sequences that could be compared because each sequence required a solved crystal structure. Additionally, comparisons between conformations and ligand interactions were between different solved structures from different sequences. This makes it more difficult to draw conclusions from comparisons between a complex and an unbound conformation at a given pH.

The surface charge calculations from the previous modeling method [55] provided a broad potential mechanism for increased CD4 binding at low pH. However, the effect of these changes in surface charge on CD4 binding was not quantified.

Overcoming Previous Limitations

The limitations from using only crystal structures could be overcome by computationally producing accurate structures from available gp120 sequences. One available tool that can help achieve this goal is MODELLER [51], which aligns a protein sequence to a template structure to quickly produce a model for the protein sequence [19]. Another helpful tool is FRODAN, which is a computationally inexpensive tool that uses geometric targeting to shift the conformation of an input model towards the conformation of a target model [21]. These

tools require target structures to which they align the input data. There are several solved gp120 structures, such as 1RZK [26] and 2B4C [25], but all models have CD4, an antibody, or both bound. The only unbound model available is 2BF1, which is the SIV gp120 subunit [13]. While there are differences in the protein sequence, this structure does provide a loose template for unbound structural alignments with HIV gp120.

APBS [5] can be used to directly calculate the electrostatic contribution to the binding energy, ΔG, of the gp120-CD4 complex based upon solvation energy calculations. This will reveal the quantitative effect pH has on the gp120-CD4 interaction. This will also provide the ability to measure entropic factors based upon the solvation energy difference between the bound and unbound conformations. Conveniently, performing these calculations also produces the data necessary for determining surface potential, so it can be compared between the generated models as well.

CHAPTER III.

METHODS

gp120 Sequences

One TF sequence and one CC sequence were analyzed from each of 24 individuals. Of these 24 pairs of sequences, 18 pairs were B clade sequences, and 6 were C clade sequences. TF sequences were defined as sequences collected within the first 6 months of infection, while CC sequences were collected after this initial period. Sequence information is provided in Appendix A.

Pipeline Configuration and Automation

Bash, Python, and R [48] scripts were used to automate the modeling and analysis of all sequences within the dataset. The scripts were designed to provide sequences and target structures to the initial modeling step (Figure 2A), and then to progress through the pipeline (Figure 2B-G) by processing the output from the current step and providing it to the next step. This pipeline is easily applicable to modeling the effect of pH and salinity for other protein-protein interactions as well.

Sequences were evenly distributed among 4 rack mounted DELL R815 servers, each containing 4 16-core 2.3GHz AMD Opteron processors, 512GB of RAM, and utilizing RedHat Enterprise Linux 6.5 OS. Each server processed a single sequence completely before beginning with the next sequence. PDB2PQR [18, 17] and APBS [5] steps were ran as 8 parallel jobs to increase throughput. APBS [5] was allowed 8 cores per job.

Homology Models

The analyzed sequences do not have solved structures, so modeling was required to create structural data to analyze. MODELLER [51] was used with a pre-constructed set of seven template gp120 structures to produce a set of homology models (Figure 2A). The structures used were 1G9M [31], 1RZK [26], 2B4C [25], 2NY7 [63], 3JWD [43], 3JWO [43], and 3LQA [16]. Ten homology models were produced for each tested sequence. This

provides ten repetitions per model to account for natural structural variations in the flexible V regions.

Models of Bound and Unbound gp120 and the gp120-CD4 Complex

To determine the binding energy of the complex, the electrostatic energy from the complex and from the individual components of the complex must be determined, so models for these structures needed to be produced. Each homology model was processed to produce a model in the bound and unbound conformations as well as a model of the gp120-CD4 complex. FRODAN is needed to correct deviations in the core structure of the models, which commonly occur in protein models produced by MODELLER [46]. It is also needed for producing unbound models because the only available solved unbound target structure is 2BF1 [13], and MODELLER cannot produce an accurate model from a single example. Lastly, FRODAN allows for the accurate docking of CD4 to gp120 to form the complex structure with the produced models.

To produce the unbound conformation gp120 model, first the MODELLER [51] salign tool was used to align the 2BF1 unbound simian gp120 structure [13] to the coordinates of the gp120 model from the produced complex; then the FRODAN [21] tool was used to shift the conformation of the bound gp120 model towards the aligned 2BF1 conformation to produce the required unbound gp120 model (Figure 2B). The VMD [27] translate feature was used to separate CD4 and gp120 from the solved complex structure 1RZK, which is a crystal structure of a CD4 bound gp120 [26]. This file was split into a separate file for each chain. The MODELLER [51] salign tool was used to align every unbound homology model from Figure 2B to the coordinates of the 1RZK gp120 chain; this model is used as free unbound gp120 in later steps.

The newly aligned chain was concatenated with the separated CD4 chain PDB file that was created when splitting the 1RZK PDB file; this creates a file with CD4 and the new model of gp120 in a reasonable proximity to simulate binding. The FRODAN [21] tool uses

this concatenated file as the initial structure and the original 1RZK PDB file as the target structure to produce the gp120-CD4 complex with the new gp120 model (Figure 2D). The 1RZK structure is the perfect target for making the complex, because it is an actual solved CD4 bound gp120 complex [26]. The resulting complex file was used to create separate files for the gp120 and CD4 chains in the same conformation and position as in the complex. This created the required CD4 and bound conformation gp120 models.

Atomic Charges and Protonation States

Charge and protonation data are required for the electrostatic energy calculations, so the models needed to be converted to PQR format. For all models, PQR files were generated over the tested pH range for CD4, the gp120-CD4 complex, and gp120 in both the bound and unbound conformations (Figure 2E). The tested pH range was from 3 to 9 in increments of 0.1. For each model/pH combination, PQR files were produced using PDB2PQR 2.0.0 [18, 17], which used PROPKA 3.0 [42, 54] to determine the partial atomic charges and protonation states (Figure 2E).

Binding Energy

Electrostatic energy for each structure was calculated for all of the PQR files by using APBS 1.4 [5] to solve the full non-linear Poisson-Boltzmann equation (Figure 2F). For each set, the number of grid points, coarse mesh lengths, fine mesh lengths, and known center were calculated using the APBS [5] psize tool with the gp120-CD4 complex PQR file from the set; the APBS [5] calculation used these values for all molecules within the corresponding set. The counter ion (e.g. NaCl) concentration was set to 0.155M for ions with a +1 charge and for ions with a -1 charge. The calculations were carried out using water as the solvent and 310K as the system energy. Surface potential data were saved in DX format for each molecule within a set at whole number pH values to conserve space, as each DX file consumed approximately 150MB of disk space. Total data usage is described in Results.

Binding energies were calculated in two ways. The bound form binding energy was calculated by subtracting the electrostatic energies of both the CD4 molecule and the bound conformation of gp120 from the electrostatic energy of the gp120-CD4 complex at a given pH. The unbound form binding energy was calculated by subtracting the electrostatic energies of both the CD4 molecule and the unbound conformation of gp120 from the electrostatic energy of the gp120-CD4 complex at a given pH. Additionally, the difference between these two binding energies were calculated by subtracting the unbound binding energy from the bound binding energy within a particular set.

Binding energy sensitivity was determined as the binding energy at low-pH (3.5, 3.6, 3.7, 3.8, 3.9, 4.0, 4.1, 4.2, 4.3, 4.4, and 4.5) subtracted from the binding energy at high-pH (7.0, 7.1, 7.2, 7.3, 7.4, 7.5, 7.6, 7.7, 7.8, 7.9, and 8.0, respectively).This produces 11 binding energy sensitivity values for each of the 10 models within each sequence.

Individual sequence sensitivity was determined by pooling the 11 sensitivities from all 10 models within a sequence and creating a boxplot from these values. Group sensitivity was determined by finding the median sensitivity across the 10 models within each sequence; this produced 11 median sensitivity values for each sequence. These sensitivity values were pooled with the sensitivity values from all sequences within the group, and a boxplot was created from these pooled values.

Charge Density

Whole molecule charge density was calculated as the sum of all charges determined by APBS divided by the total solvent accessible surface area, which was determined using VMD [27]. The median charge density was determined within each group at pH 4, 5, 7, and 8.

Residue specific charge density was calculated as the sum of all charges determined by APBS that were on the surface of the residue, divided by the solvent accessible surface

area of the residue. VMD [27] was used to assign electrostatic charge coordinates to corresponding residues and to determine the solvent accessible surface area of each residue.

To determine pH sensitivity of whole molecule charge density, first charge density differences were determined by subtracting the unbound charge density from the bound charge density at each pH value. These charge density differences were directly compared. The sensitivity was calculated as this value at pH 4 subtracted from this value at pH 7, this value at pH 5 subtracted from this value at pH 8, or the average of these two sensitivities.

Data Analysis

All data analysis and plotting was performed in R 3.2.4 [48]. Box plots were created using the included boxplot function with notches enabled to automatically calculate a 95% confidence interval utilizing the method shown in [12]. The included Wilcoxon signed rank test was used to determine all confidence intervals. The included plot and matplot functions were used for all other plotting.

K Nearest Neighbor Model Fitting

A K nearest neighbors algorithm was used to determine the distance to the 100 closest coordinates in each generated model from each coordinate in the 1RZK [26] template model. The accuracy of each K value from 1 to 100, inclusive, was tested. The contributing charge of each point was determined as the charge of the generated model coordinate divided by the distance squared. The median of contributing charges 1 through K was assigned to the template coordinate. The distribution of charges on the fitted template model was compared to the distribution of the charges on the generated model using the KolmogorovSmirnov statistic. This statistic was determined for all models and compared across K values from 1 to 100, inclusive.

Sequence Alignment

Clustal Omega [53, 23, 39] was used to align all sequences. TeXshade [8] was used to format the alignment for publication and to highlight regions of consensus and similarity.

Sequence Alignment Based Model Fitting

Within each group tested, charges were assigned to alignment positions in two ways. The first way ignored gaps when assigning charge values to each position. Within each group, the median of the residue specific pH sensitivity of charge density from all residues at each position was found and assigned to the corresponding position; if no residues were present at the position within the group, then a value of 0 was assigned. The other method assigned a charge of zero to every gap in the alignment, so that gaps would contribute to the median charge determined at each position.

Groups were compared by subtracting each assigned value from the corresponding value in the other group. The top 1% of residues were identified as residues that had an absolute sensitivity difference larger than 1% of the absolute sum of residue sensitivity differences in a given comparison.

Mapping Sequences to Structures

The top residues identified in the alignment based model fitting were mapped onto one of the model structures produced for the EU744010 sequence in this study. This sequence was chosen because it was the longest sequence analyzed, so it covered the largest percentage of the alignment. To visualize sensitivity, calculated residue sensitivity values were inserted into the temperature factor column of the EU744010 PDB file. VMD [27] was used to visualize the PDB files. The Surf drawing method was used for the entire protein. The identified sensitive residues were colored using the Beta coloring method, which visualizes the temperature factor of the PDB file. The remaining residues were set to color ID 8 (white), and the material was set to Opaque for the first image; it was changed to Transparent for the second. The included snapshot tool was used to capture the images.

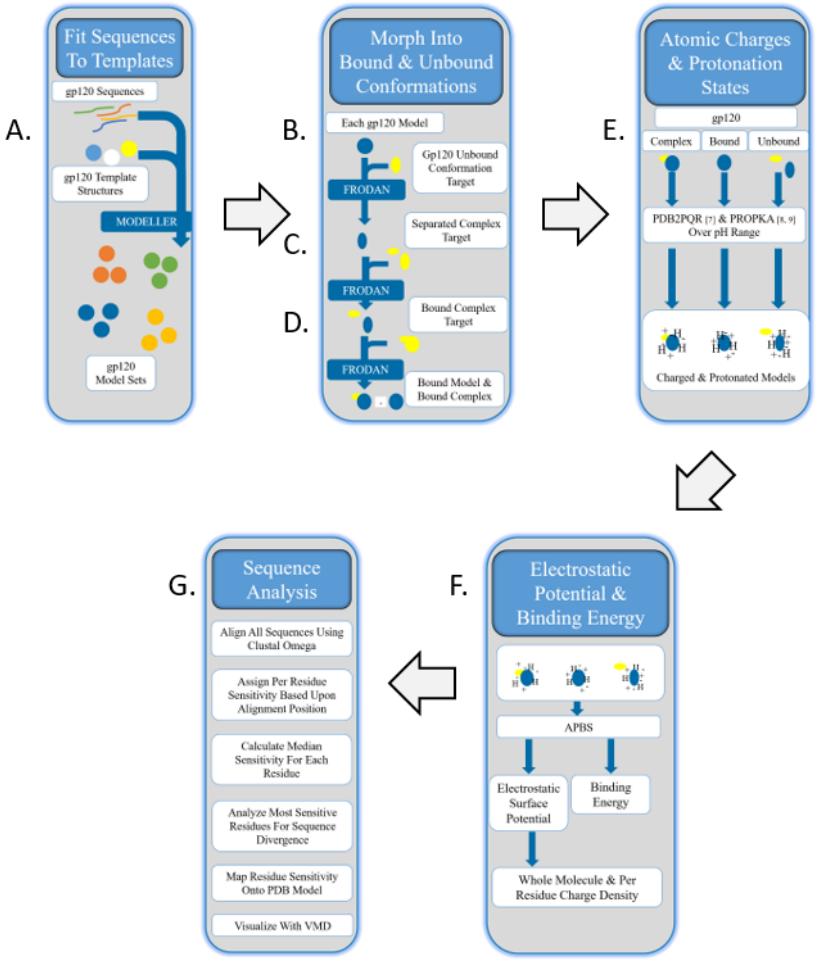


Figure 2: Pipeline for calculating and analyzing the pH sensitivity of the interaction between gp120 and CD4. A) MODELLER [51] is used to produce a model set from a gp120 sequence and several template structures. B) FRODAN [21] was used to shift the new separated complex model toward the conformation of the SIV 2BF1 unbound structure [13]. The model produced is used as the unbound gp120 structure in later steps. C) The chains from the solved 1RZK gp120-CD4 complex were separated using VMD [27]. The models produced in (2B) are aligned to the coordinates of the separated gp120 chain, and the conformation is shifted towards the bound conformation. D) The CD4 structure is added to the bound conformation and FRODAN [21] is used to shift these separated chains towards the 1RZK gp120-CD4 complex. E) PDB2PQR [18, 17] and PROPKA [42, 54] are used to produce PQR files from the bound complex model, CD4, the bound gp120 chain, and the unbound gp120 chain. This is determined across a range of pH values from 3 to 9, inclusive, in increments of 0.1. F) APBS [5] is used to determine the electrostatic surface potential of each model, as well as the binding energy of the complex. The electrostatic surface potential was used to calculate whole molecule charge density data. G) Electrostatic data and a Clustal Omega [53, 23, 39] sequence alignment were used to identify residues that potentially contribute to pH sensitivity.

CHAPTER IV.

RESULTS

Pipeline Throughput

An automated pipeline was successfully constructed for modeling pH sensitivity of a protein-protein interaction given a set of sequences to model and the required structural templates. In this work, each of the 48 sequences were used to create 10 sets of a bound conformation model, an unbound conformation model, and a complex model; each of these model sets was analyzed at 61 different pH values to determine the effect of pH on residue and molecule specific electrostatic data, as well as the binding energy of the complex. Each of the 4 machines in the cluster modeled and analyzed 12 of the 48 sequences. All models were produced within approximately 3 hours; the electrostatic and binding energy calculations were then completed within 3 weeks. With a total run time of 3 weeks using 256 cores, these calculations used approximately 130,000 CPU hours.

Electrostatic surface potential DX files were the largest use of space. Each DX file was approximately 150MB, and each model produced 4 DX files at each of 7 distinct pH values; therefore, each model consumed approximately 4.1GB of surface potential data. With 10 models per sequence and 48 sequences, 13,440 DX files were created and stored; this consumed approximately 1.9TB of surface potential data. All other generated files consumed a total of approximately 300GB, which increased the total usage to approximately 2.2TB of space.

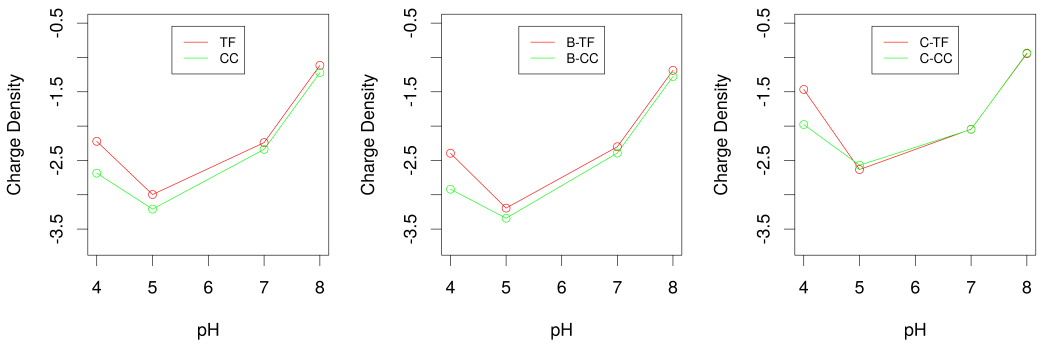
Charge Sensitivity

Whole molecule charge density was used to compare pH sensitivity between TF and CC strains at pH values 4, 5, 7, and 8; the largest difference in charge was found when comparing pH 5 to pH 8, and pH 4 was the only pH at which TF and CC differed (Figure 3A). This result was very consistent within B clade (Figure 3B). However, TF strains were found to be more positive at pH 4, 5, and 7 within C clade (Figure 3C).

Charge pH sensitivity was calculated as the difference in charge density between high and low pH values. TF and CC strains were compared by the difference between pH 4 and 7 (Figure 4A), pH 5 and 8 (Figure 4B), and the average of the two intervals (Figure 4C). CC strains were found to be more sensitive than TF strains when using the pH 4 and 7 interval and the average of the intervals (Figures 4C & 4A). This was consistent within B clade (Figures 4D & 4F), but no significant differences were found within C clade (Figures 4G, 4H & 4I).

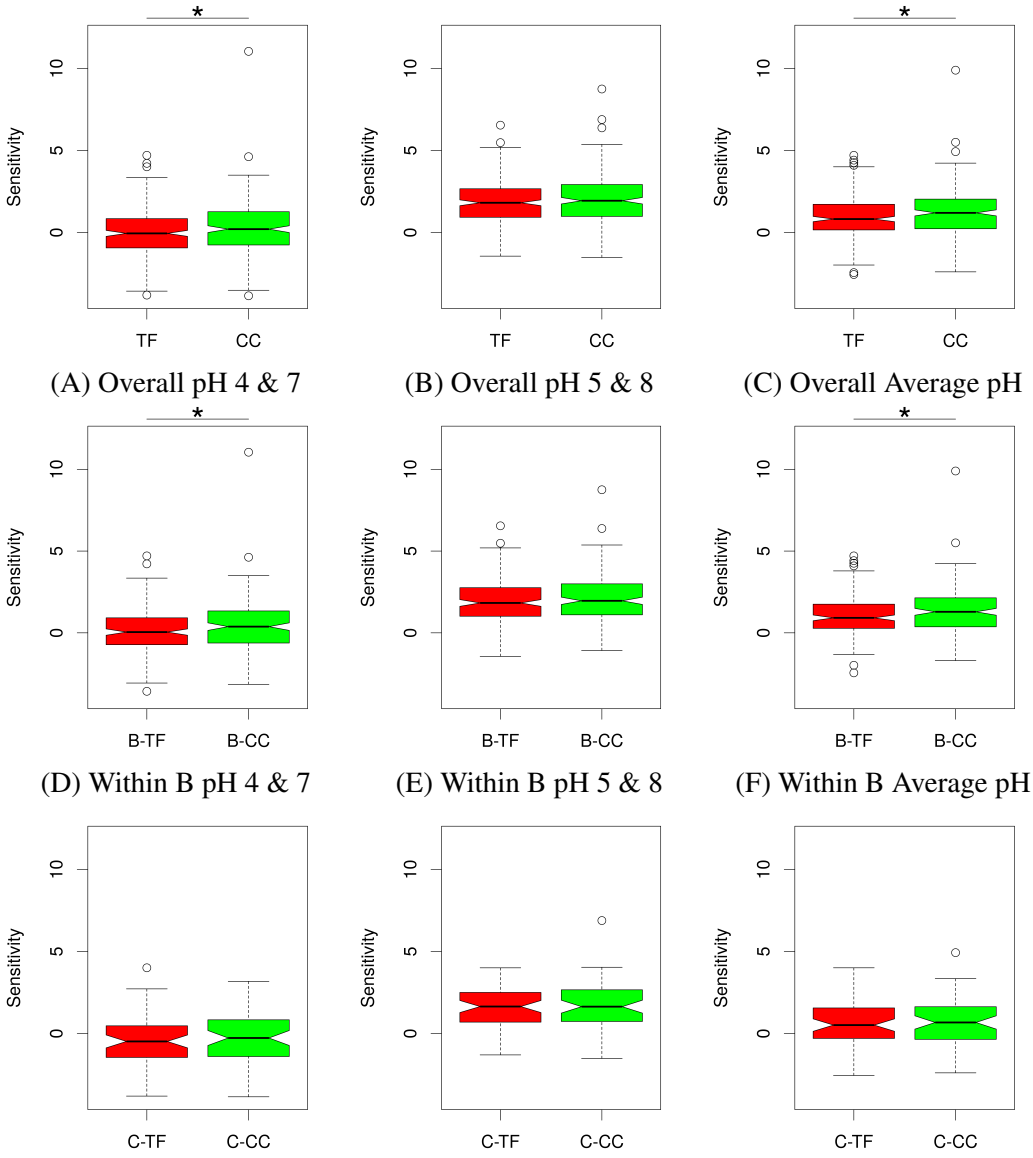
Clade B was more positive than clade C at all pH values tested; as with classes TF and CC, the largest difference in charge was found when comparing pH 5 to pH 8 (Figure 5A). This was consistent within classes TF (Figure 5B) and CC (Figure 5C).

B and C clades were compared by the difference between pH 4 and 7 (Figure 6A), pH 5 and 8 (Figure 6B), and the average of the two intervals (Figure 6C). B was significantly more positive under all of these conditions. This was consistent within all but the pH 5 and 8 difference in classes TF (Figures 6D, 6E & 6F), and CC (Figures 6G, 6H & 6I).



(A) Overall (B) Within B (C) Within C

Figure 3: TF vs CC Charge Density Over pH. The difference in charge density between the bound and unbound conformations of gp120 was calculated at pH 4, 5, 7, and 8. TF and CC were compared overall (A), and within each clade (B & C)



(A) Overall pH 4 & 7 (B) Overall pH 5 & 8 (C) Overall Average pH

(D) Within B pH 4 & 7 (E) Within B pH 5 & 8 (F) Within B Average pH

(G) Within C pH 4 & 7 (H) Within C pH 5 & 8 (I) Within C Average pH

Figure 4: TF vs CC pH Sensitivity of Charge Density. The pH sensitivity of the charge density was calculated in three ways: the difference in charge density at pH 4 subtracted from the difference in charge density at pH 7 (A, D & G), the difference in charge density at pH 5 subtracted from the difference in charge density at pH 8 (B, E & H), and the average of the previous two calculations (C, F & I). TF and CC were compared overall (A-C), and within each clade (D-I). (*$p<0.05$, **$p<0.01$, ***$p<0.001$)

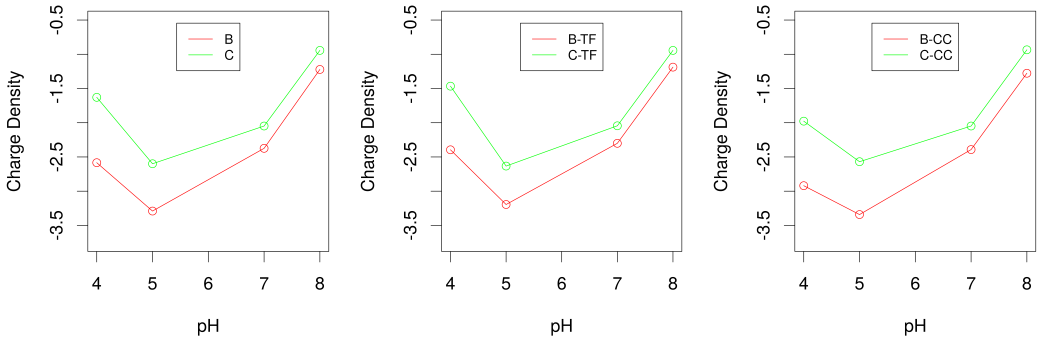
gp120-CD4 Binding Energy

Bound gp120 Conformation

The typical approach for using APBS [5] to calculate binding energy is to calculate the difference of the total electrostatic energy between the complex of interest and the individual chains from which the complex is composed. The conformation of the chains in the complex and in the separate files are identical. The required structures were produced as shown in Figures 2F and 2G.

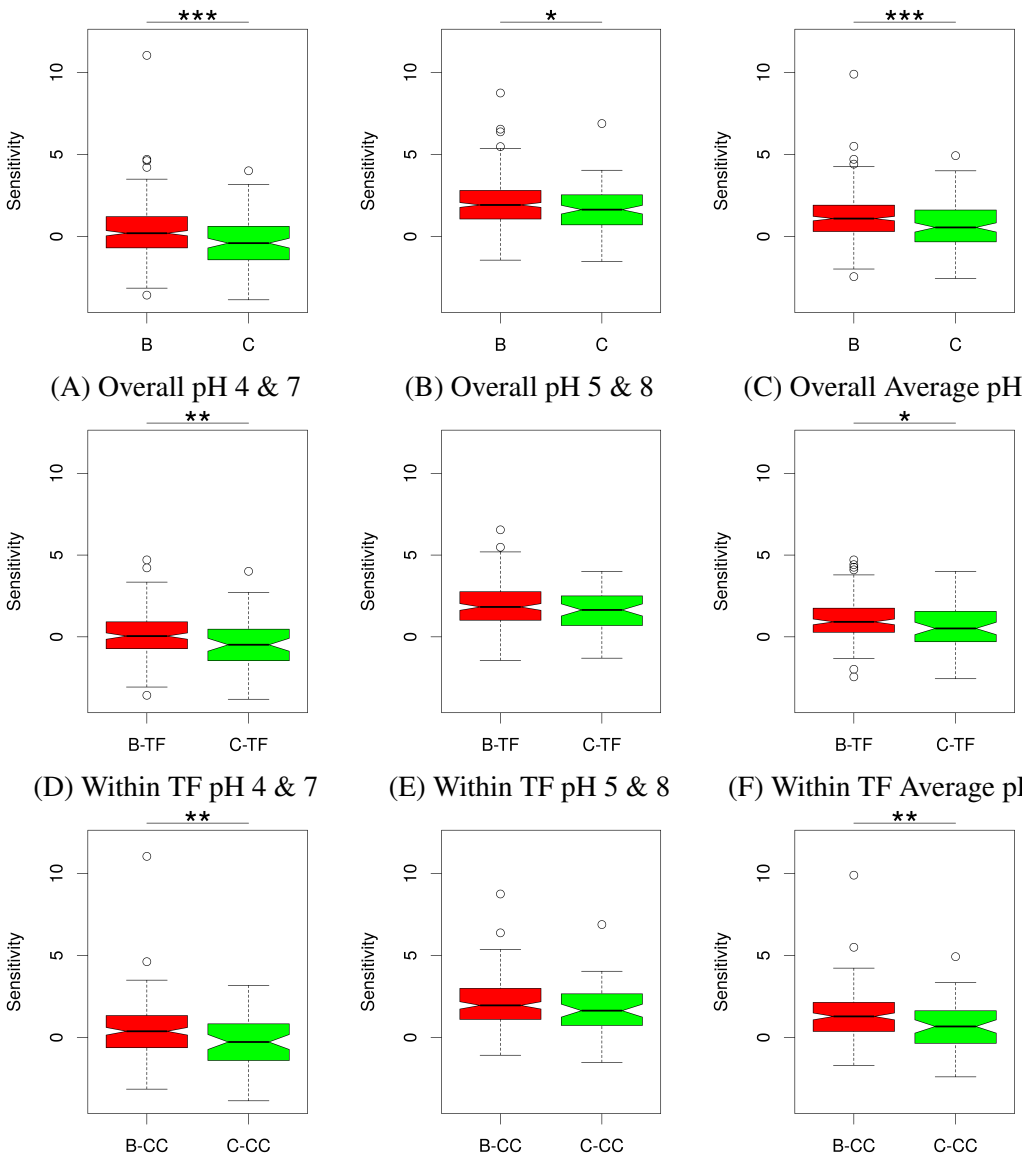
Binding energies were compared between TF and CC strains, and TF strains appeared to bind CD4 better than CC strains at pH values between approximately 3.5 and 6.5 (Figure 7A). This suggests that TF strains bind CD4 better at low pH when compared to CC strains. The results were similar within clades B (Figure 8A) and C (Figure 8B), though the latter is less consistent.

B clade appears to bind CD4 more strongly at pH values between approximately 4 and 4.5 (Figure 7B). When analyzed within classes TF (Figure 9A) and CC (Figure 9B), the trend is only found within TF.



(A) Overall (B) Within TF (C) Within CC

Figure 5: B vs C Charge Density Over pH. The difference in charge density between the bound and unbound conformations of gp120 was calculated at pH 4, 5, 7, and 8. B and C were compared overall (A), and within each class (B & C)



(A) Overall pH 4 & 7 (B) Overall pH 5 & 8 (C) Overall Average pH

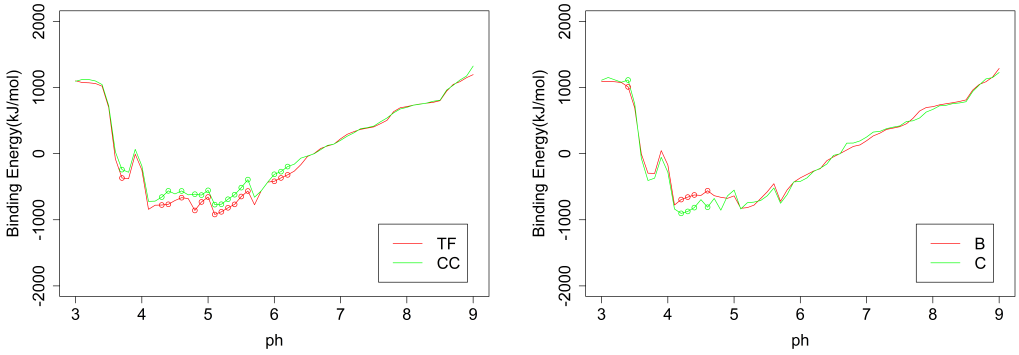
(D) Within TF pH 4 & 7 (E) Within TF pH 5 & 8 (F) Within TF Average pH

(G) Within CC pH 4 & 7 (H) Within CC pH 5 & 8 (I) Within CC Average pH

Figure 6: B vs C pH Sensitivity of Charge Density. The pH sensitivity of the charge density was calculated in three ways: the difference in charge density at pH 4 subtracted from the difference in charge density at pH 7 (A, D & G), the difference in charge density at pH 5 subtracted from the difference in charge density at pH 8 (B, E & H), and the average of the previous two calculations (C, F & I). B and C were compared overall (A-C), and within each class (D-I). (*p<0.05, **p<0.01, ***p<0.001)

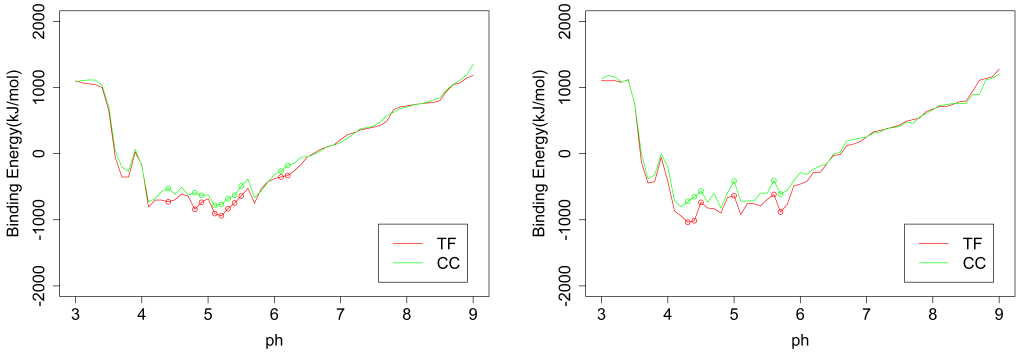
Unbound gp120 Conformation

While using the bound conformation of gp120 to determine the electrostatic energy of free gp120 corresponds with the typical approach to APBS [5] binding energy calculations,



(A) TF vs CC (B) B vs C

Figure 7: Overall Binding Energy Using gp120 Bound Conformation. Values below zero indicate a favorable binding interaction between gp120 and CD4, and more negative values indicate stronger binding. Points on the lines indicate statistically significant differences in binding energy at the given pH with p $<$0.05. Comparisons were made between TF and CC classes (A) and between B and C clades (B)



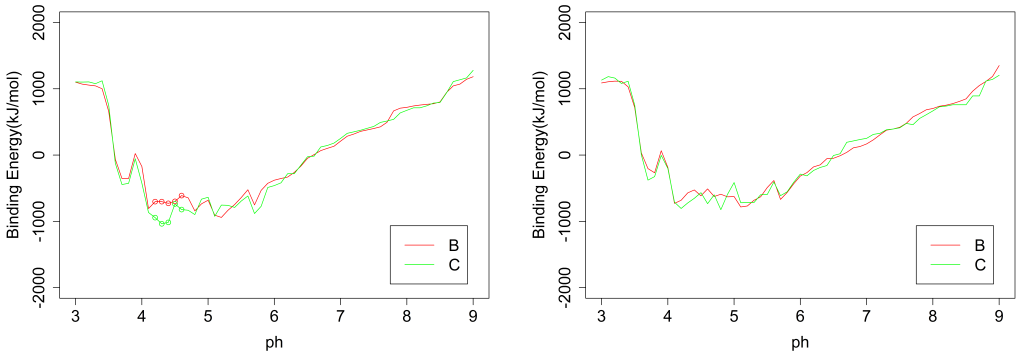
(A) Within B (B) Within C

Figure 8: TF vs CC Binding Energy Within Clades Using gp120 Bound Conformation. Values below zero indicate a favorable binding interaction between gp120 and CD4, and more negative values indicate stronger binding. Points on the lines indicate statistically significant differences in binding energy at the given pH with p $<$0.05. Comparisons between TF and CC were made within B clade (A) and within C clade (B)

it ignores the energy contribution of the conformational transition from the unbound state to the bound state. To overcome this limitation, we replaced the bound gp120 energy calculation with an unbound gp120 energy calculation; the unbound gp120 structure was produced as indicated in Figures 2H, 2I, and 2J-4.

Binding energies were compared between TF and CC strains, and CC strains appeared to bind better at higher pH conditions, though this was only significant from pH 7.9 to 8.1 (Figure 10A). Within B clade, CC strains bound CD4 significantly better than TF strains at pH values between 6.7 and 8.1 (Figure 11A). There was no significant difference within C clade (Figure 11B).

C clade sequences bound CD4 significantly better than B clade sequences pH at 3.6 and 3.7 (Figure 10B); this was consistent at pH 3.7 within the CC class (Figure 12B). Within the TF class, there is a trend of B binding CD4 better at approximately pH 6.5 and above, but the significant difference was found at pH 9 (Figure 12A).



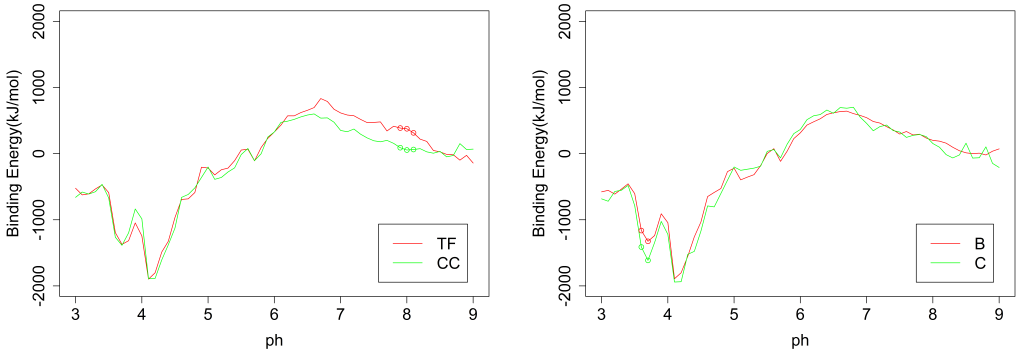
(A) Within TF

(B) Within CC

Figure 9: B vs C Binding Energy Within Classes Using gp120 Bound Conformation. Values below zero indicate a favorable binding interaction between gp120 and CD4, and more negative values indicate stronger binding. Points on the lines indicate statistically significant differences in binding energy at the given pH with $p < 0.05$. Comparisons between B and C were made within the TF class (A) and within the CC class (B)

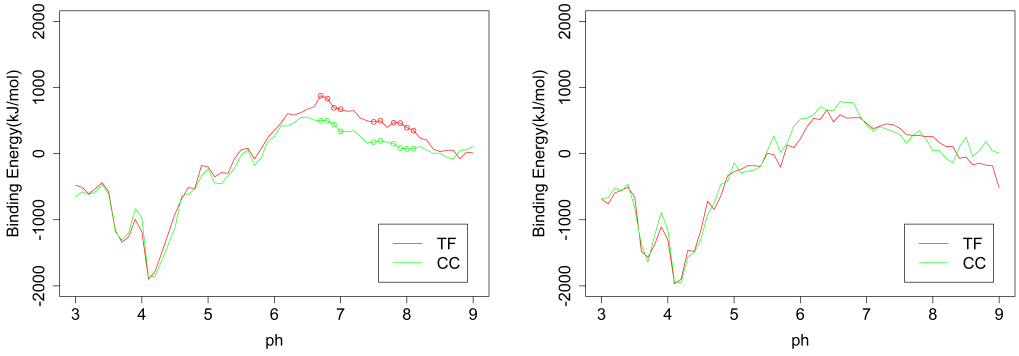
Difference Between Bound and Unbound gp120 Conformation

To determine the portion of the binding energy due to the conformational shift from unbound gp120 to bound gp120, we determined the difference in the two previously calculated



(A) TF vs CC (B) B vs C

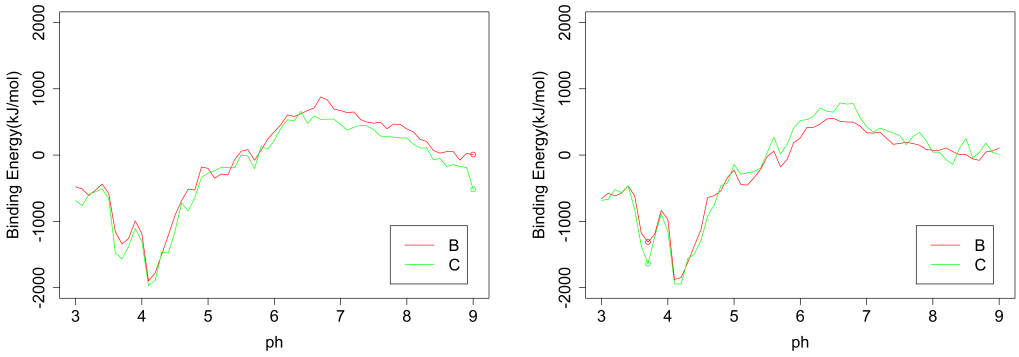
Figure 10: Overall Binding Energy Using gp120 Unbound Conformation. Values below zero indicate a favorable binding interaction between gp120 and CD4, and more negative values indicate stronger binding. Points on the lines indicate statistically significant differences in binding energy at the given pH with p $<$0.05. Comparisons were made between TF and CC classes (A) and between B and C clades (B)



(A) Within B (B) Within C

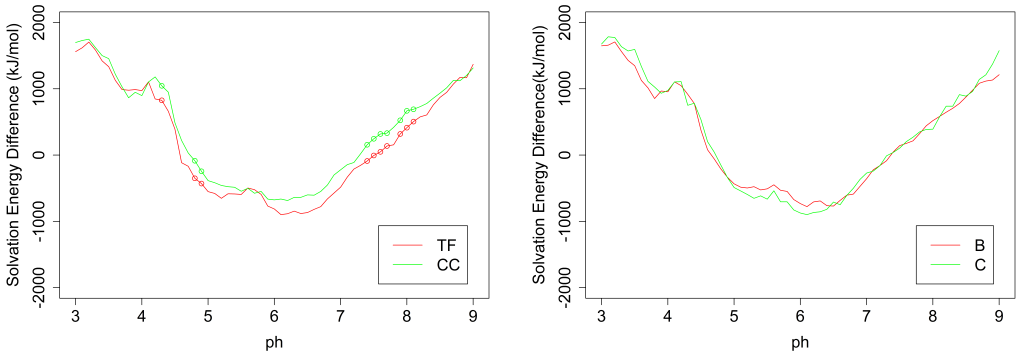
Figure 11: TF vs CC Binding Energy Within Clades Using gp120 Unbound Conformation. Values below zero indicate a favorable binding interaction between gp120 and CD4, and more negative values indicate stronger binding. Points on the lines indicate statistically significant differences in binding energy at the given pH with p $<$0.05. Comparisons between TF and CC were made within B clade (A) and within C clade (B)

binding energies for each model/pH combination (Figure 13) by subtracting the unbound calculation (Figure 10) from the bound calculations (Figure 7).



(A) Within TF (B) Within CC

Figure 12: B vs C Binding Energy Within Classes Using gp120 Unbound Conformation. Values below zero indicate a favorable binding interaction between gp120 and CD4, and more negative values indicate stronger binding. Points on the lines indicate statistically significant differences in binding energy at the given pH with p <0.05. Comparisons between B and C were made within the TF class (A) and within the CC class (B)



(A) TF vs CC (B) B vs C

Figure 13: Overall Energy Difference Between Bound and Unbound Conformations. Values below zero indicate a preference for the unbound conformation. Points on the lines indicate statistically significant differences in energy at the given pH with p <0.05. Comparisons were made between TF and CC classes (A) and between B and C clades (B)

The unbound calculation is:

$$\Delta G = G_{complex} - G_{CD4} - G_{unbound}$$

and the bound calculation is:

$$\Delta G = G_{complex} - G_{CD4} - G_{bound}$$

In these equations, $G_{complex}$, G_{CD4}, $G_{unbound}$, and G_{bound} are the electrostatic energies calculated at the given model/pH combination. Therefore,

$$(bound\ calculation) - (unbound\ calculation)$$

simplifies to

$$G_{unbound} - G_{bound}$$

This eliminates the electrostatic energy contribution of complex formation and simply leaves the difference in solvation energy for the two conformations. Because a lower electrostatic energy indicates a more preferable state, a positive result of this calculation indicates a preference for the bound conformation, while a negative result indicates a preference for the unbound conformation.

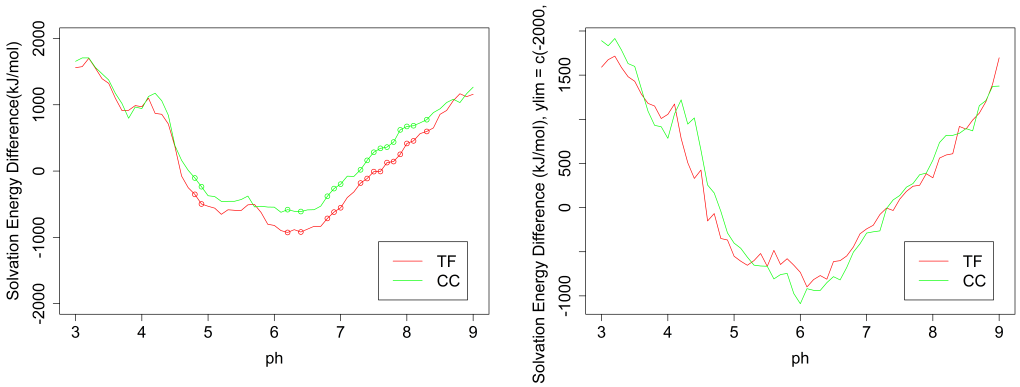
Both TF and CC classes prefer the unbound conformation approximately between pH values 4.5 and 7.5, while the bound form is preferred outside of this range. The CC class appears to be more positive than the TF class approximately between pH 4.5 and 8.5; this is significant at pH 4.2, 4.7, 4.8, and from pH 7.4 to pH 8.1 (Figure 13A). Within the B clade, this trend is significant at pH 4.7 an 4.8, as well as the majority of pH values between 6.2 and 8.3 (Figure 14A). However, the trend is mostly absent within the C clade, and there is no significant difference at any point (Figure 14B).

The B clade appears to be slightly more positive than the C clade between pH 5 and 6.3, but there are no significant differences (Figure 13B). This is also true within the CC class, which has a slightly bigger difference between clades (Figure 15B). There is also no significant difference within the TF class, though C appears to be slightly more positive than B between pH 6 and 8 (Figure 15A).

Binding Energy pH Sensitivity

Bound gp120 Conformation

The difference in binding energy between low and high pH values was used as a calculation of the pH sensitivity of this interaction. First, corresponding TF and CC bound conformation sensitivity from each individual was compared, but no consistent differences were found between corresponding sequences (Figure 16A). This was also true when sensitivity was grouped by class; the median sensitivity was greater in the TF class, but there was no significant difference (Figure 16B). This is also consistent within clades B (Figure 17A) and C (Figure 17B).



(A) Within B (B) Within C

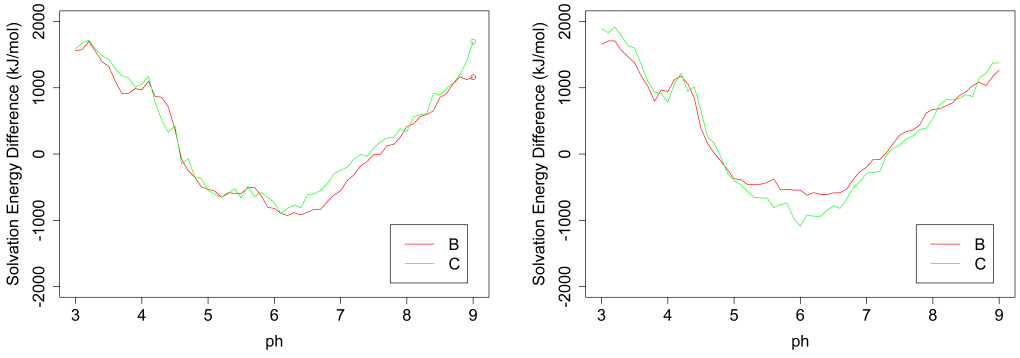
Figure 14: TF vs CC Energy Difference Between Bound and Unbound Conformations. Values below zero indicate a preference for the unbound conformation. Points on the lines indicate statistically significant differences in energy at the given pH with $p < 0.05$. Comparisons between TF and CC were made within B clade (A) and within C clade (B)

There is also no clear difference between individual B and C clade strains using the bound gp120 conformation sensitivity (Figure 16A). There is also no significant difference when sequences are pooled into their respective clades (Figure 16C). This is consistent within class TF (Figure 18A), but B clade is significantly more sensitive within class CC (Figure 18B).

Unbound gp120 Conformation

The binding sensitivity was also calculated using the unbound gp120 conformation binding energies. No consistent differences were found between corresponding sequences under these conditions, either (Figure 19A). However, when binding energy sensitivity data were grouped into classes, CD4 binding in TF strains was found to be more sensitive to pH than in CC strains (Figure 19B). This was consistent within B clade (Figure 20A), but not within C clade (Figure 20B).

There was still no significant difference in pH sensitivity of CD4 binding between overall B and C clades (Figure 19C). This was consistent within the TF class (Figure 21A). C clade



(A) Within TF

(B) Within CC

Figure 15: B vs C Energy Difference Between Bound and Unbound Conformations. Values below zero indicate a preference for the unbound conformation. Points on the lines indicate statistically significant differences in energy at the given pH with $p < 0.05$. Comparisons between B and C were made within the TF class (A) and within the CC class (B)

strains appear to be more sensitive than B clade within the CC class (Figure 21B). This is the opposite of the result observed when using the bound gp120 conformation.

Difference Between Bound and Unbound gp120 Conformation

The pH sensitivity of the difference between bound and unbound conformations indicates the pH sensitivity of the conformational change. There was no clear difference between individual corresponding sequences (Figure 22A). There was no significant difference between overall TF and CC strains (Figure 22B); these results were consistent within C clade (Figure 23B). However, within B clade, TF strains were significantly more positive than CC strains (Figure 23B). This indicates that the gp120 conformation shift from bound to unbound is significantly more pH sensitive in TF strains than in CC strains.

There was also no significant difference between overall B and C clades (Figure 22C), which was consistent within the TF class (Figure 24A). Within the CC class, C clade was significantly more sensitive (Figure 24B).

Residue Specific pH Sensitivity

In order to determine a mechanism of pH sensitivity of the gp120-CD4 interaction, residue specific sensitivity was determined to identify the most sensitive residues. Residue specific charge was used to calculate pH sensitivity. Several methods were attempted to produce class and clade models of residue specific pH sensitivities.

KNN Mapping of Coordinate Charges

One of the template models, 1RZK [26], was originally used as a target structure for mapping residue specific charges. Electrostatic field coordinates generated by APBS [5] for each model at pH 4, 5, 7, and 8 were used to determine specific charges for VMD produced coordinate for each model. These charges were mapped onto 1RZK model coordinates generated by VMD [27] using a KNN algorithm. The algorithm found 100 KNN for each of 68,701 coordinates in the target structure by calculating the distance from each of 140,000 to 170,000 points in each model. However, this approach proved to be intractable with 32

sequences running in parallel completing approximately every 1.5 days. This would have required approximately 90 days to complete the 40 pH specific models per each of the 48 sequences in the data set.

To overcome these limitations, the KNN mapping approach was refined to the binding interface of the models. The residues of the binding interface for each model and the target structure were identified by subtracting the solvent accessible surface area (SASA) of each gp120 residue in the CD4-bound model from the bound conformation of gp120 without CD4 included in the model. Any residue with a difference in SASA greater than zero was considered to be part of the binding interface. Figure 25 shows an example of the SASA differences determined for a single sequence model set. This greatly reduced the number of coordinates to between 12,000 and 25,000 for the models and 10,506 for the target. This allowed for all models to be mapped within a week using 100 KNN.

The KolmogorovSmirnov (KS) statistic between the fitted template model and the original models was used to determine the optimal number of nearest neighbors from 1 to 100 26. For all models, as the value of K increased, the KS statistic also increased. This indicates that the model fit became worse as the value of K increased. It was unlikely that a single nearest neighbor based mapping would produce a strong fit, and this approach ignored any pH sensitivity outside of the binding region. Consequently, an alternative approach was explored.

Sequence Alignment Mapping

To ensure that all residues were consistent and residues were matched appropriately, a sequence alignment was used to map residue specific sensitivity to particular positions. Because of the differences observed in the overall charge density at different pH values (Figures 3 & 3), sensitives were calculated using only pH 4 and 7, only pH 5 and 8, and all 4 pH values together. Clustal Omega [53, 23, 39] was used to align all sequences in the data set along with 1RZK [26] (Appendix B). The median sensitivity of the residues at each

Table 1: TF vs CC Top Residue Positions. Residue positions were chosen if the absolute value of the charge at that position was greater than 1% of the sum of the absolute charge of every position.

pH Values	Overall	Within B Clade	Within C Clade
7 & 4	343, 348, 376, 406, 409, 413	285, 290, 335, 338, 343, 387, 402, 427	199, 253, 255, 262, 299, 348, 390, 391, 397, 453
8 & 5	289, 342, 343, 348, 358, 359, 404, 405, 406, 407, 413	285, 289, 342, 343, 402, 404, 405, 407, 413	254, 280, 299, 314, 348, 358, 359, 391
Average	263, 289, 301, 343, 348, 349, 403, 404, 406, 409, 414	263, 285, 343, 387, 395, 402, 409, 414, 427	240, 255, 289, 348, 349, 390, 391, 397, 449, 455, 494

position in the alignment were determined for TF and CC classes overall and within each clade, and for B and C clades overall and within each class. A gap at a position did not contribute to the median of the position, and a sensitivity of zero was used if all sequences within a group had a gap at a position.

TF and CC classes were compared by subtracting CC sensitivity from TF sensitivity at each position in the alignment (Figure 27A). This was also compared within B clade (Figure 27B) and C clade (Figure 27C). Positive values indicate greater sensitivity in TF strains, while negative values indicate greater sensitivity in CC strains. There was a greater number of residues that were more sensitive in CC than in TF under all conditions. These general patterns remained consistent when considering only pH 4 and 7, or only pH 5 and 8 (Appendix C.1 & C.2). The top 1% of residues were determined based upon absolute sensitivity difference (Table 1).

B and C clades were compared by subtracting C sensitivity from B sensitivity at each position in the alignment (Figure 28A). This was also compared within classes TF (Figure

Table 2: B vs C Top Residue Positions. Residue positions were chosen if the absolute value of the charge at that position was greater than 1% of the sum of the absolute charge of every position.

pH Values	Overall	Within TF Class	Within CC Class
7 & 4	343	285, 342, 343, 402	474
8 & 5	405	285, 342,343, 402, 403, 412, 413	405
Average	342, 343, 358, 373, 409, 453	342, 343, 358, 402, 427, 453	342, 347, 358, 409, 474

28B) and CC (Figure 28C). Positive values indicate greater sensitivity in B strains, while negative values indicate greater sensitivity in C strains. These general patterns remained consistent when considering only pH 4 and 7, or only pH 5 and 8 (Appendix C.3 & C.4). There was a greater number of residues that were more sensitive in B clade than in C clade. The top 1% of residues were determined based upon absolute sensitivity difference (Table 2).

To ensure that small sample sizes from certain positions in the alignment were not biasing the results, the median charge of each position was also determined with gaps contributing a value of zero to a position. The overall trend remained the same between TF and CC classes (Figure 29 and Appendix D.1 & D.2). The top 1% of residues were determined based upon absolute sensitivity difference (Table 3).

The overall trend also remained the same between B and C clades (Figure 30 and Appendix D.3 & D.4). The top 1% of residues were determined based upon absolute sensitivity difference (Table 4).

Sequence Comparison

Identified pH sensitive alignment positions were analyzed for sequence composition using R 2.38.4 [48], and the package RWebLogo [58]. While some sequence differences

Table 3: TF vs CC Top Residue Positions Considering Gaps. Residue positions were chosen if the absolute value of the charge at that position was greater than 1% of the sum of the absolute charge of every position.

pH Values	Overall	Within B Clade	Within C Clade
7 & 4	255, 263, 299, 376, 391	263, 285, 290, 335, 338, 339, 387, 427	199, 243, 253, 254, 255, 262, 299, 314, 375, 390, 391, 453
8 & 5	289, 299, 338	207, 285, 289, 290, 291, 338, 387, 419, 427	254, 255, 280, 289, 299, 314, 375, 390, 458
Average	243, 263, 289, 301, 320, 322, 395, 455	263, 285, 289, 290, 302, 322, 335, 338, 339, 387, 394, 395, 425, 427	240, 253, 255, 262, 289, 301, 315, 353, 390, 391, 449, 453, 455, 494

Table 4: B vs C Top Residue Positions Considering Gaps. Residue positions were chosen if the absolute value of the charge at that position was greater than 1% of the sum of the absolute charge of every position.

pH Values	Overall	Within TF Class	Within CC Class
7 & 4	263, 290, 373, 453	262, 285, 315, 373	263, 373
8 & 5	263	285, 373, 427	262
Average	263, 296, 373, 453, 457	262, 285, 373, 387, 427, 453	263, 338, 350

were found between TF and CC strains at the sensitive alignment positions, none of these differences indicate a clear mechanism for pH sensitivity (Figure 31). The most striking difference is at position 343; a positive lysine is found in the TF class, while a negative glutamate is found in the CC class (Figures 31A, 31B & 31C, and Appendix E.1). However, this difference is not found within the C clade (Appendix E.2), and the bit score is low under the conditions in which it is found; this is because this is an insertion that only occurs within the TF and CC forms of the WEAU sequence (Appendix B).

There was also no clear insight into the mechanism for the difference pH sensitivity between B and C clades (Figure 32 and Appendix E.3 & E.4). There were low bit scores (Position 350 in Appendix E.3C & E.4C) and the differences were often small changes in the ratio of similar amino acids (Position 427 in Appendix E.3B & E.4B).

These problems persisted when gaps were considered in the identification of sensitive residues (Appendix F.1, F.3, F.4, F.2, F.5, & F.6).

Mapping Sequences to Structures

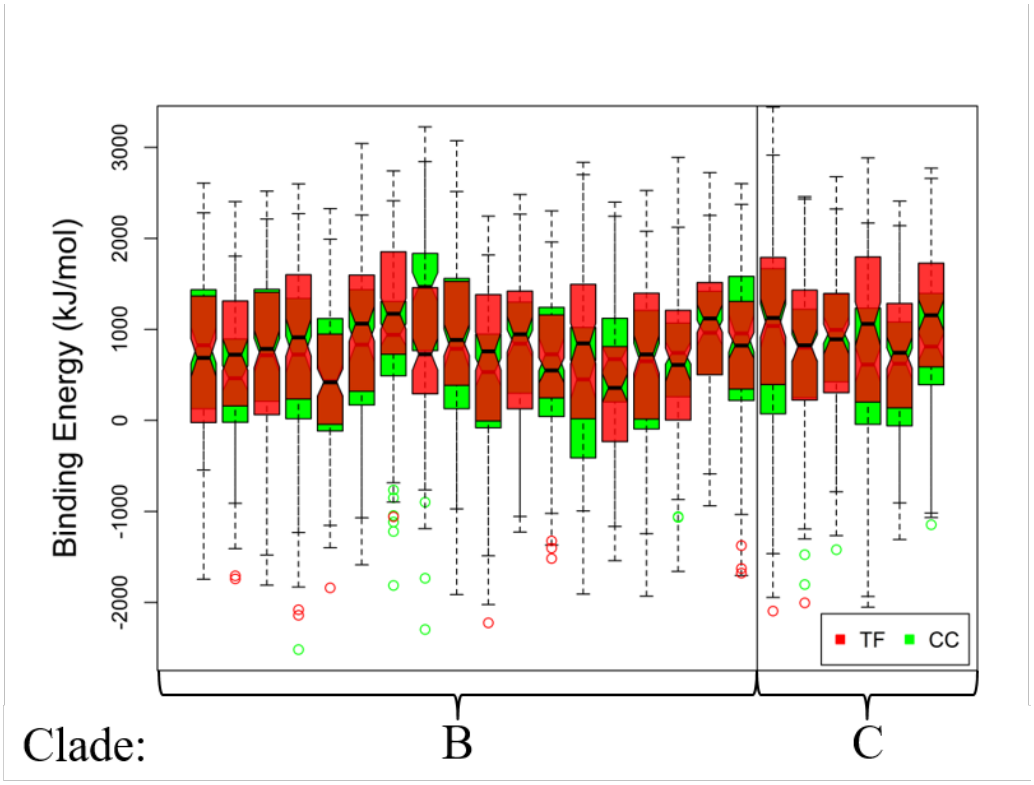
The sensitivities of the selected residues were mapped onto a single model of the EU744010 sequence because it was the longest sequence analyzed (503 AA). The binding interface was also mapped onto this structure to determine if any of the identified residues likely interacts with CD4 (Figure 33). None of the residues identified within the TF/CC or B/C groups was a member of the CD4 binding interface under any conditions tested (Figures 34, 35 & 36). However, many of the identified residues in the TF/CC comparisons were near the binding interface, so they could indirectly affect the gp120-CD4 interaction; these residues were marked with a yellow arrow and the alignment position was indicated (Figures 34 & 35). Interestingly, most of the residues were found to be more pH sensitive in CC strains.

When comparing B and C clades, there were much fewer sensitive residues exposed near the binding surface (Figures 34 & 36). Within each the TF class, C clade had considerably more pH sensitive residues exposed near the binding site than B clade, and the most significant residues within class CC were not present on the model (Figure 36).

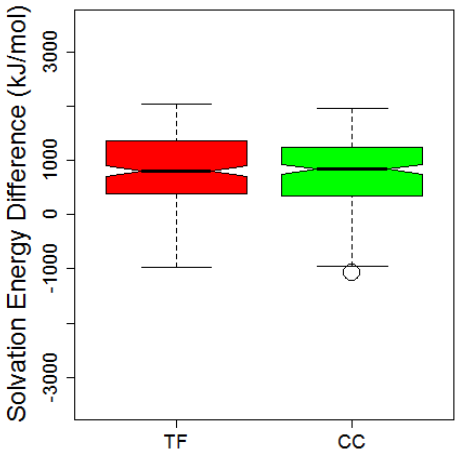
The sensitivities calculated with gaps considered were also mapped onto this gp120 structure. As with the previous mapping, no binding interface residues were found among the identified sensitive positions. The TF/CC comparison found fewer residues located near the binding interface than with the previous sensitivities. Residue 391 was found under many conditions in both sensitivity sets, and occurred primarily within C clade (Figures

35B, 35D, 35F, 37A, 38B & 38F) Additionally, CC strains were found to have a greater number of increased sensitivity residues.

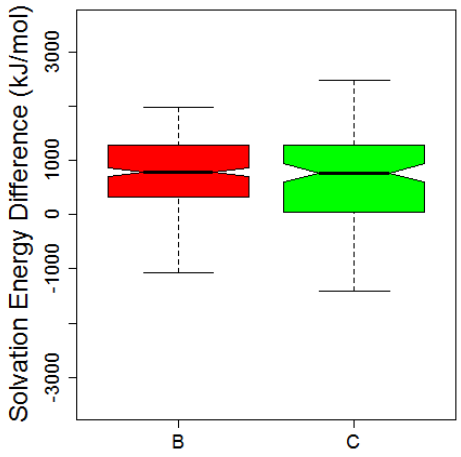
When comparing B and C clades, the residues missing from the model were no longer significant, and the majority of significant residues were more sensitive in B clade (Figures 37 & 39).



(A) Individual Sequences

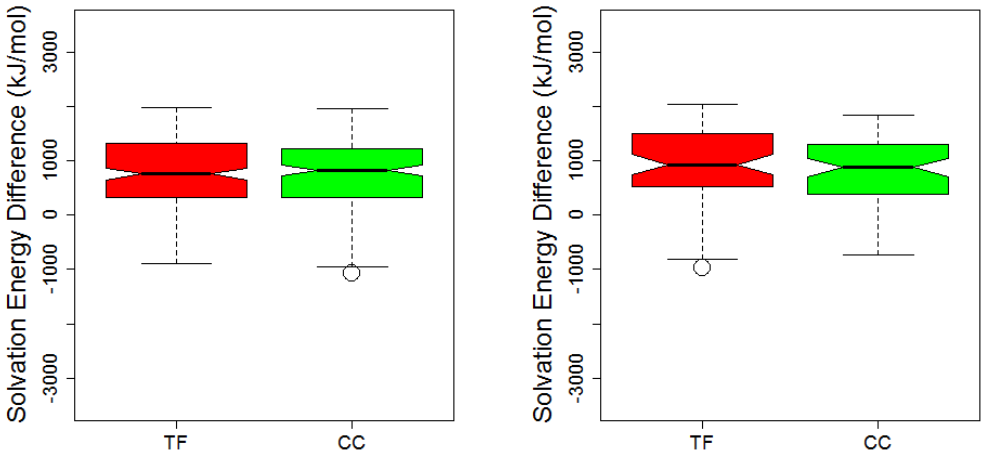


(B) TF vs CC



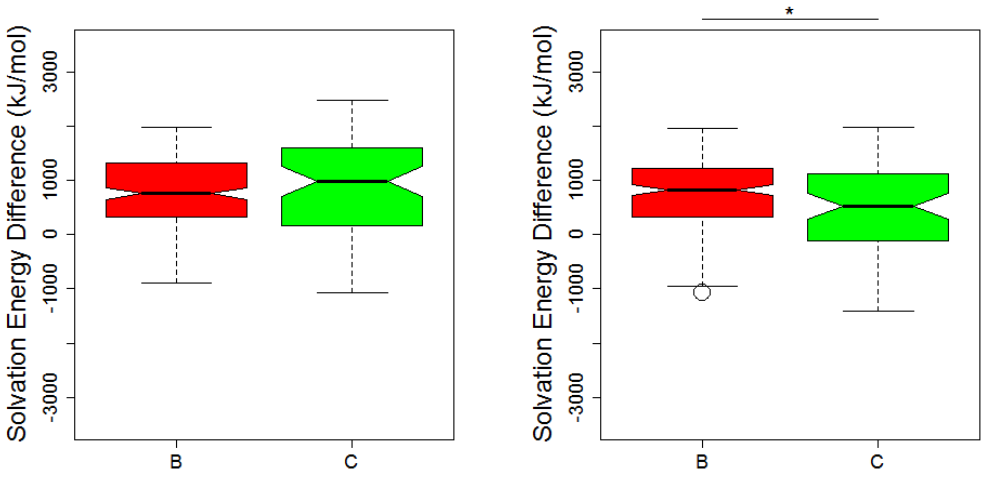
(C) B vs C

Figure 16: pH Sensitivity of Binding Energy Using gp120 Bound Conformation. Sensitivity was calculated as the difference in binding energy between low and high pH conditions. Comparisons were made between individual sequences (A), between TF and CC (B) and between B and C (C)



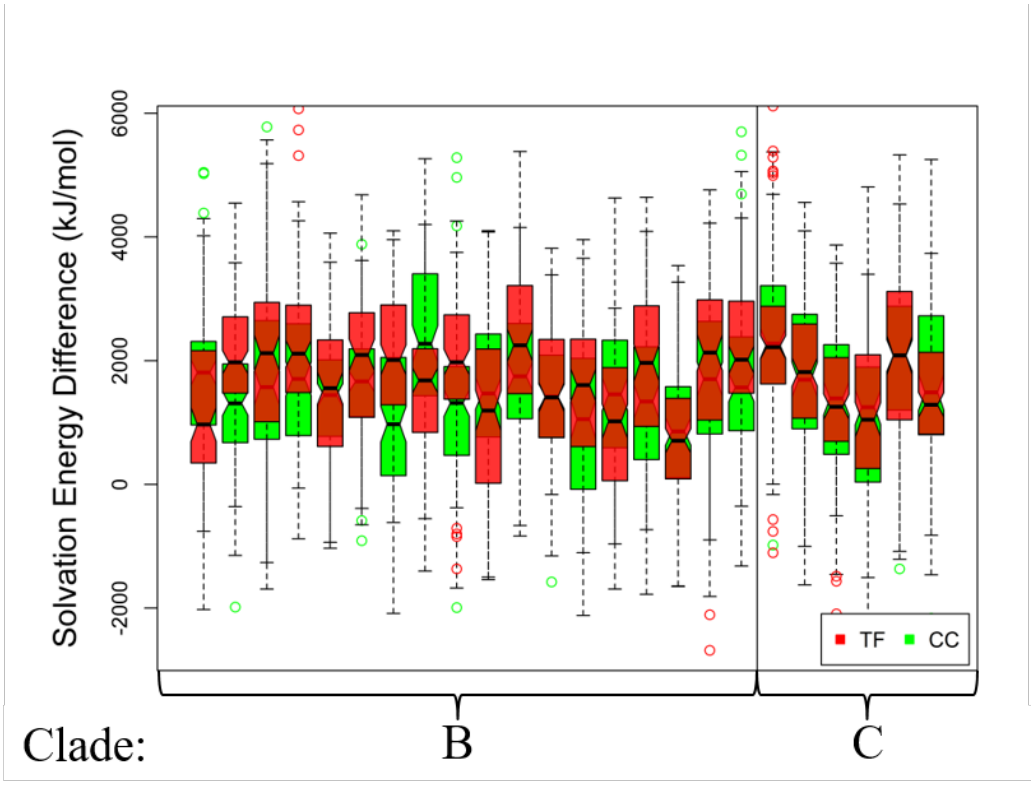
(A) Within B (B) Within C

Figure 17: TF vs CC pH Sensitivity of Binding Energy Within Clades Using gp120 Bound Conformation. Sensitivity was calculated as the difference in binding energy between low and high pH conditions. TF and CC were compared within B clade (A) and within C clade (B)

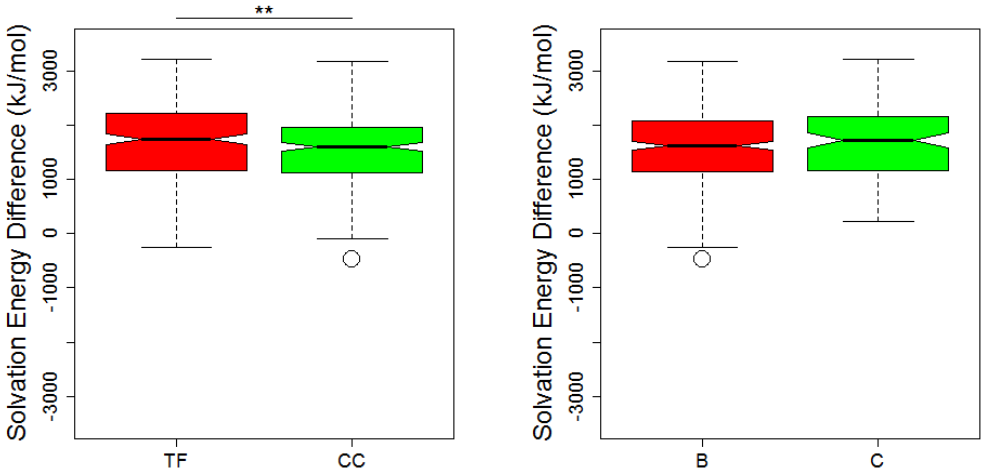


(A) Within TF (B) Within CC

Figure 18: B vs C pH Sensitivity of Binding Energy Within Classes Using gp120 Bound Conformation. Sensitivity was calculated as the difference in binding energy between low and high pH conditions. B and C clades were compared within the TF class (A) and within the CC class (B)



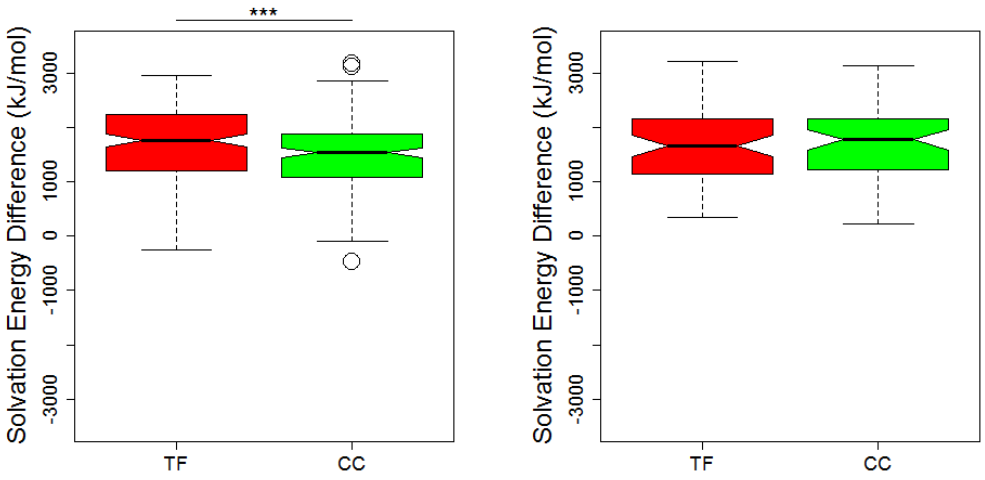
(A) Individual Sequence



(B) TF vs CC Sensitivity

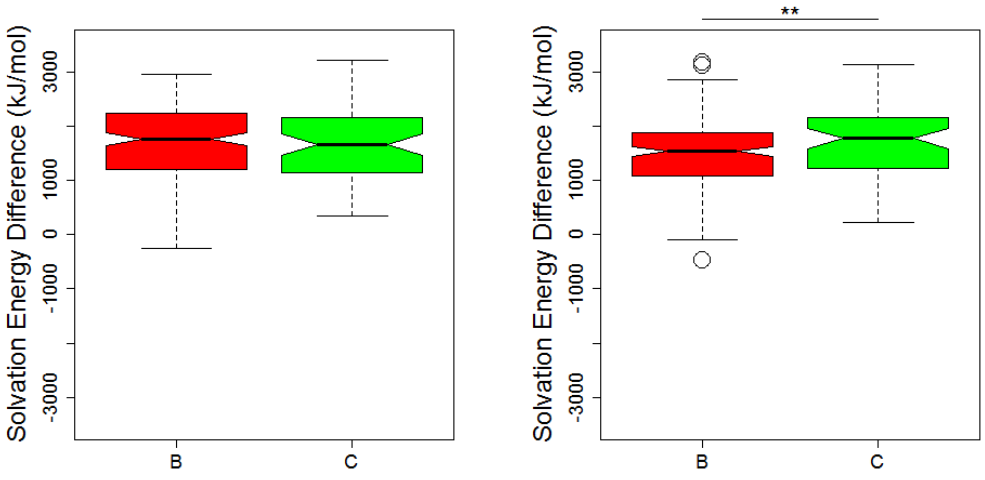
(C) B vs C Sensitivity

Figure 19: pH Sensitivity of Binding Energy Using gp120 Unbound Conformation. Sensitivity was calculated as the difference in binding energy between low and high pH conditions. Comparisons were made between individual sequences (A), between TF and CC (B) and between B and C (C) (***p<0.001)



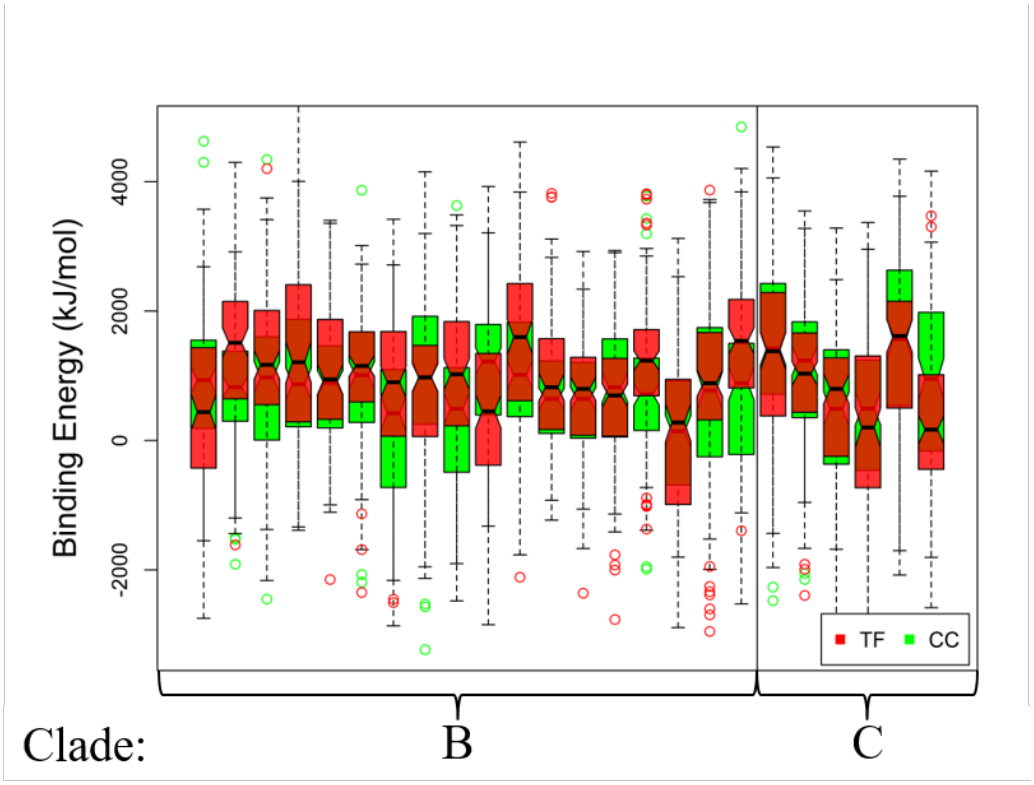
(A) Within B (B) Within C

Figure 20: TF vs CC pH Sensitivity of Binding Energy Within Clades Using gp120 Unbound Conformation. Sensitivity was calculated as the difference in binding energy between low and high pH conditions. TF and CC were compared within B clade (A) and within C clade (B) (**p<0.01)

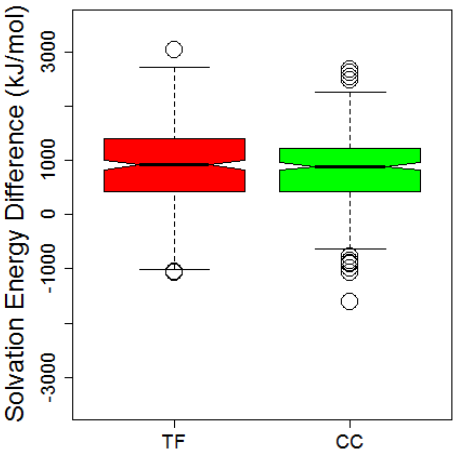


(A) Within TF (B) Within CC

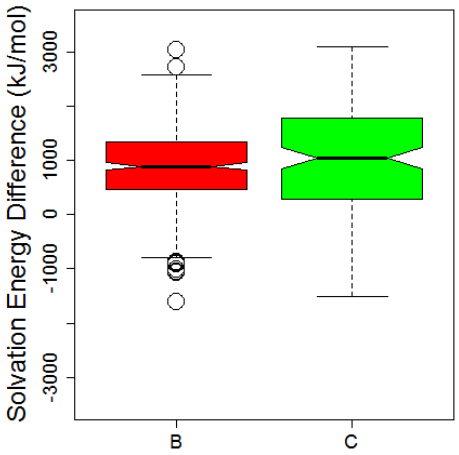
Figure 21: B vs C pH Sensitivity of Binding Energy Within Classes Using gp120 Unbound Conformation. Sensitivity was calculated as the difference in binding energy between low and high pH conditions. B and C clades were compared within the TF class (A) and within the CC class (B)



(A) Individual Sequence

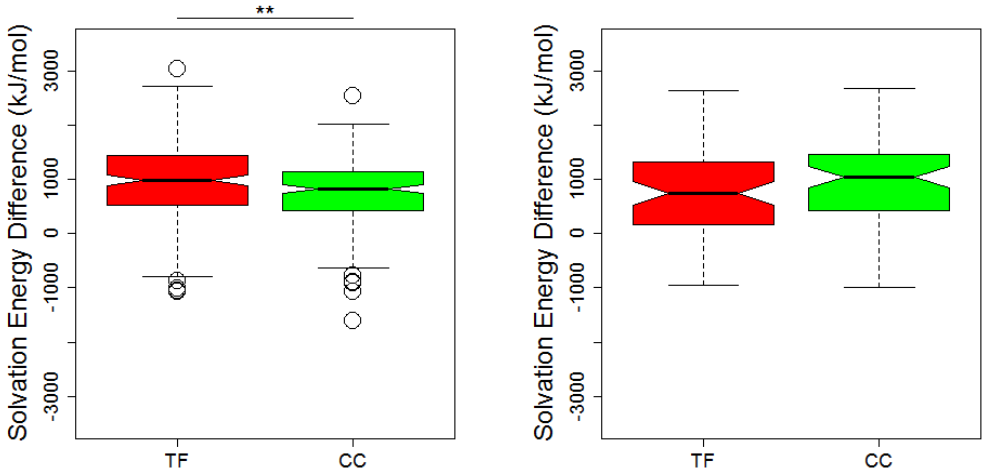


(B) TF vs CC



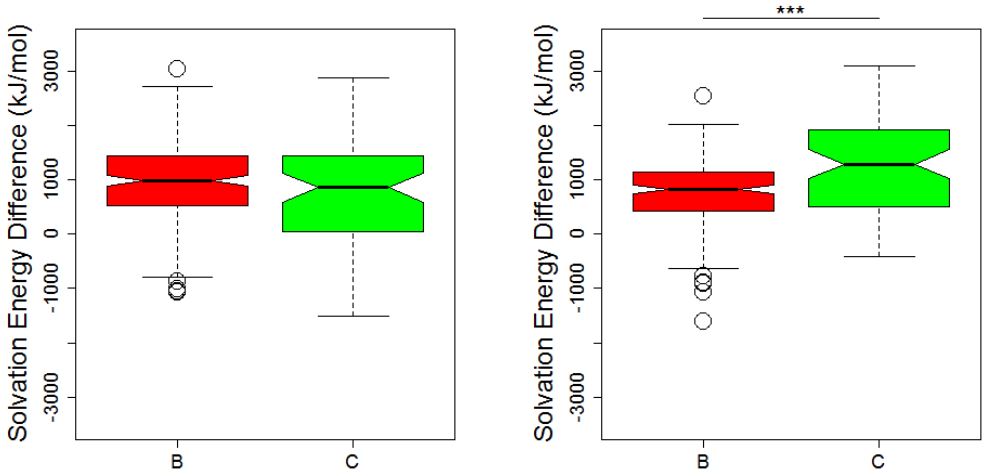
(C) B vs C

Figure 22: Overall pH Sensitivity of Energy Difference Between Bound and Unbound Conformations. Values above zero indicate a shift toward a preference for the bound conformation when pH is shifted from high to low. Sensitivity was calculated as the difference in energy between low and high pH conditions. Comparisons were made between individual sequences (A), between TF and CC (B) and between B and C (C)



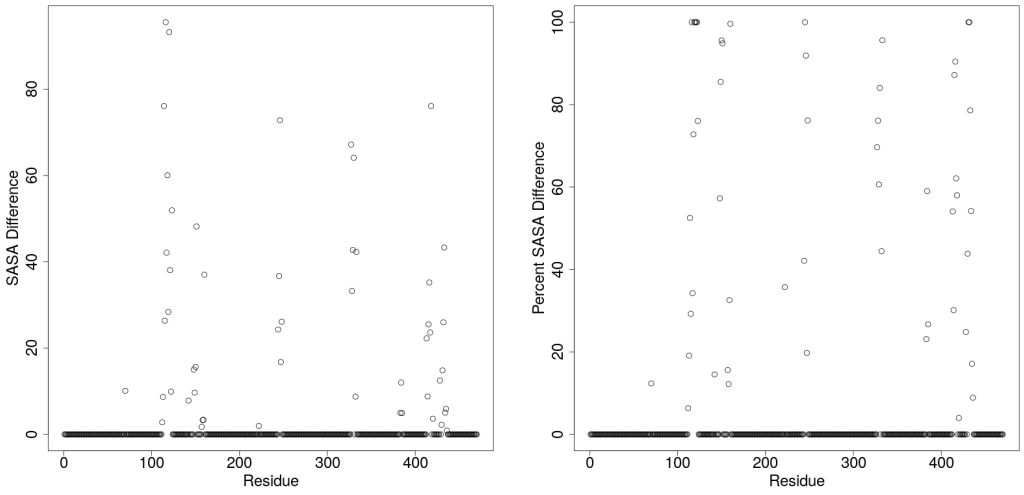
(A) Within B (B) Within C

Figure 23: TF vs CC pH Sensitivity of Energy Difference Between Bound and Unbound Conformations. Values above zero indicate a shift toward a preference for the bound conformation when pH is shifted from high to low. Sensitivity was calculated as the difference in energy between low and high pH conditions. TF and CC were compared within B clade (A) and within C clade (B)



(A) Within TF (B) Within CC

Figure 24: B vs C pH Sensitivity of Energy Difference Between Bound and Unbound Conformations. Values above zero indicate a shift toward a preference for the bound conformation when pH is shifted from high to low. Sensitivity was calculated as the difference in energy between low and high pH conditions. TF and CC were compared within the TF class (A) and within the CC class (B)



(A) Example SASA Difference (B) Example Percent SASA Difference

Figure 25: Binding Interface Residue Identification Example. A) Difference between solvent accessible surface area of each residue before and after ligand is bound. B) Percentage of solvent accessible surface area that remains after ligand is bound.

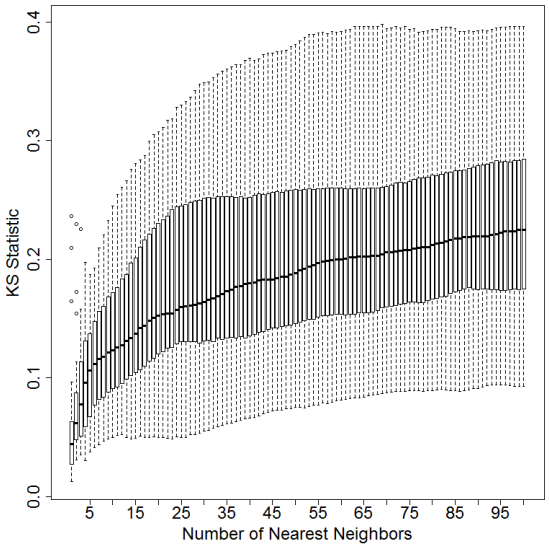
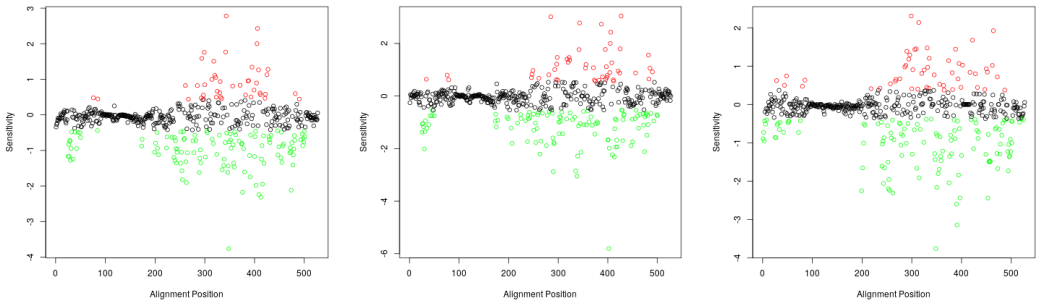
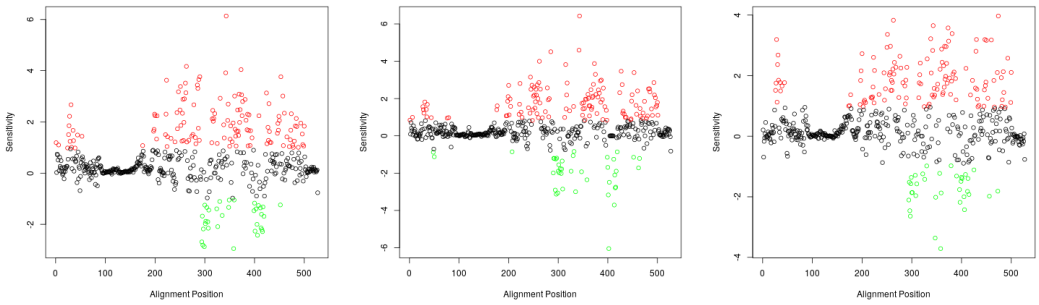


Figure 26: Analysis of KNN Mapping Using KS Statistic. The KS statistic was calculated between the charge distribution in the mapping of each sequence and the original model charge distribution for each value of K from 1 to 100, inclusive.



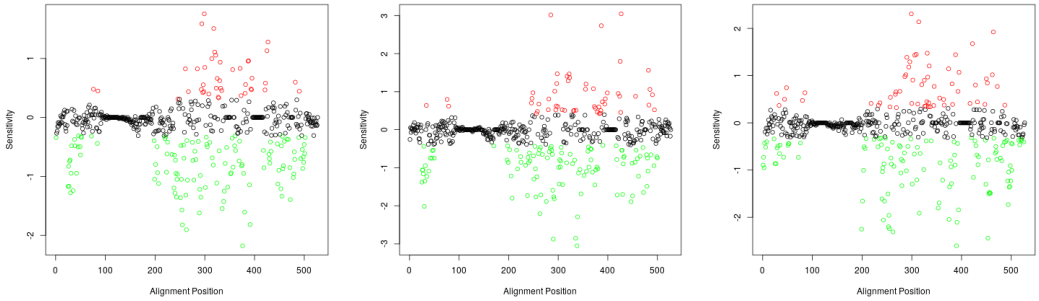
(A) Overall (B) Within B (C) Within C

Figure 27: TF vs CC Relative Residue Specific pH Sensitivity. Median charges of residues at each position in the alignment were computed. Sensitivity was calculated as the average of the difference between the charge at pH 4 subtracted from the charge at pH 7 and the charge at pH 5 subtracted from the charge at pH 8. CC charges were subtracted from TF charges. Values above zero indicate greater sensitivity in TF strains, while values below zero indicate greater sensitivity in CC strains. Red points are greater than one interquartile range above zero, and green points are below one interquartile range below zero.



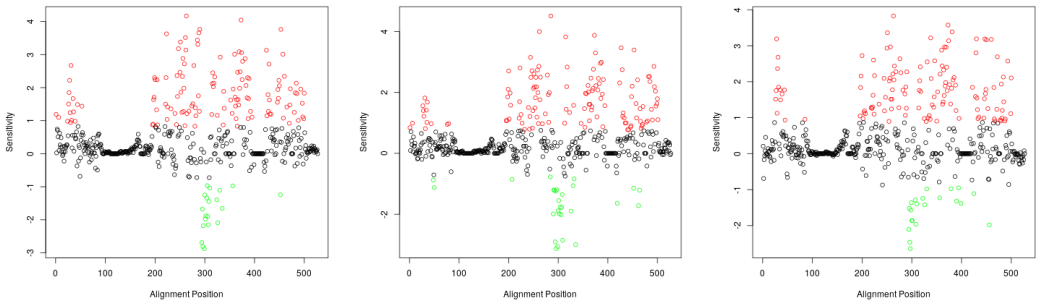
(A) Overall (B) Within TF (C) Within CC

Figure 28: B vs C Relative Residue Specific pH Sensitivity. Median charges of residues at each position in the alignment were computed. Sensitivity was calculated as the average of the difference between the charge at pH 4 subtracted from the charge at pH 7 and the charge at pH 5 subtracted from the charge at pH 8. C charges were subtracted from B charges. Values above zero indicate greater sensitivity in B strains, while values below zero indicate greater sensitivity in C strains. Red points are greater than one interquartile range above zero, and green points are below one interquartile range below zero



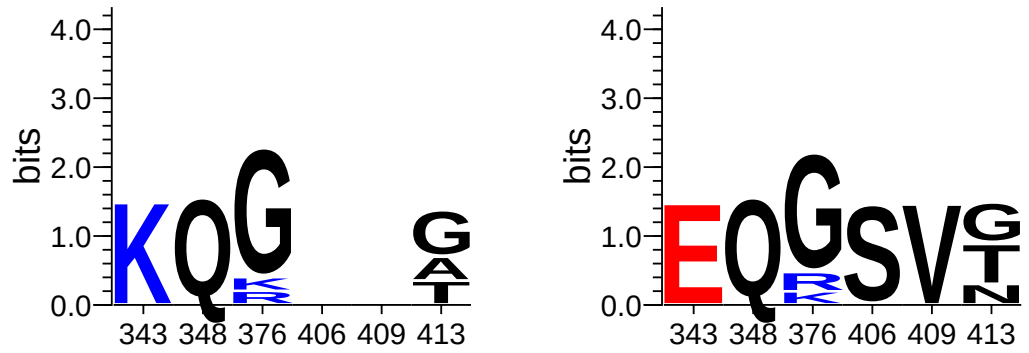
(A) Overall (B) Within B (C) Within C

Figure 29: TF vs CC Relative Residue Specific pH Sensitivity Considering Gaps. Median charges of residues at each position in the alignment were computed with gaps being considered a charge value of zero. Sensitivity was calculated as the average of the difference between the charge at pH 4 subtracted from the charge at pH 7 and the charge at pH 5 subtracted from the charge at pH 8. CC charges were subtracted from TF charges. Values above zero indicate greater sensitivity in TF strains, while values below zero indicate greater sensitivity in CC strains. Red points are greater than one interquartile range above zero, and green points are below one interquartile range below zero.

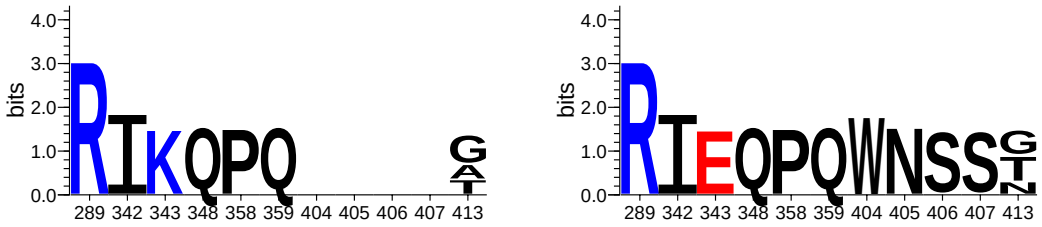


(A) Overall (B) Within TF (C) Within CC

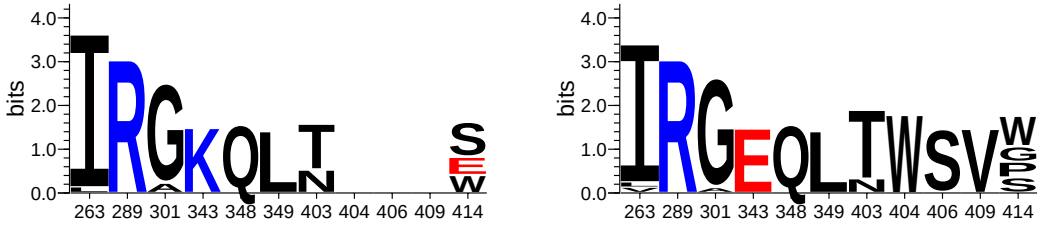
Figure 30: B vs C Relative Residue Specific pH Sensitivity Considering Gaps. Median charges of residues at each position in the alignment were computed with gaps being considered a charge value of zero. Sensitivity was calculated as the average of the difference between the charge at pH 4 subtracted from the charge at pH 7 and the charge at pH 5 subtracted from the charge at pH 8. C charges were subtracted from B charges. Values above zero indicate greater sensitivity in B strains, while values below zero indicate greater sensitivity in C strains. Red points are greater than one interquartile range above zero, and green points are below one interquartile range below zero.



(A) Average pH



(B) pH 4 & 7



(C) pH 5 & 8

Figure 31: TF vs CC Sensitive Residue Composition. A) Comparison of the top 1% of residues identified identified from Figure 27A. B) Comparison of the top 1% of residues identified in Figure C.1A. C) Comparison of the top 1% of residues identified in Figure C.2A.

(A) B vs C - Average pH

(B) B vs C - pH 4 & 7

(C) B vs C - pH 5 & 8

Figure 32: B vs C Sensitive Residue Composition. A) Comparison of the top 1% of residues identified identified from Figure 28A. B) Comparison of the top 1% of residues identified in Figure C.3A. C) Comparison of the top 1% of residues identified in Figure C.4A.

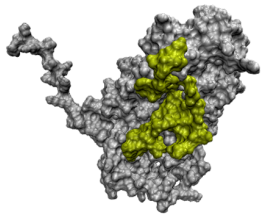
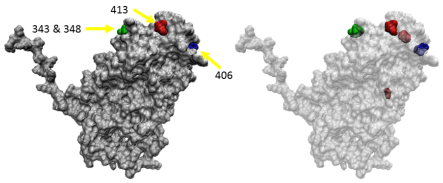
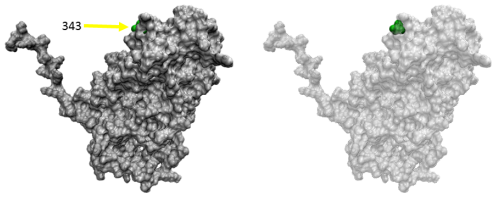


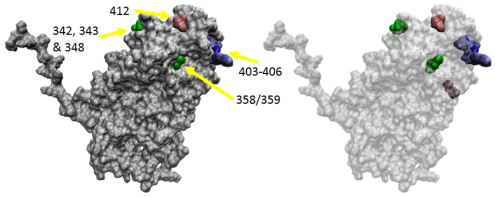
Figure 33: CD4 Binding Interface Mapped onto EU744010. The CD4 binding interface was determined as described in the KNN section of the methods. Identified binding interface residues are indicated in yellow



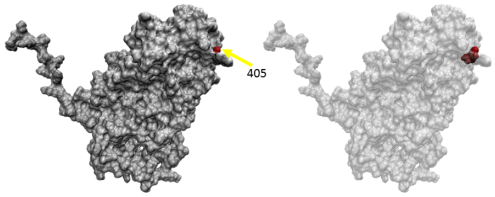
(A) TF vs CC Average pH



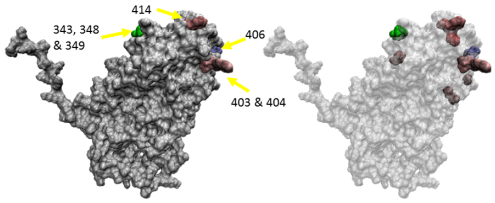
(B) B vs C Average pH



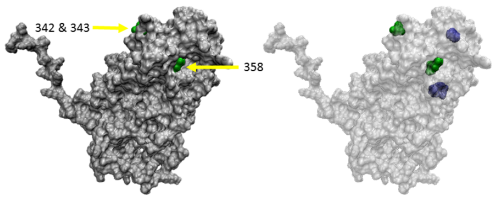
(C) TF vs CC pH 4 & 7



(D) B vs C pH 4 & 7

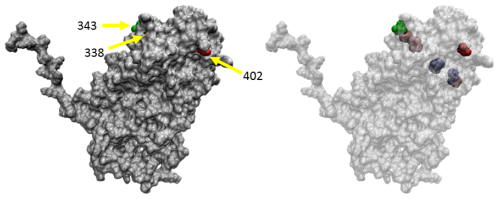


(E) TF vs CC pH 5 & 8

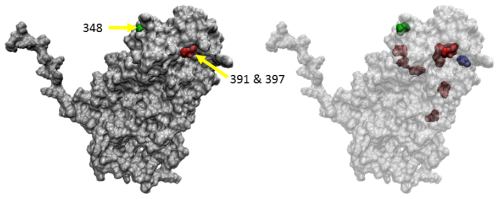


(F) B vs C pH 5 & 8

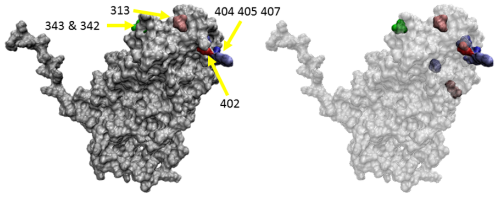
Figure 34: Structural Mapping of Residue Sensitivities for Overall Classes and Clades. A) Binding site interface residues are colored yellow. B-G) Each sub-figure contains two images of the same mapping. The left image shows a surface view of the binding interface side of the model, and the right one has all residues transparent except the selected sensitive residues. Blue indicates greater sensitivity in TF (B, D & F) or B Clade (C, E & G). Red indicates greater sensitivity in CC (B, D & F) or C Clade (C, E & G). Green identifies the general location of identified residues that are not present in the model structure; green residues are the closest residue within two positions to the missing residue. Arrows indicate residues that are exposed near the CD4 binding interface.



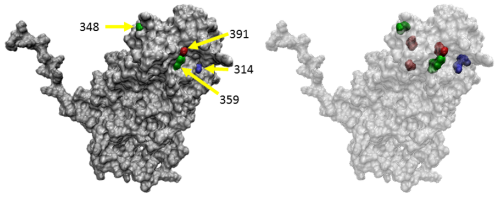
(A) Within B Clade Average pH



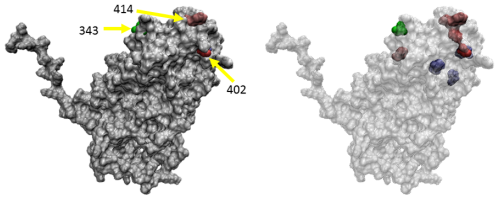
(B) Within C Clade Average pH



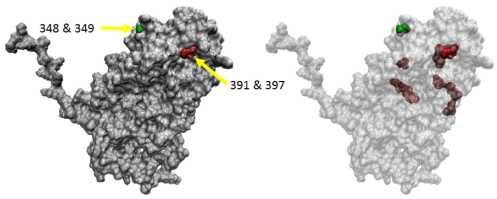
(C) Within B Clade pH 4 & 7



(D) Within C Clade pH 4 & 7

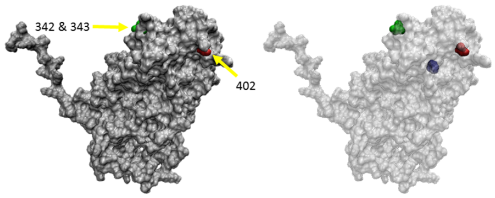


(E) Within B Clade pH 5 & 8

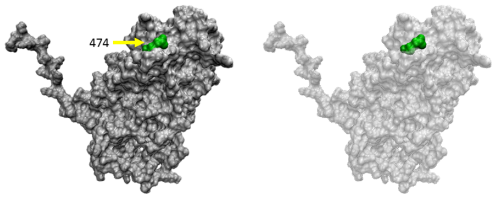


(F) Within C Clade pH 5 & 8

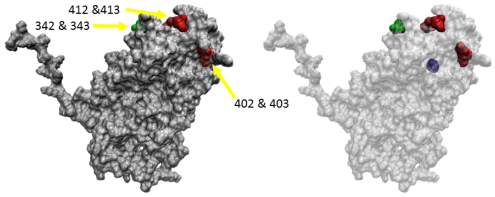
Figure 35: TF vs CC Structural Mapping of Residue Sensitivities Within Clades. Each sub-figure contains two images of the same mapping. The left image shows a surface view of the binding interface side of the model, and the right one has all residues transparent except the selected sensitive residues. Blue indicates greater sensitivity in TF (A, C & E) or B Clade (B, D & F). Red indicates greater sensitivity in CC (A, C & E) or C Clade (B, D & F). Green identifies the general location of identified residues that are not present in the model structure; green residues are the closest residue within two positions to the missing residue. Arrows indicate residues that are exposed near the CD4 binding interface.



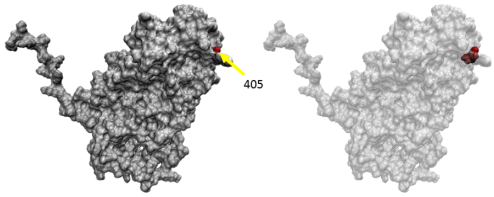
(A) Within TF Class Average pH



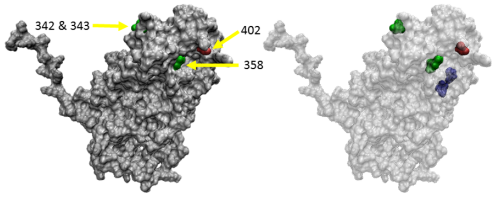
(B) Within CC Class Average pH



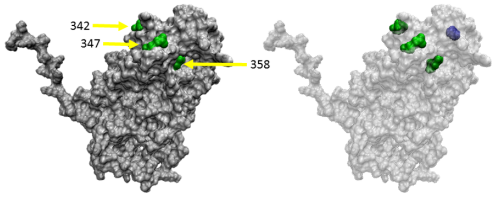
(C) Within TF Class pH 4 & 7



(D) Within CC Class pH 4 & 7

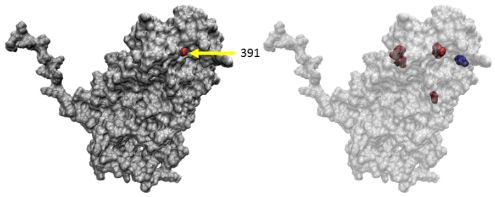


(E) Within TF Class pH 5 & 8

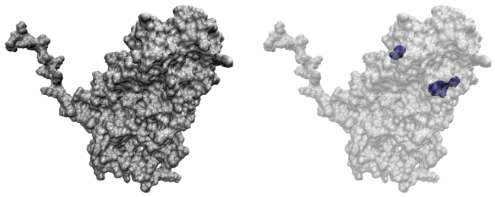


(F) Within CC Class pH 5 & 8

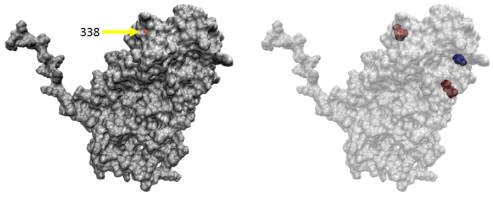
Figure 36: B vs C Structural Mapping of Residue Sensitivities Within Classes. Each subfigure contains two images of the same mapping. The left image shows a surface view of the binding interface side of the model, and the right one has all residues transparent except the selected sensitive residues. Blue indicates greater sensitivity in TF (A, C & E) or B Clade (B, D & F). Red indicates greater sensitivity in CC (A, C & E) or C Clade (B, D & F). Green identifies the general location of identified residues that are not present in the model structure; green residues are the closest residue within two positions to the missing residue. Arrows indicate residues that are exposed near the CD4 binding interface.



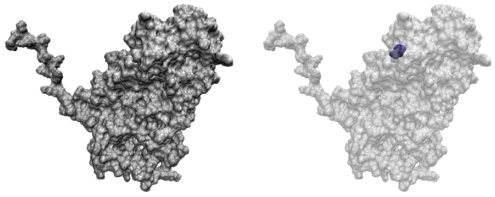
(A) TF vs CC Average pH



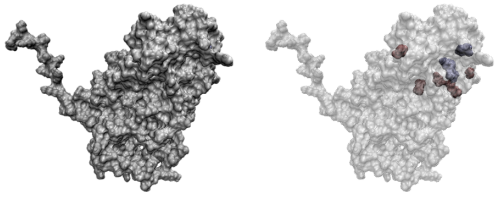
(B) B vs C Average pH



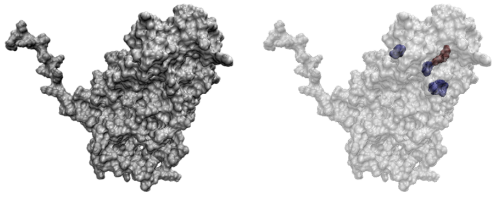
(C) TF vs CC pH 4 & 7



(D) B vs C pH 4 & 7

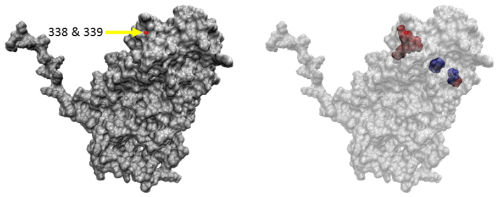


(E) TF vs CC pH 5 & 8

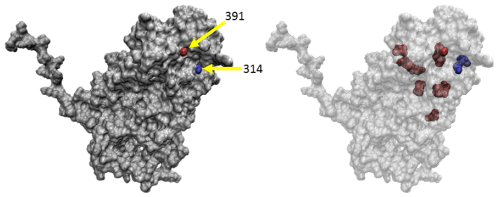


(F) B vs C pH 5 & 8

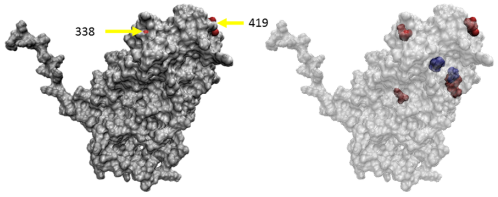
Figure 37: Structural Mapping of Gap Included Residue Sensitivities for Overall Classes and Clades. A) Binding site interface residues are colored yellow. B-G) Each sub-figure contains two images of the same mapping. The left image shows a surface view of the binding interface side of the model, and the right one has all residues transparent except the selected sensitive residues. Blue indicates greater sensitivity in TF (B, D & F) or B Clade (C, E & G). Red indicates greater sensitivity in CC (B, D & F) or C Clade (C, E & G). Green identifies the general location of identified residues that are not present in the model structure; green residues are the closest residue within two positions to the missing residue. Arrows indicate residues that are exposed near the CD4 binding interface.



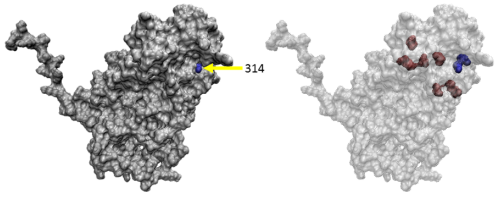
(A) Within B Clade Average pH



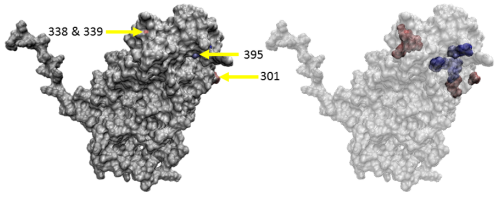
(B) Within C Clade Average pH



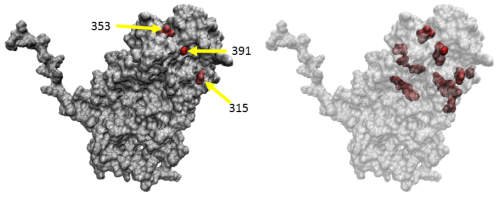
(C) Within B Clade pH 4 & 7



(D) Within C Clade pH 4 & 7

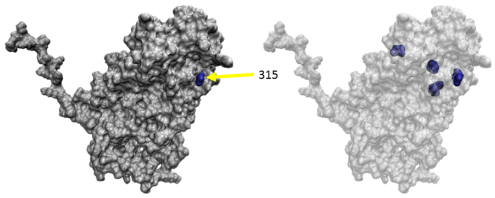


(E) Within B Clade pH 5 & 8

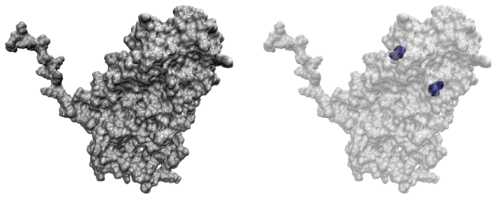


(F) Within C Clade pH 5 & 8

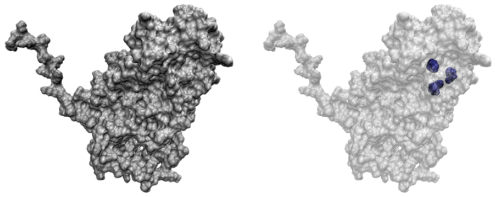
Figure 38: TF vs CC Structural Mapping of Gap Included Residue Sensitivities Within Clades. Each sub-figure contains two images of the same mapping. The left image shows a surface view of the binding interface side of the model, and the right one has all residues transparent except the selected sensitive residues. Blue indicates greater sensitivity in TF (A, C & E) or B Clade (B, D & F). Red indicates greater sensitivity in CC (A, C & E) or C Clade (B, D & F). Green identifies the general location of identified residues that are not present in the model structure; green residues are the closest residue within two positions to the missing residue. Arrows indicate residues that are exposed near the CD4 binding interface.



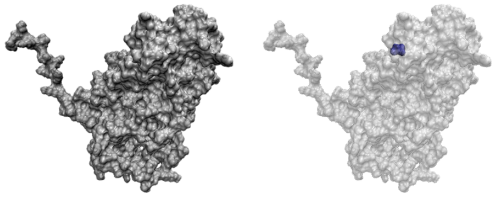
(A) Within TF Class Average pH



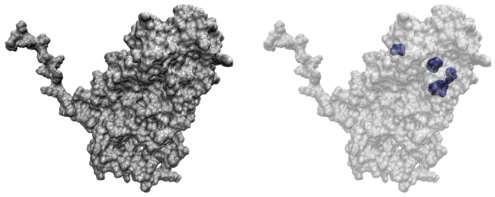
(B) Within CC Class Average pH



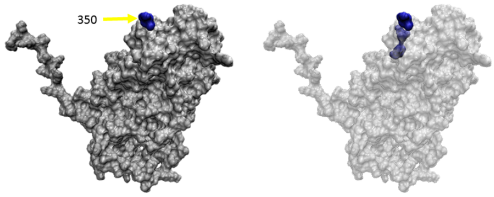
(C) Within TF Class pH 4 & 7



(D) Within CC Class pH 4 & 7



(E) Within TF Class pH 5 & 8



(F) Within CC Class pH 5 & 8

Figure 39: B vs C Structural Mapping of Gap Included Residue Sensitivities Within Classes. Each sub-figure contains two images of the same mapping. The left image shows a surface view of the binding interface side of the model, and the right one has all residues transparent except the selected sensitive residues. Blue indicates greater sensitivity in TF (A, C & E) or B Clade (B, D & F). Red indicates greater sensitivity in CC (A, C & E) or C Clade (B, D & F). Green identifies the general location of identified residues that are not present in the model structure; green residues are the closest residue within two positions to the missing residue. Arrows indicate residues that are exposed near the CD4 binding interface.

CHAPTER V.

DISCUSSION

Though AIDS and HIV have been studied for several decades, a viable vaccine has yet to be produced. Since the significance of the acidic pH of the typical mucosal transmission site has been broadly overlooked, we constructed a pipeline to analyze the pH sensitivity of the gp120-CD4 interaction in TF and CC strains. Here 24 sets of corresponding TF and CC gp120 sequences were analyzed for pH sensitivity of the gp120-CD4 interaction. B clade and C clade were compared as well, with 18 sets from B clade and 6 sets from C clade. These comparisons were also performed within clades, and within classes, respectively.

Overall and within B clade, the charge density of CC strains was found to be more pH sensitive when using the pH interval 4 and 7, and when using the average of the two intervals (Figure 4). This result differs from what was found previously [55], which may result from the increased number of sequences analyzed. The surface charge of B clade sequences was significantly more sensitive to pH than that of C clade sequences; this was consistent within both classes, though the pH 5 and 8 interval was only significant for the overall comparison 4.

Calculations using the bound conformation found TF strains to bind CD4 significantly better than CC strains at low pH (Figures 16B & 17); At high pH values, CC was found to bind CD4 significantly better in the calculation using the unbound gp120 conformation within B clade (Figure 20A); this was also present as a trend in the overall comparison (Figure 10A). The CC class within B clade also significantly prefers the unbound conformation at high pH values, and at several acidic pH values (Figure 23A). These results suggest that the increased CD4 binding at low pH in TF strains is not due to increased pressure to assume the bound conformation, and is more likely due to a more favorable interaction between gp120 and CD4; conversely, the increased binding ability of CC strains at higher pH values appears to be influenced by an increased preference to assume the bound conformation.

B clade was significantly better at binding CD4 at pH values between 4.1 and 4.6 when using the bound gp120 conformation (Figures 16C) & 18A). The only significant differences found using the unbound conformation were at pH 3.6 and 3.7 within the overall group, but there was a trend within class TF in which C clade bound CD4 better at pH values above 6 (Figure 21B). B clade preferred the unbound conformation over C clade from pH 5 to 7 within the overall group (Figure 22C) and within CC (Figure 24), but C clade prefers the unbound conformation within class TF from pH 6 to 8 (Figure 24A). The trend within class TF suggests that C clade may bind better at higher pH due to a preference for the bound conformation; however, this is based upon an observed trend, so a larger number of C clade samples would be needed to evaluate its significance.

In all binding energy calculations, gp120 was found to bind CD4 better at low pH values. This is also consistent with previous experimental results [55], which further supports the accuracy and utility of this pipeline.

It was not possible to distinguish corresponding TF and CC sequences or overall groups at the individual sequence level for any condition tested (Figures 16A, 19A & 22A). When calculating sensitivity with the bound gp120 conformations, no significant differences could be found between TF and CC (Figures 16 & 17). Using the unbound conformation for the calculation, the TF gp120-CD4 interaction was significantly more sensitive to pH within B clade (Figure 20A). This was consistent with previous experimental results [55], which supports the accuracy of this pipeline as a method of modeling pH sensitivity of protein-protein interactions.

Within the CC class, CD4 binding was found to be significantly more sensitive in B clade when using the bound conformation (Figure 18B), but significantly more sensitive in C clade when using the unbound conformation (Figure 21B). Also within CC, the preference for the bound conformation was found to be significantly more sensitive in C clade (Figure 24B).

These results suggest that within the CC class, pH affects CD4 binding through changes in the binding interface in B clade, while it affects the conformational shift in C clade.

Efforts to understand a mechanism of binding sensitivity identified multiple residues that may contribute to the observed differences (Tables 1, 2, 3 & 4). Unfortunately, sequence comparisons did not indicate any clear sequence difference that could contribute to the observed sensitivity differences (Figures 31 & 32, and Appendix E & F).

Mapping the residues onto a gp120 structure identified regions of the protein that likely contribute to the pH sensitivity mechanism. Though none of the most sensitive residues were found in the binding interface, many were found in close proximity (Figures 34, 35, 36, 37, 38 & 39). This provides potential targets for future investigations into this mechanism.

A possible alternative method of identifying important residues for the mechanism of pH sensitivity would be to systematically remove each residue from each sequence, and then use the altered sequences to calculate the binding energy of the gp120-CD4 interactions. However, this would greatly increase the number of models that would have to be produced and the amount of calculation time. To reduce the magnitude of this task, a method such as residue specific surface charge pH sensitivity could be used to identify potential residues to remove. Regardless, this approach would require increased computational power. Additionally, it would also be interesting to look at the pH sensitivity of the interaction between gp120 and bnAbs because their typical target is the CD4 binding site of gp120 [62].

While a mechanism for the pH sensitivity of gp120 surface charge density and CD4 binding was not determined, this work does show the importance of pH in this critical interaction. This is particularly important for HIV vaccine research because the CD4 binding site is an important vaccine target, and pH has been shown to affect antibody binding at the mucosa [20].

Additionally, this work shows the effectiveness of this pipeline in analyzing pH sensitivity of protein-protein interactions. The pipeline was capable of efficiently creating multiple

models for a large set of sequences, as well as calculate electrostatic information across a large set of conditions. Computed gp120-CD4 binding energy sensitivity were also consistent with previous work [55]. This tool could be applied to additional studies involving pH, as well. Studies involving the optimization or engineering of proteins for specific pH binding could utilize this pipeline to evaluate the binding interaction within the desired pH range. Additionally, mutational studies seeking to alter the pH at which a particular protein conformation occurs could analyze multiple altered sequences to determine the effect of pH on the conformation of the given sequence. This pipeline is a useful, generalizable tool for any study involving the effect of pH on conformation or protein-protein interactions.

BIBLIOGRAPHY

[1] ABDOOL KARIM, Q., ABDOOL KARIM, S. S., FROHLICH, J. A., GROBLER, A. C., BAXTER, C., MANSOOR, L. E., KHARSANY, A. B. M., SIBEKO, S., MLISANA, K. P., OMAR, Z., GENGIAH, T. N., MAARSCHALK, S., ARULAPPAN, N., MLOTSHWA, M., MORRIS, L., AND TAYLOR, D. Effectiveness and safety of tenofovir gel, an antiretroviral microbicide, for the prevention of HIV infection in women. *Science (New York, N.Y.) 329*, 5996 (Sep 2010), 1168–74.

[2] ABRAHAMS, M.-R., ANDERSON, J. A., GIORGI, E. E., SEOIGHE, C., MLISANA, K., PING, L.-H., ATHREYA, G. S., TREURNICHT, F. K., KEELE, B. F., WOOD, N., SALAZAR-GONZALEZ, J. F., BHATTACHARYA, T., CHU, H., HOFFMAN, I., GALVIN, S., MAPANJE, C., KAZEMBE, P., THEBUS, R., FISCUS, S., HIDE, W., COHEN, M. S., KARIM, S. A., HAYNES, B. F., SHAW, G. M., HAHN, B. H., KORBER, B. T., SWANSTROM, R., WILLIAMSON, C., CAPRISA ACUTE INFECTION STUDY TEAM, AND CENTER FOR HIV-AIDS VACCINE IMMUNOLOGY CONSORTIUM. 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.

[3] ARIËN, K. K., VANHAM, G., AND ARTS, E. J. Is HIV-1 evolving to a less virulent form in humans? *Nature Reviews. Microbiology 5*, 2 (2007), 141–51.

[4] ARTHOS, J., CICALA, C., MARTINELLI, E., MACLEOD, K., VAN RYK, D., WEI, D., XIAO, Z., VEENSTRA, T. D., CONRAD, T. P., LEMPICKI, R. A., MCLAUGHLIN, S., PASCUCCIO, M., GOPAUL, R., MCNALLY, J., CRUZ, C. C., CENSOPLANO, N., CHUNG, E., REITANO, K. N., KOTTILIL, S., GOODE, D. J., AND FAUCI, A. S. HIV-1 envelope protein binds to and signals through integrin alpha4beta7, the gut mucosal homing receptor for peripheral T cells. *Nature Immunology 9*, 3 (2008), 301–309.

[5] BAKER, N. A., SEPT, D., JOSEPH, S., HOLST, M. J., AND MCCAMMON, J. A. 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.

[6] BAR, K. J., TSAO, C.-Y., IYER, S. S., DECKER, J. M., YANG, Y., BONSIGNORI, M., CHEN, X., HWANG, K.-K., MONTEFIORI, D. C., LIAO, H.-X., HRABER, P., FISCHER, W., LI, H., WANG, S., STERRETT, S., KEELE, B. F., GANUSOV, V. V., PERELSON, A. S., KORBER, B. T., GEORGIEV, I., MCLELLAN, J. S., PAVLICEK, J. W., GAO, F., HAYNES, B. F., HAHN, B. H., KWONG, P. D., AND SHAW, G. M. Early low-titer neutralizing antibodies impede HIV-1 replication and select for virus escape. *PLoS Pathogens 8*, 5 (2012), e1002721.

[7] BAROUCH, D. H., STEPHENSON, K. E., BORDUCCHI, E. N., SMITH, K., STANLEY, K., MCNALLY, A. G., LIU, J., ABBINK, P., MAXFIELD, L. F., SEAMAN, M. S., DUGAST, A.-S., ALTER, G., FERGUSON, M., LI, W., EARL, P. L., MOSS, B., GIORGI, E. E., SZINGER, J. J., ELLER, L. A., BILLINGS, E. A., RAO, M., TOVANABUTRA, S., SANDERS-BUELL, E., WEIJTENS, M., PAU, M. G., SCHUITEMAKER, H., ROBB, M. L., KIM, J. H., KORBER, B. T., AND MICHAEL, N. L. Protective efficacy of a global HIV-1 mosaic vaccine against heterologous SHIV challenges in rhesus monkeys. *Cell 155*, 3 (Oct 2013), 531–9.

[8] BEITZ, E. Texshade: Shading and labeling of multiple sequence alignments using latex2e. *Bioinformatics 16*, 2 (2000), 135.

[9] BUNNIK, E. M., PISAS, L., VAN NUENEN, A. C., AND SCHUITEMAKER, H. 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.

[10] BURTON, D. R., POIGNARD, P., STANFIELD, R. L., AND WILSON, I. A. Broadly neutralizing antibodies present new prospects to counter highly antigenically diverse viruses. *Science (New York, N.Y.) 337*, 6091 (Jul 2012), 183–6.

[11] CHABIKULI NO, MBCHB, MCFP, MFAMMED, MSC, DATONYE DO, MBCHB, MPH, MFAMMED, NACHEGA J, MD, PHD, ANSONG D, MBCHB, M. Adherence to antiretroviral therapy, virologic failure and workload at the Rustenburg Provincial Hospital. *SA Fam Pract 52*, 4 (2010), 350–355.

[12] CHAMBERS, J. *Graphical methods for data analysis*. Chapman & Hall statistics series. Wadsworth International Group, 1983.

[13] CHEN, B., VOGAN, E. M., GONG, H., SKEHEL, J. J., WILEY, D. C., AND HARRISON, S. C. Structure of an unliganded simian immunodeficiency virus gp120 core., 2005.

[14] CICALA, C., MARTINELLI, E., MCNALLY, J. P., GOODE, D. J., GOPAUL, R., HIATT, J., JELICIC, K., KOTTILIL, S., MACLEOD, K., O'SHEA, A., PATEL, N., VAN RYK, D., WEI, D., PASCUCCIO, M., YI, L., MCKINNON, L., IZULLA, P., KIMANI, J., KAUL, R., FAUCI, A. S., AND ARTHOS, J. The integrin alpha4beta7 forms a complex with cell-surface CD4 and defines a T-cell subset that is highly susceptible to infection by HIV-1. *Proceedings of the National Academy of Sciences of the United States of America 106*, 49 (2009), 20877–82.

[15] DACHEUX, L., MOREAU, A., ATAMAN-ONAL, Y., BIRON, F., VERRIER, B., AND BARIN, F. 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.

[16] DISKIN, R., MARCOVECCHIO, P. M., AND BJORKMAN, P. J. Structure of a clade C HIV-1 gp120 bound to CD4 and CD4-induced antibody reveals anti-CD4 polyreactivity. *Nature Structural Molecular Biology 17*, 5 (May 2010), 608–13.

[17] DOLINSKY, T. J., CZODROWSKI, P., LI, H., NIELSEN, J. E., JENSEN, J. H., KLEBE, G., AND BAKER, N. A. PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research 35*, SUPPL.2 (2007), 522–525.

[18] DOLINSKY, T. J., NIELSEN, J. E., MCCAMMON, J. A., AND BAKER, N. A. PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Research 32*, WEB SERVER ISS. (2004), 665–667.

[19] ESWAR, N., WEBB, B., MARTI-RENOM, M. A., MADHUSUDHAN, M., ERAMIAN, D., SHEN, M.-Y., PIEPER, U., AND SALI, A. *Comparative protein structure modeling using modeller*. John Wiley & Sons, Inc., 2002.

[20] FAHRBACH, K. M., MALYKHINA, O., STIEH, D. J., AND HOPE, T. J. Differential binding of igg and iga to mucus of the female reproductive tract. *PLOS ONE 8*, 10 (10 2013), 1–11.

[21] FARRELL, D. W., SPERANSKIY, K., AND THORPE, M. F. Generating stereochemically acceptable protein pathways. *Proteins: Structure, Function and Bioinformatics 78*, 14 (2010), 2908–2921.

[22] FISCHER, W., PERKINS, S., THEILER, J., BHATTACHARYA, T., YUSIM, K., FUNKHOUSER, R., KUIKEN, C., HAYNES, B., LETVIN, N. L., WALKER, B. D., HAHN, B. H., AND KORBER, B. T. Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nature Medicine 13*, 1 (Jan 2007), 100–6.

[23] GOUJON, M., MCWILLIAM, H., LI, W., VALENTIN, F., SQUIZZATO, S., PAERN, J., AND LOPEZ, R. A new bioinformatics analysis tools framework at emblebi. *Nucleic Acids Research 38*, suppl.2 (2010), W695.

[24] HARTLEY, O., KLASSE, P. J., SATTENTAU, Q. J., AND MOORE, J. P. V3: HIV's Switch-Hitter. *AIDS Research and Human Retroviruses 21*, 2 (2005), 171–189.

[25] HUANG, C.-C., TANG, M., ZHANG, M.-Y., MAJEED, S., MONTABANA, E., STANFIELD, R. L., DIMITROV, D. S., KORBER, B., SODROSKI, J., WILSON, I. A., WYATT, R., AND KWONG, P. D. Structure of a V3-containing HIV-1 gp120 core. *Science (New York, N.Y.) 310*, 5750 (Nov 2005), 1025–8.

[26] HUANG, C. C., VENTURI, M., MAJEED, S., MOORE, M. J., PHOGAT, S., ZHANG, M. Y., DIMITROV, D. S., HENDRICKSON, W. A., ROBINSON, J., SODROSKI, J., WYATT, R., CHOE, H., FARZAN, M., AND KWONG, P. D. Structural basis of tyrosine sulfation and VH-gene usage in antibodies that recognize the HIV type 1 coreceptor-binding site on gp120. *Proc Natl Acad Sci U S A 101*, 9 (2004), 2706–2711.

[27] HUMPHREY, W., DALKE, A., AND SCHULTEN, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics 14* (1996), 33–38.

[28] KEELE, B. F., GIORGI, E. E., SALAZAR-GONZALEZ, J. F., DECKER, J. M., PHAM, K. T., SALAZAR, M. G., SUN, C., GRAYSON, T., WANG, S., LI, H., WEI, X., JIANG, C., KIRCHHERR, J. L., GAO, F., ANDERSON, J. A., PING, L.-H., SWANSTROM, R., TOMARAS, G. D., BLATTNER, W. A., GOEPFERT, P. A., KILBY, J. M., SAAG, M. S., DELWART, E. L., BUSCH, M. P., COHEN, M. S., MONTEFIORI, D. C., HAYNES, B. F., GASCHEN, B., ATHREYA, G. S., LEE, H. Y., WOOD, N., SEOIGHE, C., PERELSON, A. S., BHATTACHARYA, T., KORBER, B. T., HAHN, B. H., AND SHAW, G. M. 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.

[29] KORBER, B., AND GNANAKARAN, S. Converging on an hiv vaccine. *Science 333*, 6049 (2011), 1589–1590.

[30] KOTHE, D. L., LI, Y., DECKER, J. M., BIBOLLET-RUCHE, F., ZAMMIT, K. P., SALAZAR, M. G., CHEN, Y., WENG, Z., WEAVER, E. A., GAO, F., HAYNES, B. F., SHAW, G. M., KORBER, B. T. M., AND HAHN, B. H. Ancestral and consensus envelope immunogens for HIV-1 subtype C. *Virology 352*, 2 (Sep 2006), 438–49.

[31] KWONG, P. D., WYATT, R., MAJEED, S., ROBINSON, J., SWEET, R. W., SODROSKI, J., AND HENDRICKSON, W. A. Structures of HIV-1 gp120 envelope glycoproteins from laboratory-adapted and primary isolates. *Structure (London, England : 1993) 8*, 12 (Dec 2000), 1329–39.

[32] LEONARD, C. K., SPELLMAN, M. W., RIDDLE, L., HARRIS, R. J., THOMAS, J. N., AND GREGORY, T. J. Assignment of intrachain disulfide bonds and characterization of potential glycosylation sites of the type 1 recombinant human immunodeficiency virus envelope glycoprotein (gp120) expressed in Chinese hamster ovary cells. *The Journal of Biological Chemistry 265*, 18 (1990), 10373–10382.

[33] LI, B., DECKER, J. M., JOHNSON, R. W., BIBOLLET-RUCHE, F., WEI, X., MULENGA, J., ALLEN, S., HUNTER, E., HAHN, B. H., SHAW, G. M., BLACKWELL, J. L., AND DERDEYN, C. A. 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.

[34] LI, H., BAR, K. J., WANG, S., DECKER, J. M., CHEN, Y., SUN, C., SALAZAR-GONZALEZ, J. F., SALAZAR, M. G., LEARN, G. H., MORGAN, C. J., SCHUMACHER, J. E., HRABER, P., GIORGI, E. E., BHATTACHARYA, T., KORBER, B. T., PERELSON, A. S., ERON, J. J., COHEN, M. S., HICKS, C. B., HAYNES, B. F., MARKOWITZ, M., KEELE, B. F., HAHN, B. H., AND SHAW, G. M. High Multiplicity Infection by HIV-1 in Men Who Have Sex with Men. *PLoS Pathogens 6*, 5 (May 2010), e1000890.

[35] LIAO, H.-X., BONSIGNORI, M., ALAM, S. M., MCLELLAN, J. S., TOMARAS, G. D., MOODY, M. A., KOZINK, D. M., HWANG, K.-K., CHEN, X., TSAO, C.-Y., LIU, P., LU, X., PARKS, R. J., MONTEFIORI, D. C., FERRARI, G., POLLARA, J., RAO, M., PEACHMAN, K. K., SANTRA, S., LETVIN, N. L., KARASAVVAS, N., YANG, Z.-Y., DAI, K., PANCERA, M., GORMAN, J., WIEHE, K., NICELY, N. I., RERKS-NGARM, S., NITAYAPHAN, S., KAEWKUNGWAL, J., PITISUTTITHUM, P., TARTAGLIA, J., SINANGIL, F., KIM, J. H., MICHAEL, N. L., KEPLER, T. B., KWONG, P. D., MASCOLA, J. R., NABEL, G. J., PINTER, A., ZOLLA-PAZNER, S., AND HAYNES, B. F. 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.

[36] LIAO, H.-X., LYNCH, R., ZHOU, T., GAO, F., ALAM, S. M., BOYD, S. D., FIRE, A. Z., ROSKIN, K. M., SCHRAMM, C. A., ZHANG, Z., ZHU, J., SHAPIRO, L., MULLIKIN, J. C., GNANAKARAN, S., HRABER, P., WIEHE, K., KELSOE, G., YANG, G., XIA, S.-M., MONTEFIORI, D. C., PARKS, R., LLOYD, K. E., SCEARCE, R. M., SODERBERG, K. A., COHEN, M., KAMANGA, G., LOUDER, M. K., TRAN, L. M., CHEN, Y., CAI, F., CHEN, S., MOQUIN, S., DU, X., JOYCE, M. G., SRIVATSAN, S., ZHANG, B., ZHENG, A., SHAW, G. M., HAHN, B. H., KEPLER, T. B., KORBER, B. T. M., KWONG, P. D., MASCOLA, J. R., AND HAYNES, B. F. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature 496*, 7446 (2013), 469–76.

[37] LIU, M. K. P., HAWKINS, N., RITCHIE, A. J., GANUSOV, V. V., WHALE, V., BRACKENRIDGE, S., LI, H., PAVLICEK, J. W., CAI, F., ROSE-ABRAHAMS, M., TREURNICHT, F., HRABER, P., RIOU, C., GRAY, C., FERRARI, G., TANNER, R., PING, L.-H., ANDERSON, J. A., SWANSTROM, R., CHAVI CORE B, COHEN, M., KARIM, S. S. A., HAYNES, B., BORROW, P., PERELSON, A. S., SHAW, G. M., HAHN, B. H., WILLIAMSON, C., KORBER, B. T., GAO, F., SELF, S., MCMICHAEL, A., AND GOONETILLEKE, N. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *The Journal of clinical investigation 123*, 1 (jan 2013), 380–93.

[38] LU, M., BLACKLOW, S. C., AND KIM, P. S. A trimeric subdomain of the simian immunodeficiency virus envelope glycoprotein. *Nature Structural Biology 34*, 46 (Nov 1995), 14955–62.

[39] MCWILLIAM, H., LI, W., ULUDAG, M., SQUIZZATO, S., PARK, Y. M., BUSO, N., COWLEY, A. P., AND LOPEZ, R. Analysis tool web services from the embl-ebi. *Nucleic Acids Research 41*, W1 (2013), W597.

[40] NAGOT, N., OUEDRAOGO, A., WEISS, H. A., KONATE, I., SANON, A., DEFER, M.-C., SAWADOGO, A., ANDONABA, J.-B., VALLO, R., BECQUART, P., SEGONDY, M., MAYAUD, P., AND VAN DE PERRE, P. Longitudinal effect following initiation of highly active antiretroviral therapy on plasma and cervico-vaginal HIV-1 RNA among women in Burkina Faso. *Sexually Transmitted Infections 84*, 3 (2008), 167–70.

[41] OJIKUTU, B., MAKADZANGE, A. T., AND GAOLATHE, T. Scaling up ART treatment capacity: lessons learned from South Africa, Zimbabwe, and Botswana. *Current HIV/AIDS Reports 5*, 2 (May 2008), 94–8.

[42] OLSSON, M. H. M., SØNDERGAARD, C. R., ROSTKOWSKI, M., AND JENSEN, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *Journal of Chemical Theory and Computation 7*, 2 (2011), 525–537.

[43] PANCERA, M., MAJEED, S., BAN, Y.-E. A., CHEN, L., HUANG, C.-C., KONG, L., KWON, Y. D., STUCKEY, J., ZHOU, T., ROBINSON, J. E., SCHIEF, W. R., SODROSKI, J., WYATT, R., AND KWONG, P. D. Structure of HIV-1 gp120 with gp41-interactive region reveals layered envelope architecture and basis of conformational mobility. *Proceedings of the National Academy of Sciences of the United States of America 107*, 3 (Jan 2010), 1166–71.

[44] PANTOPHLET, R., AND BURTON, D. R. GP120: Target for neutralizing HIV-1 antibodies. *Annual Review of Immunology 24* (2006), 739–769.

[45] PARRISH, N. F., GAO, F., LI, H., GIORGI, E. E., BARBIAN, H. J., PARRISH, E. H., ZAJIC, L., IYER, S. S., DECKER, J. M., KUMAR, A., HORA, B., BERG, A., CAI, F., HOPPER, J., DENNY, T. N., DING, H., OCHSENBAUER, C., KAPPES, J. C., GALIMIDI, R. P., WEST, A. P., BJORKMAN, P. J., WILEN, C. B., DOMS, R. W., O'BRIEN, M., BHARDWAJ, N., BORROW, P., HAYNES, B. F., MULDOON, M., THEILER, J. P., KORBER, B., SHAW, G. M., AND HAHN, B. H. 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.

[46] PHILLIPS, J. L., AND GNANAKARAN, S. A data-driven approach to modeling the tripartite structure of multidrug resistance efflux pumps. *Proteins: Structure, Function and Bioinformatics 83*, 1 (2015), 46–65.

[47] POIGNARD, P., SAPHIRE, E. O., PARREN, P. W., AND BURTON, D. R. gp120: Biologic aspects of structural features. *Annual Review of Immunology 19* (2001), 253–274.

[48] R CORE TEAM. *R: A Language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[49] ROTTA, I., AND DE ALMEIDA, S. M. Genotypical diversity of HIV clades and central nervous system impairment. *Arquivos de Neuro-psiquiatria 69*, 6 (Dec 2011), 964–72.

[50] SALAZAR-GONZALEZ, J. F., SALAZAR, M. G., KEELE, B. F., LEARN, G. H., GIORGI, E. E., LI, H., DECKER, J. M., WANG, S., BAALWA, J., KRAUS, M. H., PARRISH, N. F., SHAW, K. S., GUFFEY, M. B., BAR, K. J., DAVIS, K. L., OCHSENBAUER-JAMBOR, C., KAPPES, J. C., SAAG, M. S., COHEN, M. S., MULENGA, J., DERDEYN, C. A., ALLEN, S., HUNTER, E., MARKOWITZ, M., HRABER, P., PERELSON, A. S., BHATTACHARYA, T., HAYNES, B. F., KORBER, B. T., HAHN, B. H., AND SHAW, G. M. 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.

[51] SALI, A., AND BLUNDELL, T. L. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology 234*, 3 (1993), 779–815.

[52] SAPHIRE, A. C. S., BOBARDT, M. D., ZHANG, Z., DAVID, G., AND GALLAY, P. A. Syndecans Serve as Attachment Receptors for Human Immunodeficiency Virus Type 1 on Macrophages. *Journal of Virology 75*, 19 (Oct 2001), 9187–9200.

[53] SIEVERS, F., WILM, A., DINEEN, D., GIBSON, T. J., KARPLUS, K., LI, W., LOPEZ, R., MCWILLIAM, H., REMMERT, M., SÖDING, J., THOMPSON, J. D., AND HIGGINS, D. G. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology 7*, 1 (2011).

[54] SØNDERGAARD, C. R., OLSSON, M. H. M., ROSTKOWSKI, M., AND JENSEN, J. H. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of p K a values. *Journal of Chemical Theory and Computation 7*, 7 (2011), 2284–2295.

[55] STIEH, D. J., PHILLIPS, J. L., ROGERS, P. M., KING, D. F., CIANCI, G. C., JEFFS, S. A., GNANAKARAN, S., AND SHATTOCK, R. J. Dynamic electrophoretic fingerprinting of the hiv-1 envelope glycoprotein. *Retrovirology 10*, 1 (2013), 33.

[56] TAYLOR, B. S., SOBIESZCZYK, M. E., MCCUTCHAN, F. E., AND HAMMER, S. M. The challenge of HIV-1 subtype diversity. *The New England Journal of Medicine 358*, 15 (Apr 2008), 1590–602.

[57] TURNBULL, E. L., WONG, M., WANG, S., WEI, X., JONES, N. A., CONROD, K. E., ALDAM, D., TURNER, J., PELLEGRINO, P., KEELE, B. F., WILLIAMS, I., SHAW, G. M., AND BORROW, P. 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.

[58] WAGIH, O. *RWebLogo: Plotting custom sequence logos*, 2014. R package version 1.0.3.

[59] WEI, X., DECKER, J. M., WANG, S., HUI, H., KAPPES, J. C., WU, X., SALAZAR-GONZALEZ, J. F., SALAZAR, M. G., KILBY, J. M., SAAG, M. S., KOMAROVA, N. L., NOWAK, M. A., HAHN, B. H., KWONG, P. D., AND SHAW, G. M. Antibody neutralization and escape by HIV-1. *Nature 422*, 6929 (Mar 2003), 307–12.

[60] WILEN, C. B., TILTON, J. C., AND DOMS, R. W. Hiv: Cell binding and entry. *Cold Spring Harbor Perspectives in Medicine 2*, 8 (2012).

[61] WU, X., ZHOU, T., ZHU, J., ZHANG, B., GEORGIEV, I., WANG, C., CHEN, X., LONGO, N. S., LOUDER, M., MCKEE, K., O'DELL, S., PERFETTO, S., SCHMIDT, S. D., SHI, W., WU, L., YANG, Y., YANG, Z.-Y., YANG, Z., ZHANG, Z., BONSIGNORI, M., CRUMP, J. A., KAPIGA, S. H., SAM, N. E., HAYNES, B. F., SIMEK, M., BURTON, D. R., KOFF, W. C., DORIA-ROSE, N. A., CONNORS, M., MULLIKIN, J. C., NABEL, G. J., ROEDERER, M., SHAPIRO, L., KWONG, P. D., AND MASCOLA, J. R. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science (New York, N.Y.) 333*, 6049 (Sep 2011), 1593–602.

[62] WYATT, R., KWONG, P. D., DESJARDINS, E., SWEET, R. W., ROBINSON, J., HENDRICKSON, W. A., AND SODROSKI, J. G. The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature 393*, 6686 (Jun 1998), 705–11.

[63] ZHOU, T., XU, L., DEY, B., HESSELL, A. J., VAN RYK, D., XIANG, S.-H., YANG, X., ZHANG, M.-Y., ZWICK, M. B., ARTHOS, J., BURTON, D. R., DIMITROV, D. S., SODROSKI, J., WYATT, R., NABEL, G. J., AND KWONG, P. D. Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature 445*, 7129 (Feb 2007), 732–7.

APPENDICES

APPENDIX A

Sequence Information

The gp120 regions of the following sequences were used in this study.

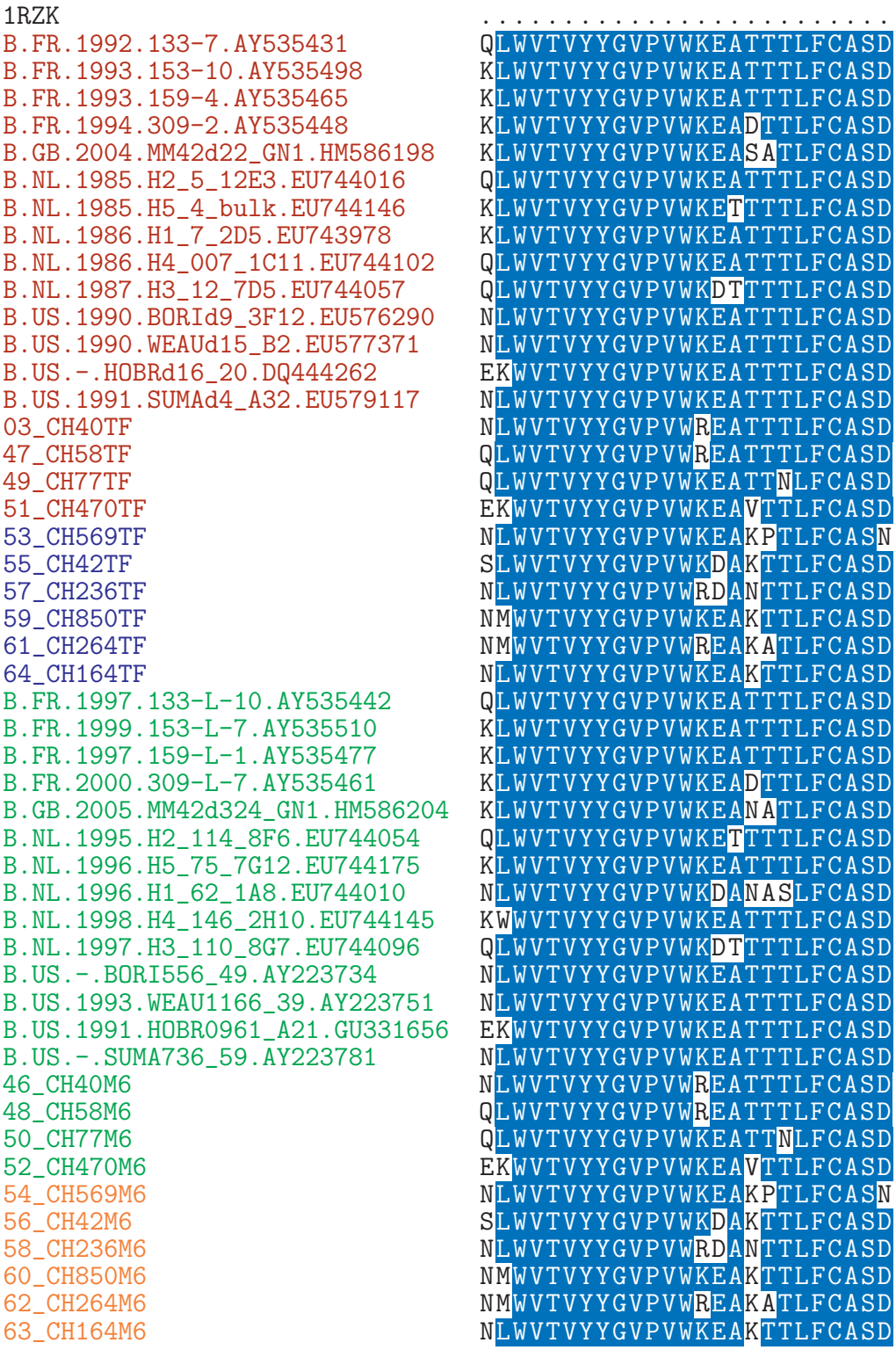
Table A.1: Sequence Information

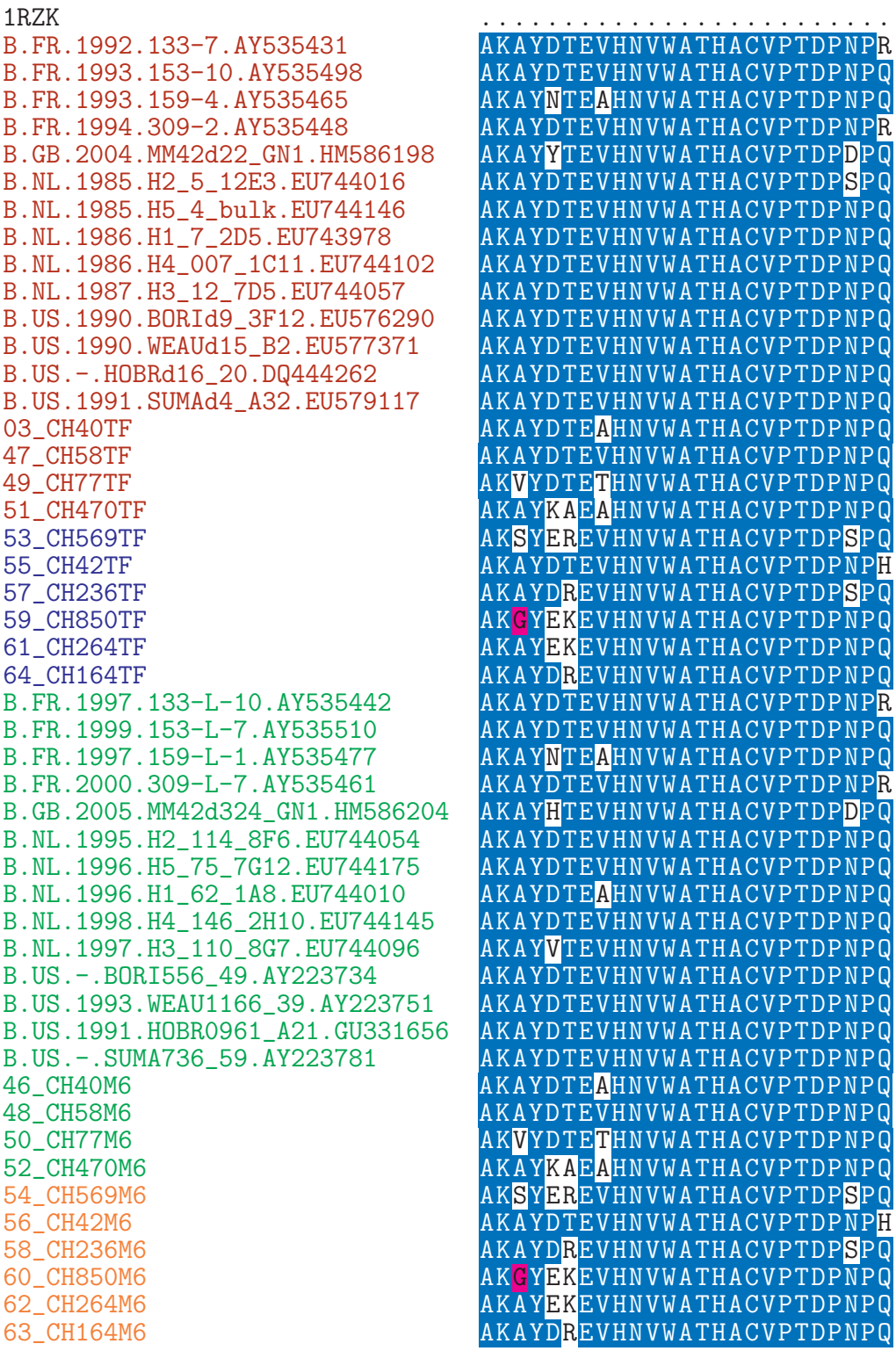
Sequence Name	Protein Accession	Clade	Class	Reference
03_CH40TF	ACD41465	B	TF	[28]
46_CH40M6	AFK87864	B	CC	[6]
47_CH58TF	ACE68159	B	TF	[28]
48_CH58M6	AFK88130	B	CC	[6]
49_CH77TF	ACD41595	B	TF	[28]
50_CH77M6	ACR52213	B	CC	[50]
51_CH470TF	AGG92565	B	TF	[37]
52_CH470M6	AGG92637	B	CC	[37]
53_CH569TF	Unavailable	C	TF	[30]
54_CH569M6	Unavailable	C	CC	[30]
55_CH42TF	ACS67441	C	TF	[2]
56_CH42M6	AGF30459	C	CC	Unpublished
57_CH236TF	ACS67726	C	TF	[2]
58_CH236M6	Unavailable	C	CC	[30]
59_CH850TF	Unavailable	C	TF	[30]
60_CH850M6	Unavailable	C	CC	[30]
61_CH264TF	Unavailable	C	TF	[30]
62_CH264M6	Unavailable	C	CC	[30]
64_CH164TF	AGG99748	C	TF	[37]
63_CH164M6	AGG99898	C	CC	[37]
B.FR.1992.133-7.AY535431	AAS58774	B	TF	[15]
B.FR.1997.133-L-10.AY535442	AAS58785	B	CC	[15]
B.FR.1993.153-10.AY535498	AAS58841	B	TF	[15]
B.FR.1999.153-L-7.AY535510	AAS58853	B	CC	[15]
B.FR.1993.159-4.AY535465	AAS58808	B	TF	[15]
B.FR.1997.159-L-1.AY535477	AAS58820	B	CC	[15]
B.FR.1994.309-2.AY535448	AAS58791	B	TF	[15]
B.FR.2000.309-L-7.AY535461	AAS58804	B	CC	[15]
B.GB.2004.MM42d22_GN1.HM586198	ADK75299	B	TF	[57]
B.GB.2005.MM42d324_GN1.HM586204	ADK75319	B	CC	[57]
B.NL.1985.H2_5_12E3.EU744016	ACE76354	B	TF	[9]
B.NL.1995.H2_114_8F6.EU744054	ACE76392	B	CC	[9]
B.NL.1985.H5_4_bulk.EU744146	ACE76484	B	TF	[9]
B.NL.1996.H5_75_7G12.EU744175	ACE76513	B	CC	[9]
B.NL.1986.H1_7_2D5.EU743978	ACE76316	B	TF	[9]
B.NL.1996.H1_62_1A8.EU744010	ACE76348	B	CC	[9]
B.NL.1986.H4_007_1C11.EU744102	ACE76440	B	TF	[9]
B.NL.1998.H4_146_2H10.EU744145	ACE76483	B	CC	[9]
B.NL.1987.H3_12_7D5.EU744057	ACE76395	B	TF	[9]
B.NL.1997.H3_110_8G7.EU744096	ACE76434	B	CC	[9]
B.US.1990.BORId9_3F12.EU576290	ACE67727	B	TF	[28]
B.US.-.BORI556_49.AY223734	AAP57334	B	CC	[59]
B.US.-.HOBRd16_20.DQ444262	ABD96594	B	TF	[33]
B.US.1991.HOBR0961_A21.GU331656	AEO86126	B	CC	[34]
B.US.1991.SUMAd4_A32.EU579117	ACD42000	B	TF	[28]
B.US.-.SUMA736_59.AY223781	AAP57376	B	CC	[59]
B.US.1990.WEAUd15_B2.EU577371	ACE71521	B	TF	[28]
B.US.1993.WEAU1166_39.AY223751	AAP57351	B	CC	Unpublished

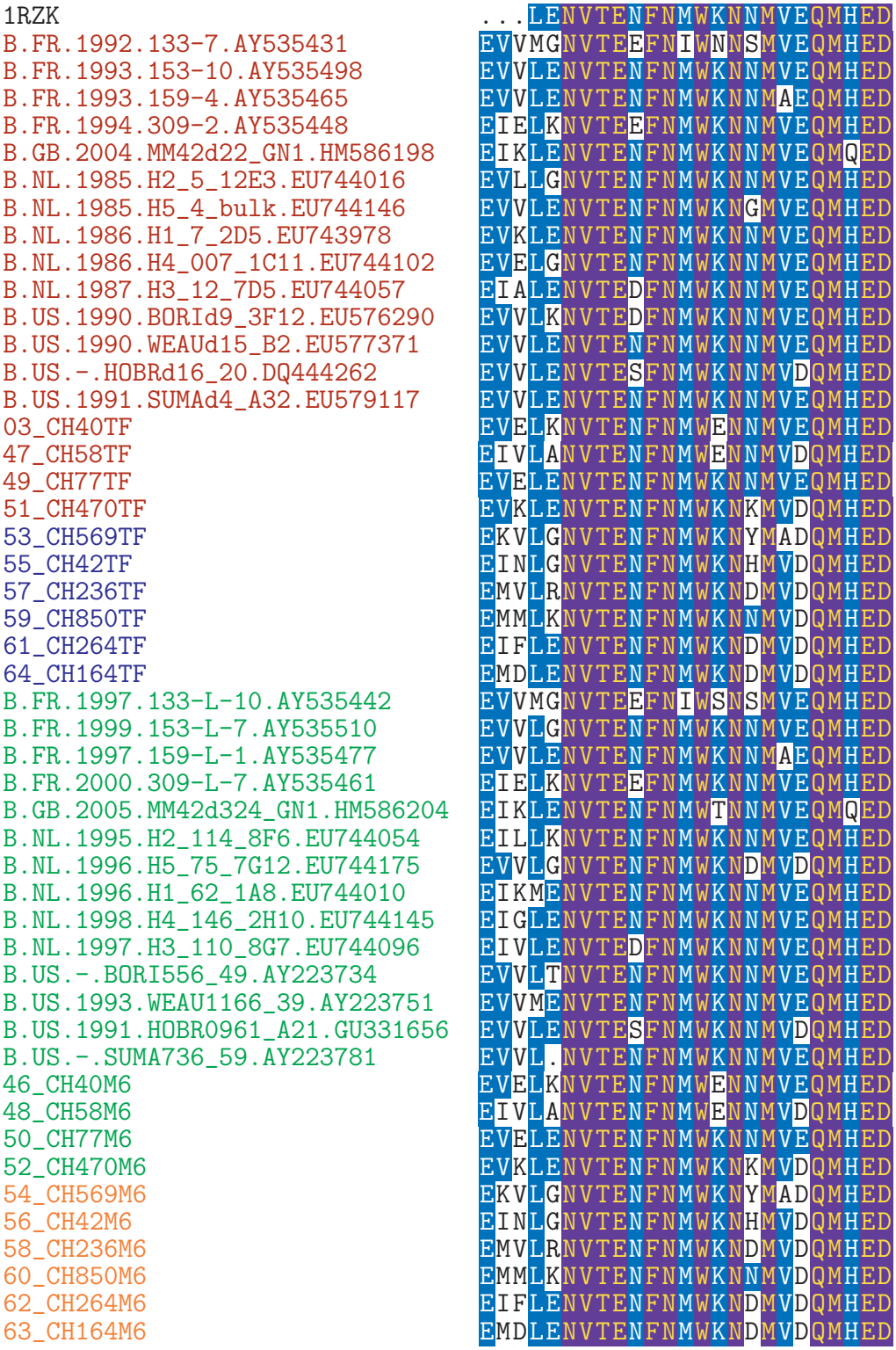
APPENDIX B

gp120 Sequence Alignment

All sequences used in this study were aligned using Clustal Omega [53, 23, 39]. TeXshade [8] was used to format the alignment for publication and to highlight regions of consensus and similarity. In black is the template sequence 1RZK [26]. Names in red are B clade TF strains. Names in Blue are C clade TF strains. Names in green are B clade CC strains. Names in orange are C clade CC strains. The alignment is color coded by similarity. A purple column indicates perfect consensus A blue column indicates majority consensus Pink indicates consensus based upon similarity of amino acids. The alignment is 22 pages because the alignment length is 527 amino acids, and 25 amino acids are included on each page.

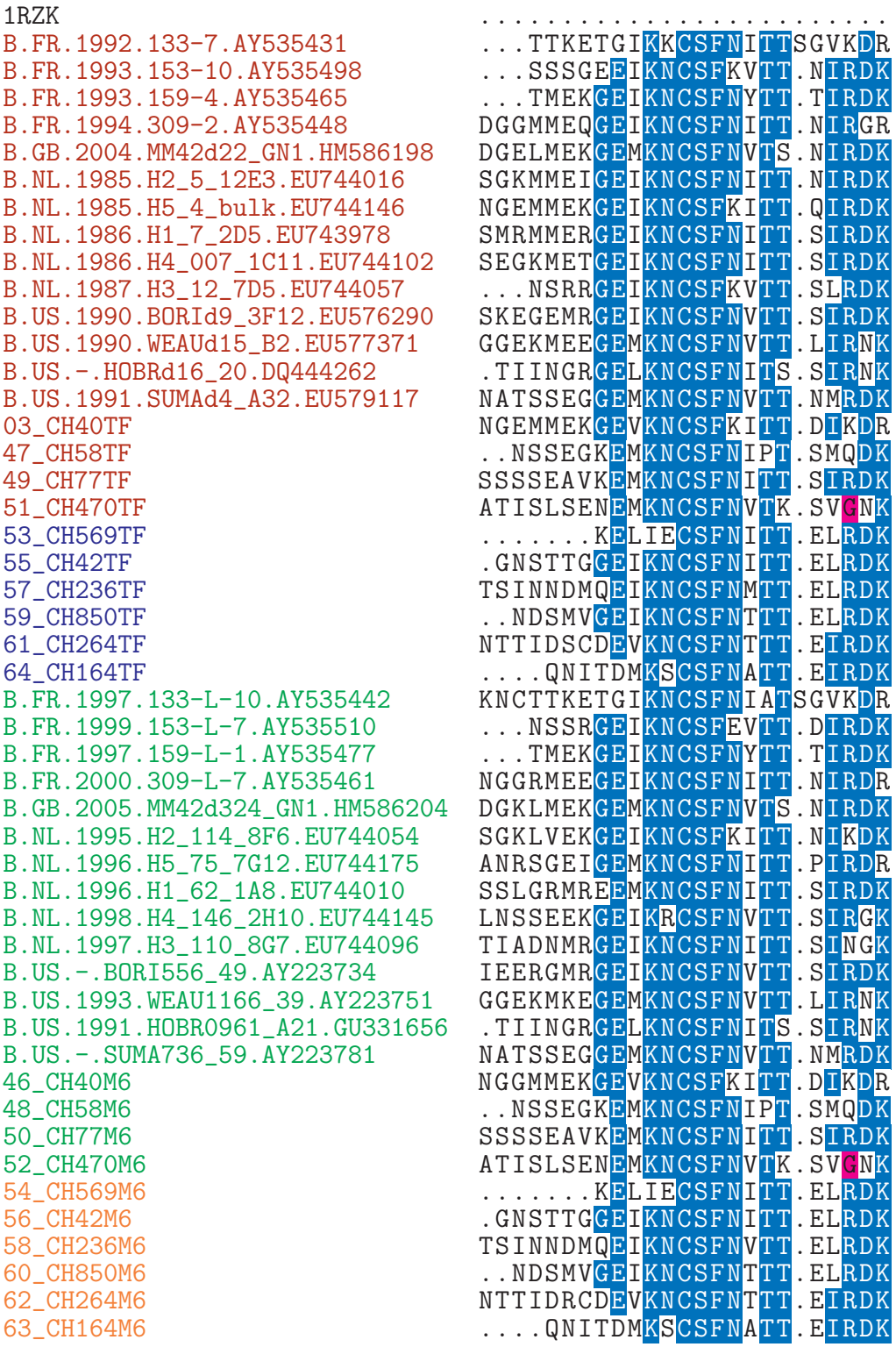

1RZK
B.FR.1992.133-7.AY535431
B.FR.1993.153-10.AY535498
B.FR.1993.159-4.AY535465
B.FR.1994.309-2.AY535448
B.GB.2004.MM42d22_GN1.HM586198
B.NL.1985.H2_5_12E3.EU744016
B.NL.1985.H5_4_bulk.EU744146
B.NL.1986.H1_7_2D5.EU743978
B.NL.1986.H4_007_1C11.EU744102
B.NL.1987.H3_12_7D5.EU744057
B.US.1990.BORId9_3F12.EU576290
B.US.1990.WEAUd15_B2.EU577371
B.US.-.HOBRd16_20.DQ444262
B.US.1991.SUMAd4_A32.EU579117
03_CH40TF
47_CH58TF
49_CH77TF
51_CH470TF
53_CH569TF
55_CH42TF
57_CH236TF
59_CH850TF
61_CH264TF
64_CH164TF
B.FR.1997.133-L-10.AY535442
B.FR.1999.153-L-7.AY535510
B.FR.1997.159-L-1.AY535477
B.FR.2000.309-L-7.AY535461
B.GB.2005.MM42d324_GN1.HM586204
B.NL.1995.H2_114_8F6.EU744054
B.NL.1996.H5_75_7G12.EU744175
B.NL.1996.H1_62_1A8.EU744010
B.NL.1998.H4_146_2H10.EU744145
B.NL.1997.H3_110_8G7.EU744096
B.US.-.BORI556_49.AY223734
B.US.1993.WEAU1166_39.AY223751
B.US.1991.HOBR0961_A21.GU331656
B.US.-.SUMA736_59.AY223781
46_CH40M6
48_CH58M6
50_CH77M6
52_CH470M6
54_CH569M6
56_CH42M6
58_CH236M6
60_CH850M6
62_CH264M6
63_CH164M6

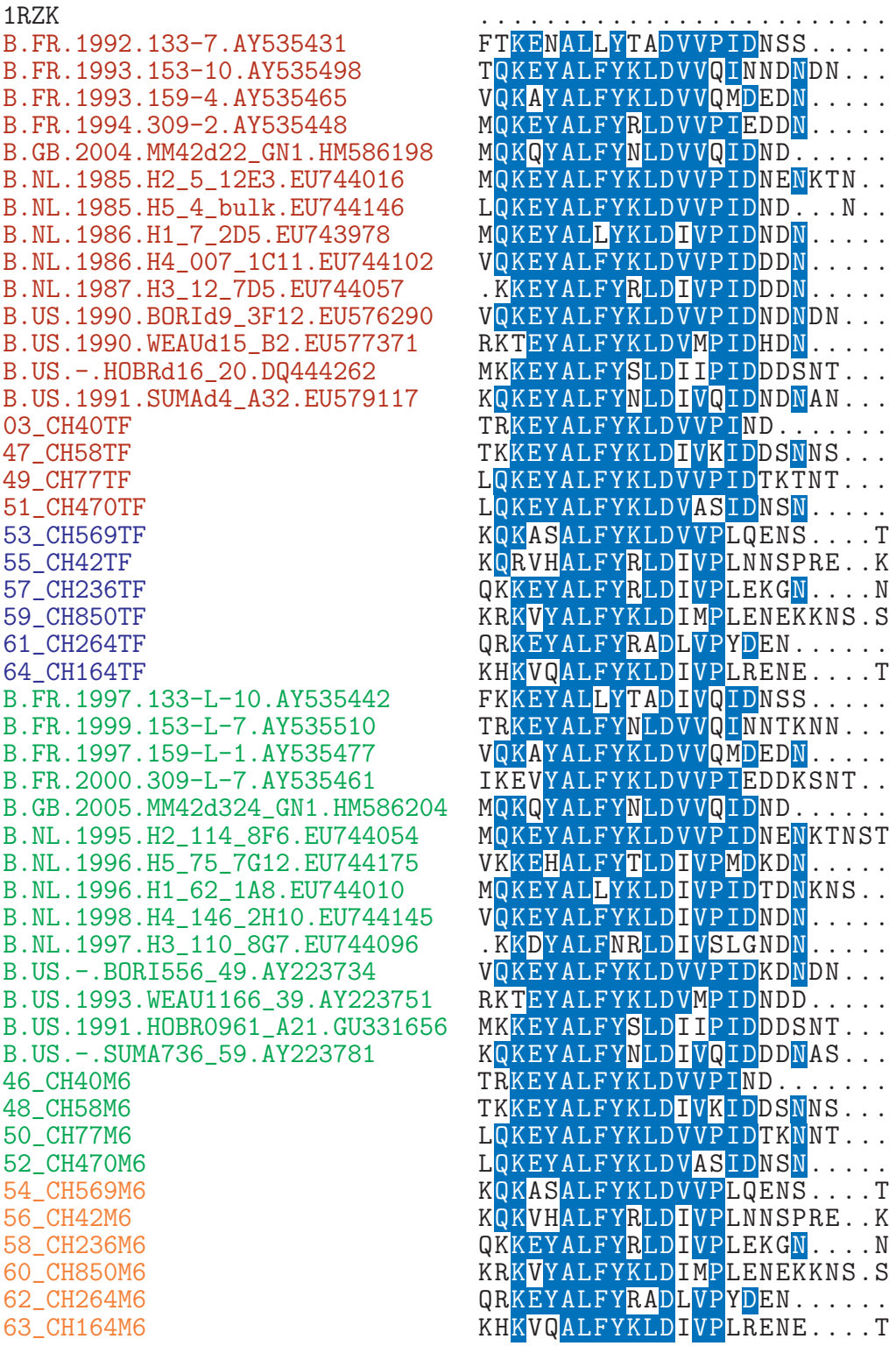

1RZK
B.FR.1992.133-7.AY535431
B.FR.1993.153-10.AY535498
B.FR.1993.159-4.AY535465
B.FR.1994.309-2.AY535448
B.GB.2004.MM42d22_GN1.HM586198
B.NL.1985.H2_5_12E3.EU744016
B.NL.1985.H5_4_bulk.EU744146
B.NL.1986.H1_7_2D5.EU743978
B.NL.1986.H4_007_1C11.EU744102
B.NL.1987.H3_12_7D5.EU744057
B.US.1990.BORId9_3F12.EU576290
B.US.1990.WEAUd15_B2.EU577371
B.US.-.HOBRd16_20.DQ444262
B.US.1991.SUMAd4_A32.EU579117
03_CH40TF
47_CH58TF
49_CH77TF
51_CH470TF
53_CH569TF
55_CH42TF
57_CH236TF
59_CH850TF
61_CH264TF
64_CH164TF
B.FR.1997.133-L-10.AY535442
B.FR.1999.153-L-7.AY535510
B.FR.1997.159-L-1.AY535477
B.FR.2000.309-L-7.AY535461
B.GB.2005.MM42d324_GN1.HM586204
B.NL.1995.H2_114_8F6.EU744054
B.NL.1996.H5_75_7G12.EU744175
B.NL.1996.H1_62_1A8.EU744010
B.NL.1998.H4_146_2H10.EU744145
B.NL.1997.H3_110_8G7.EU744096
B.US.-.BORI556_49.AY223734
B.US.1993.WEAU1166_39.AY223751
B.US.1991.HOBR0961_A21.GU331656
B.US.-.SUMA736_59.AY223781
46_CH40M6
48_CH58M6
50_CH77M6
52_CH470M6
54_CH569M6
56_CH42M6
58_CH236M6
60_CH850M6
62_CH264M6
63_CH164M6

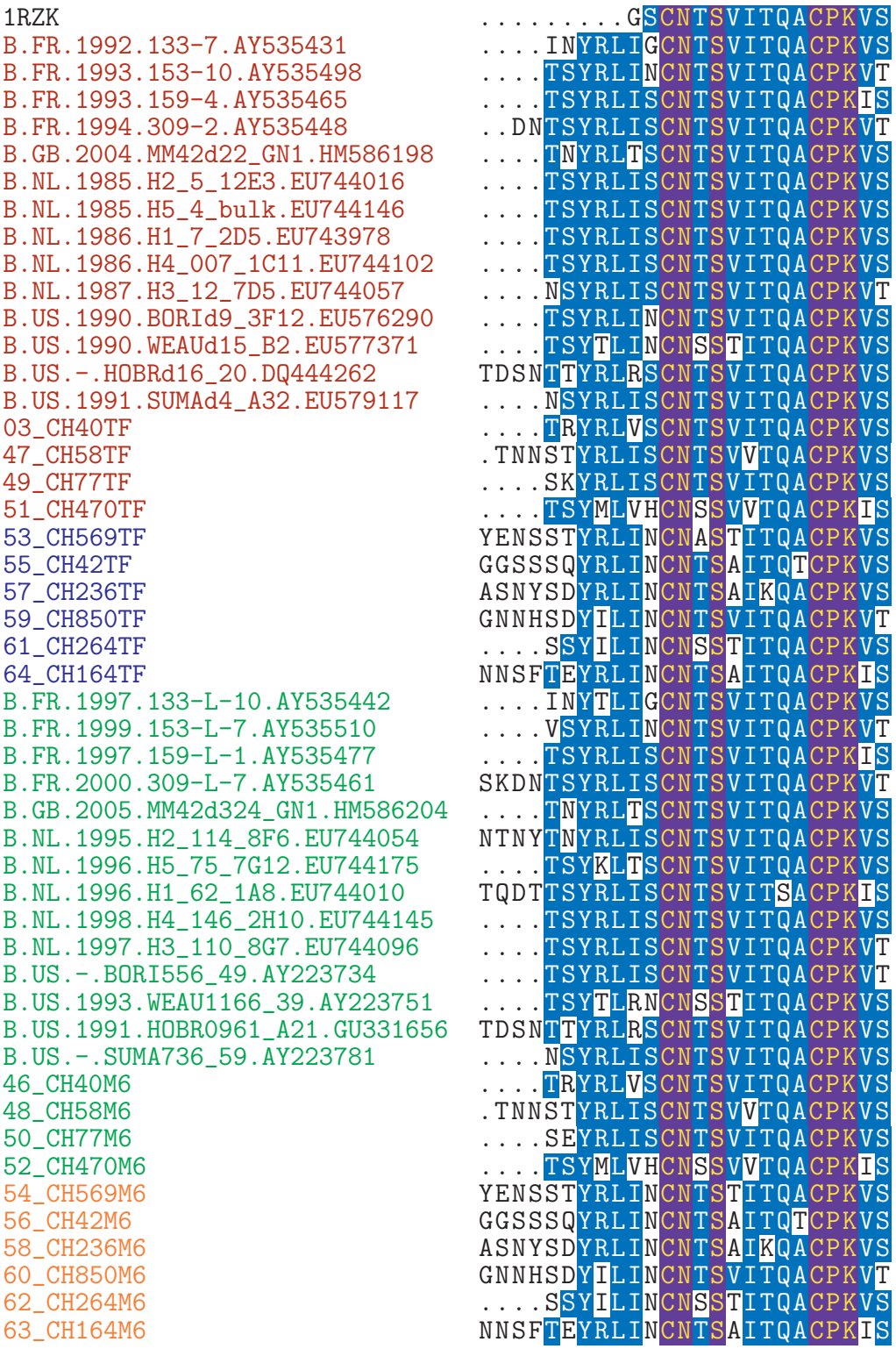

1RZK ...LENVTENFNMWKNNMVEQMHED
B.FR.1992.133-7.AY535431 EVVMGNVTEEFNIWNNSMVEQMHED
B.FR.1993.153-10.AY535498 EVVLENVTENFNMWKNNMVEQMHED
B.FR.1993.159-4.AY535465 EVVLENVTENFNMWKNNMAEQMHED
B.FR.1994.309-2.AY535448 EIELKNVTEEFNMWKNNMVEQMHED
B.GB.2004.MM42d22_GN1.HM586198 EIKLENVTENFNMWKNNMVEQMQED
B.NL.1985.H2_5_12E3.EU744016 EVLLGNVTENFNMWKNNMVEQMHED
B.NL.1985.H5_4_bulk.EU744146 EVVLENVTENFNMWKNGMVEQMHED
B.NL.1986.H1_7_2D5.EU743978 EVKLENVTENFNMWKNNMVEQMHED
B.NL.1986.H4_007_1C11.EU744102 EVELGNVTENFNMWKNNMVEQMHED
B.NL.1987.H3_12_7D5.EU744057 EIALENVTEDFNMWKNNMVEQMHED
B.US.1990.BORId9_3F12.EU576290 EVVLKNVTEDFNMWKNNMVEQMHED
B.US.1990.WEAUd15_B2.EU577371 EVVLENVTENFNMWKNNMVEQMHED
B.US.-.HOBRd16_20.DQ444262 EVVLENVTESFNMWKNNMVDQMHED
B.US.1991.SUMAd4_A32.EU579117 EVVLENVTENFNMWKNNMVEQMHED
03_CH40TF EVELKNVTENFNMWENNMVEQMHED
47_CH58TF EIVLANVTENFNMWENNMVDQMHED
49_CH77TF EVELENVTENFNMWKNNMVEQMHED
51_CH470TF EVKLENVTENFNMWKNKMVDQMHED
53_CH569TF EKVLGNVTENFNMWKNYMADQMHED
55_CH42TF EINLGNVTENFNMWKNHMVDQMHED
57_CH236TF EMVLRNVTENFNMWKNDMVDQMHED
59_CH850TF EMMLKNVTENFNMWKNNMVDQMHED
61_CH264TF EIFLENVTENFNMWKNDMVDQMHED
64_CH164TF EMDLENVTENFNMWKNDMVDQMHED
B.FR.1997.133-L-10.AY535442 EVVMGNVTEEFNIWSNSMVEQMHED
B.FR.1999.153-L-7.AY535510 EVVLGNVTENFNMWKNNMVEQMHED
B.FR.1997.159-L-1.AY535477 EVVLENVTENFNMWKNNMAEQMHED
B.FR.2000.309-L-7.AY535461 EIELKNVTEEFNMWKNNMVEQMHED
B.GB.2005.MM42d324_GN1.HM586204 EIKLENVTENFNMWTNNMVEQMQED
B.NL.1995.H2_114_8F6.EU744054 EILLKNVTENFNMWKNNMVEQMHED
B.NL.1996.H5_75_7G12.EU744175 EVVLGNVTENFNMWKNDMVDQMHED
B.NL.1996.H1_62_1A8.EU744010 EIKMENVTENFNMWKNNMVEQMHED
B.NL.1998.H4_146_2H10.EU744145 EIGLENVTENFNMWKNNMVEQMHED
B.NL.1997.H3_110_8G7.EU744096 EIVLENVTEDFNMWKNNMVEQMHED
B.US.-.BORI556_49.AY223734 EVVLTNVTENFNMWKNNMVEQMHED
B.US.1993.WEAU1166_39.AY223751 EVVMENVTENFNMWKNNMVEQMHED
B.US.1991.HOBR0961_A21.GU331656 EVVLENVTESFNMWKNNMVDQMHED
B.US.-.SUMA736_59.AY223781 EVVL.NVTENFNMWKNNMVEQMHED
46_CH40M6 EVELKNVTENFNMWENNMVEQMHED
48_CH58M6 EIVLANVTENFNMWENNMVDQMHED
50_CH77M6 EVELENVTENFNMWKNNMVEQMHED
52_CH470M6 EVKLENVTENFNMWKNKMVDQMHED
54_CH569M6 EKVLGNVTENFNMWKNYMADQMHED
56_CH42M6 EINLGNVTENFNMWKNHMVDQMHED
58_CH236M6 EMVLRNVTENFNMWKNDMVDQMHED
60_CH850M6 EMMLKNVTENFNMWKNNMVDQMHED
62_CH264M6 EIFLENVTENFNMWKNDMVDQMHED
63_CH164M6 EMDLENVTENFNMWKNDMVDQMHED

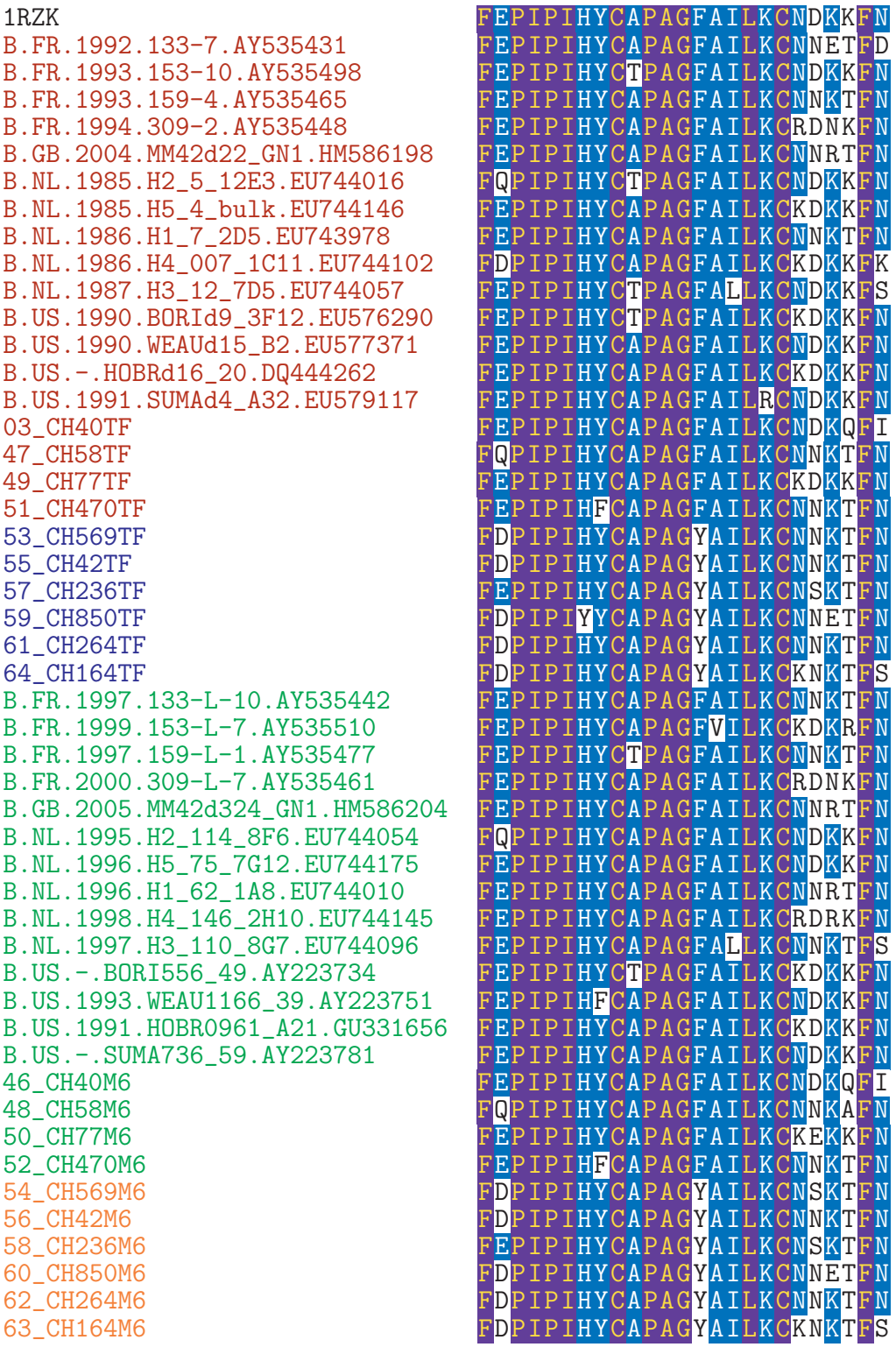

1RZK
B.FR.1992.133-7.AY535431
B.FR.1993.153-10.AY535498
B.FR.1993.159-4.AY535465
B.FR.1994.309-2.AY535448
B.GB.2004.MM42d22_GN1.HM586198
B.NL.1985.H2_5_12E3.EU744016
B.NL.1985.H5_4_bulk.EU744146
B.NL.1986.H1_7_2D5.EU743978
B.NL.1986.H4_007_1C11.EU744102
B.NL.1987.H3_12_7D5.EU744057
B.US.1990.BORId9_3F12.EU576290
B.US.1990.WEAUd15_B2.EU577371
B.US.-.HOBRd16_20.DQ444262
B.US.1991.SUMAd4_A32.EU579117
03_CH40TF
47_CH58TF
49_CH77TF
51_CH470TF
53_CH569TF
55_CH42TF
57_CH236TF
59_CH850TF
61_CH264TF
64_CH164TF
B.FR.1997.133-L-10.AY535442
B.FR.1999.153-L-7.AY535510
B.FR.1997.159-L-1.AY535477
B.FR.2000.309-L-7.AY535461
B.GB.2005.MM42d324_GN1.HM586204
B.NL.1995.H2_114_8F6.EU744054
B.NL.1996.H5_75_7G12.EU744175
B.NL.1996.H1_62_1A8.EU744010
B.NL.1998.H4_146_2H10.EU744145
B.NL.1997.H3_110_8G7.EU744096
B.US.-.BORI556_49.AY223734
B.US.1993.WEAU1166_39.AY223751
B.US.1991.HOBR0961_A21.GU331656
B.US.-.SUMA736_59.AY223781
46_CH40M6
48_CH58M6
50_CH77M6
52_CH470M6
54_CH569M6
56_CH42M6
58_CH236M6
60_CH850M6
62_CH264M6
63_CH164M6

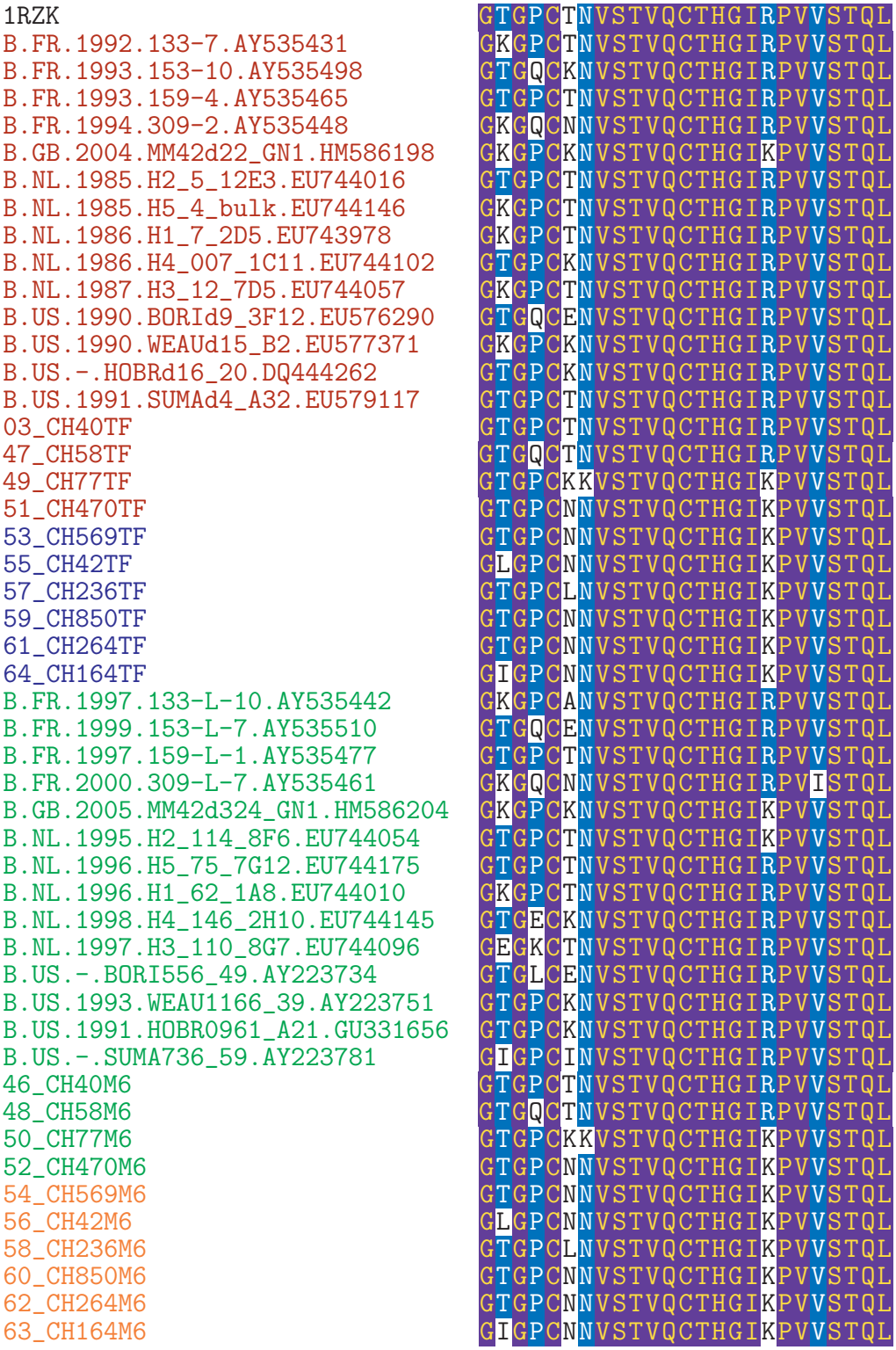
```
1RZK                                .........................
B.FR.1992.133-7.AY535431            NYNSTSND.................
B.FR.1993.153-10.AY535498           NAGNITNN.................
B.FR.1993.159-4.AY535465            DLGNATN..................
B.FR.1994.309-2.AY535448            DVKANST.NT............TNS
B.GB.2004.MM42d22_GN1.HM586198      DVNSTRTNSSS.............G
B.NL.1985.H2_5_12E3.EU744016        DLGNATNTTNS..............
B.NL.1985.H5_4_bulk.EU744146        DLMNTTNTNSS..............
B.NL.1986.H1_7_2D5.EU743978         DLRNATNNS................
B.NL.1986.H4_007_1C11.EU744102      DLNNATNTPNS..............
B.NL.1987.H3_12_7D5.EU744057        ELENTINIT................
B.US.1990.BORId9_3F12.EU576290      DHLWNVTNTMR.......NATNTTS
B.US.1990.WEAUd15_B2.EU577371       NVNVTNLKN........ETNTNSSS
B.US.-.HOBRd16_20.DQ444262          HNVTATNG.................
B.US.1991.SUMAd4_A32.EU579117       DYVKNVTNATST.............
03_CH40TF                           DLGNVTNTTN.............S
47_CH58TF                           ELNNNSTTTT...............
49_CH77TF                           DSNGDSSIAN...............
51_CH470TF                          DASAGNGTDAIA........NNGTN
53_CH569TF                          NTVKGNKS.................
55_CH42TF                           NAKNDNATVD...............
57_CH236TF                          NANITNTNANSTNS....TSTNANK
59_CH850TF                          NADVNFTSYN...............
61_CH264TF                          SVKNNSTACNSTAS....NSTA...
64_CH164TF                          TAIAHNASN................
B.FR.1997.133-L-10.AY535442         NYNGTRNGTTTEPP.........EV
B.FR.1999.153-L-7.AY535510          NVNSNITN.................
B.FR.1997.159-L-1.AY535477          DLGNATN..................
B.FR.2000.309-L-7.AY535461          DVKANSTTNT............TNS
B.GB.2005.MM42d324_GN1.HM586204     DVNSTRTNSSS.............G
B.NL.1995.H2_114_8F6.EU744054       DVRNATNTTNS..............
B.NL.1996.H5_75_7G12.EU744175       DYLGNGTNITI.............T
B.NL.1996.H1_62_1A8.EU744010        DPRNDTSNSTI..........NYGN
B.NL.1998.H4_146_2H10.EU744145      DLNNNATDLNN..........TNSA
B.NL.1997.H3_110_8G7.EU744096       ELQINDTSVTSGNKTDSNNSTSNNR
B.US.-.BORI556_49.AY223734          DL........K.......NATNTTI
B.US.1993.WEAU1166_39.AY223751      NVNVTNLKN........ETNTNSRS
B.US.1991.HOBR0961_A21.GU331656     HNVTATNG.................
B.US.-.SUMA736_59.AY223781          DYVKNVTNATST.............
46_CH40M6                           DLGNVTNTTN.............S
48_CH58M6                           ELNNNSTTTT...............
50_CH77M6                           DSNGDSSIAN...............
52_CH470M6                          DASAGNGTDAIA........NNGTN
54_CH569M6                          DTVKGNKS.................
56_CH42M6                           NAKNDNATVD...............
58_CH236M6                          NANITNTNANSTNS....TSTNANK
60_CH850M6                          NADVNFTSYN...............
62_CH264M6                          SVKNNSTACNSTAS....NSTA...
63_CH164M6                          TAIAHNASN................
```

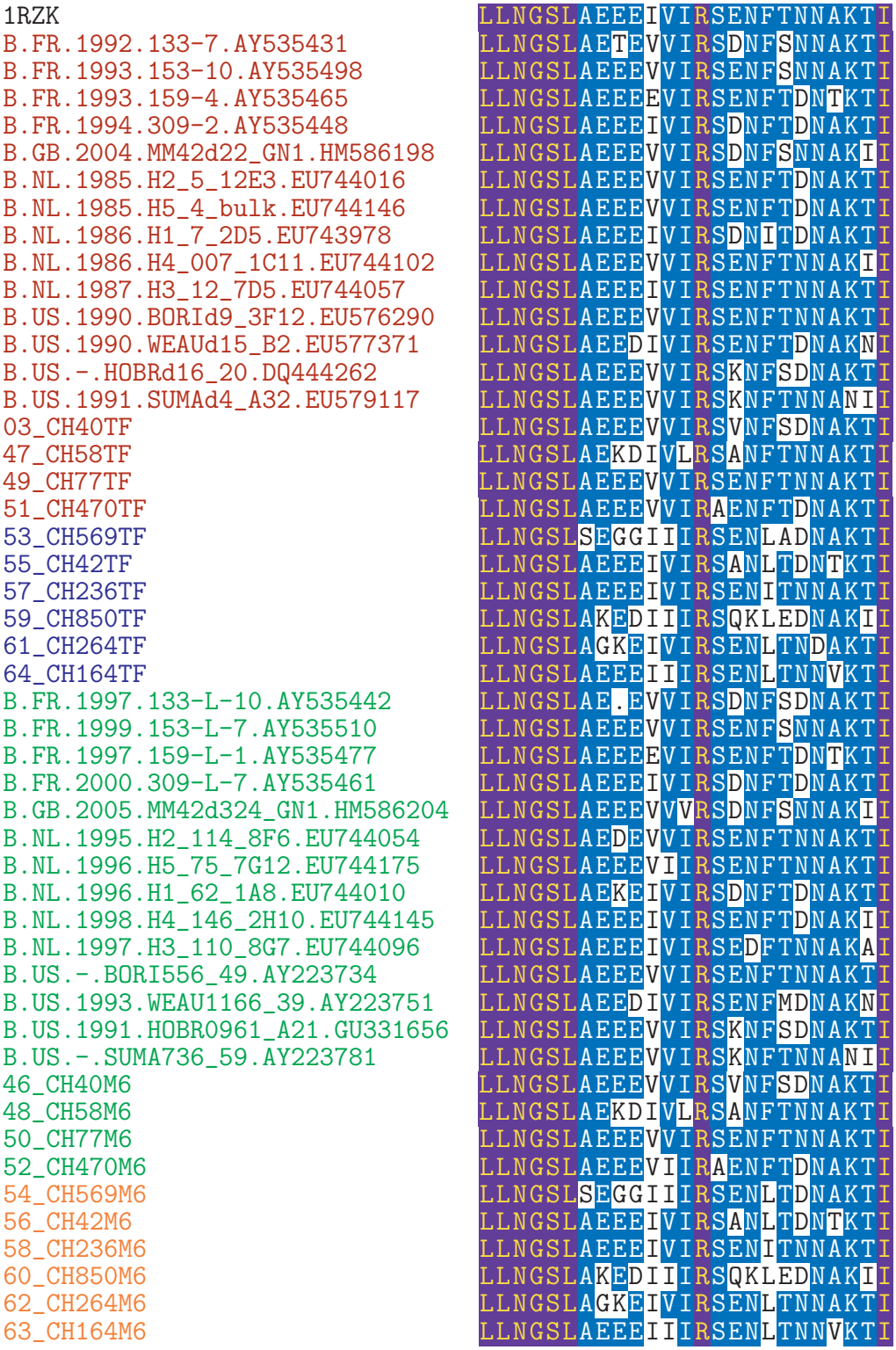

1RZK
B.FR.1992.133-7.AY535431
B.FR.1993.153-10.AY535498
B.FR.1993.159-4.AY535465
B.FR.1994.309-2.AY535448
B.GB.2004.MM42d22_GN1.HM586198
B.NL.1985.H2_5_12E3.EU744016
B.NL.1985.H5_4_bulk.EU744146
B.NL.1986.H1_7_2D5.EU743978
B.NL.1986.H4_007_1C11.EU744102
B.NL.1987.H3_12_7D5.EU744057
B.US.1990.BORId9_3F12.EU576290
B.US.1990.WEAUd15_B2.EU577371
B.US.-.HOBRd16_20.DQ444262
B.US.1991.SUMAd4_A32.EU579117
03_CH40TF
47_CH58TF
49_CH77TF
51_CH470TF
53_CH569TF
55_CH42TF
57_CH236TF
59_CH850TF
61_CH264TF
64_CH164TF
B.FR.1997.133-L-10.AY535442
B.FR.1999.153-L-7.AY535510
B.FR.1997.159-L-1.AY535477
B.FR.2000.309-L-7.AY535461
B.GB.2005.MM42d324_GN1.HM586204
B.NL.1995.H2_114_8F6.EU744054
B.NL.1996.H5_75_7G12.EU744175
B.NL.1996.H1_62_1A8.EU744010
B.NL.1998.H4_146_2H10.EU744145
B.NL.1997.H3_110_8G7.EU744096
B.US.-.BORI556_49.AY223734
B.US.1993.WEAU1166_39.AY223751
B.US.1991.HOBR0961_A21.GU331656
B.US.-.SUMA736_59.AY223781
46_CH40M6
48_CH58M6
50_CH77M6
52_CH470M6
54_CH569M6
56_CH42M6
58_CH236M6
60_CH850M6
62_CH264M6
63_CH164M6

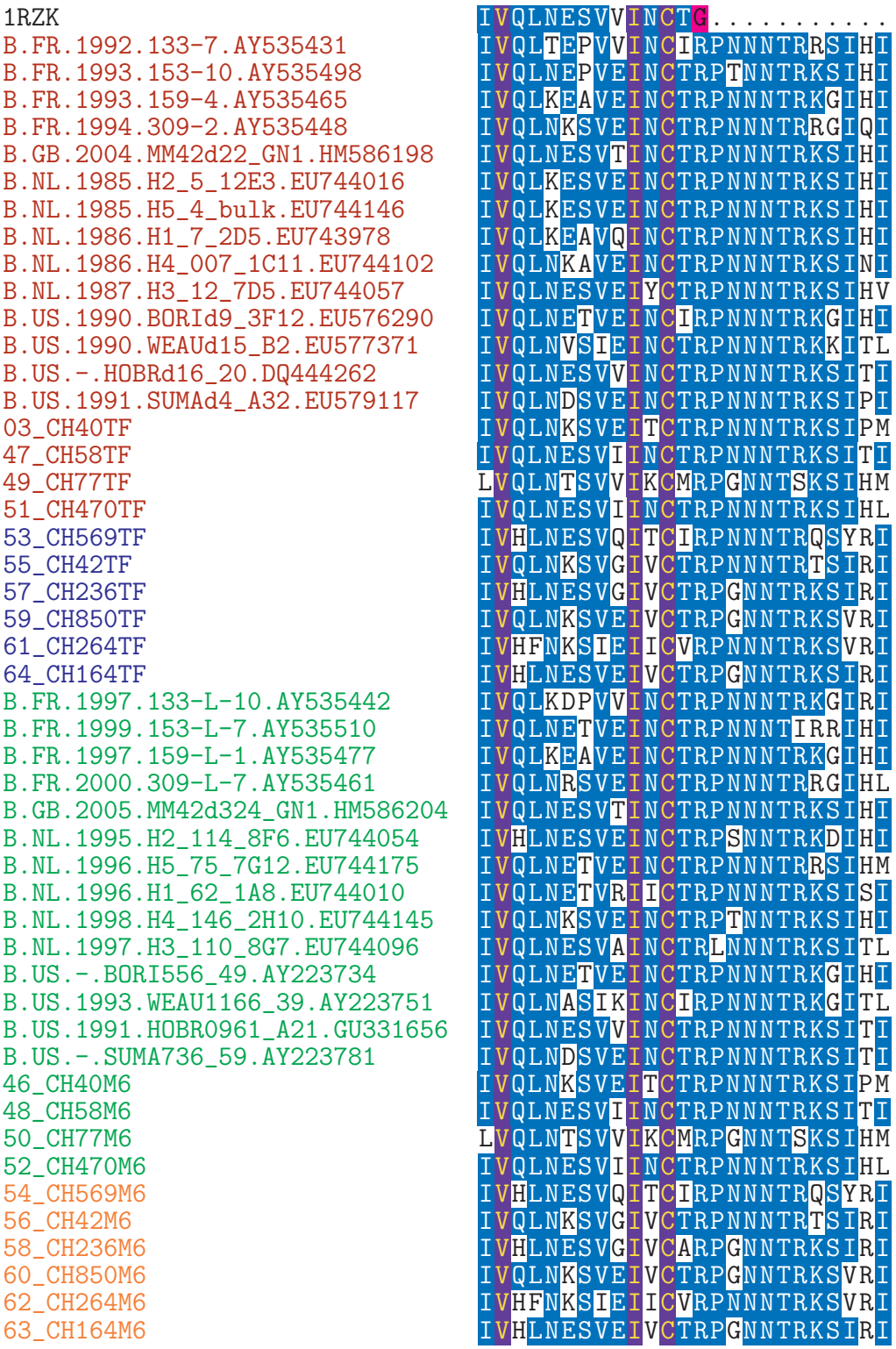

1RZK
B.FR.1992.133-7.AY535431
B.FR.1993.153-10.AY535498
B.FR.1993.159-4.AY535465
B.FR.1994.309-2.AY535448
B.GB.2004.MM42d22_GN1.HM586198
B.NL.1985.H2_5_12E3.EU744016
B.NL.1985.H5_4_bulk.EU744146
B.NL.1986.H1_7_2D5.EU743978
B.NL.1986.H4_007_1C11.EU744102
B.NL.1987.H3_12_7D5.EU744057
B.US.1990.BORId9_3F12.EU576290
B.US.1990.WEAUd15_B2.EU577371
B.US.-.HOBRd16_20.DQ444262
B.US.1991.SUMAd4_A32.EU579117
03_CH40TF
47_CH58TF
49_CH77TF
51_CH470TF
53_CH569TF
55_CH42TF
57_CH236TF
59_CH850TF
61_CH264TF
64_CH164TF
B.FR.1997.133-L-10.AY535442
B.FR.1999.153-L-7.AY535510
B.FR.1997.159-L-1.AY535477
B.FR.2000.309-L-7.AY535461
B.GB.2005.MM42d324_GN1.HM586204
B.NL.1995.H2_114_8F6.EU744054
B.NL.1996.H5_75_7G12.EU744175
B.NL.1996.H1_62_1A8.EU744010
B.NL.1998.H4_146_2H10.EU744145
B.NL.1997.H3_110_8G7.EU744096
B.US.-.BORI556_49.AY223734
B.US.1993.WEAU1166_39.AY223751
B.US.1991.HOBR0961_A21.GU331656
B.US.-.SUMA736_59.AY223781
46_CH40M6
48_CH58M6
50_CH77M6
52_CH470M6
54_CH569M6
56_CH42M6
58_CH236M6
60_CH850M6
62_CH264M6
63_CH164M6

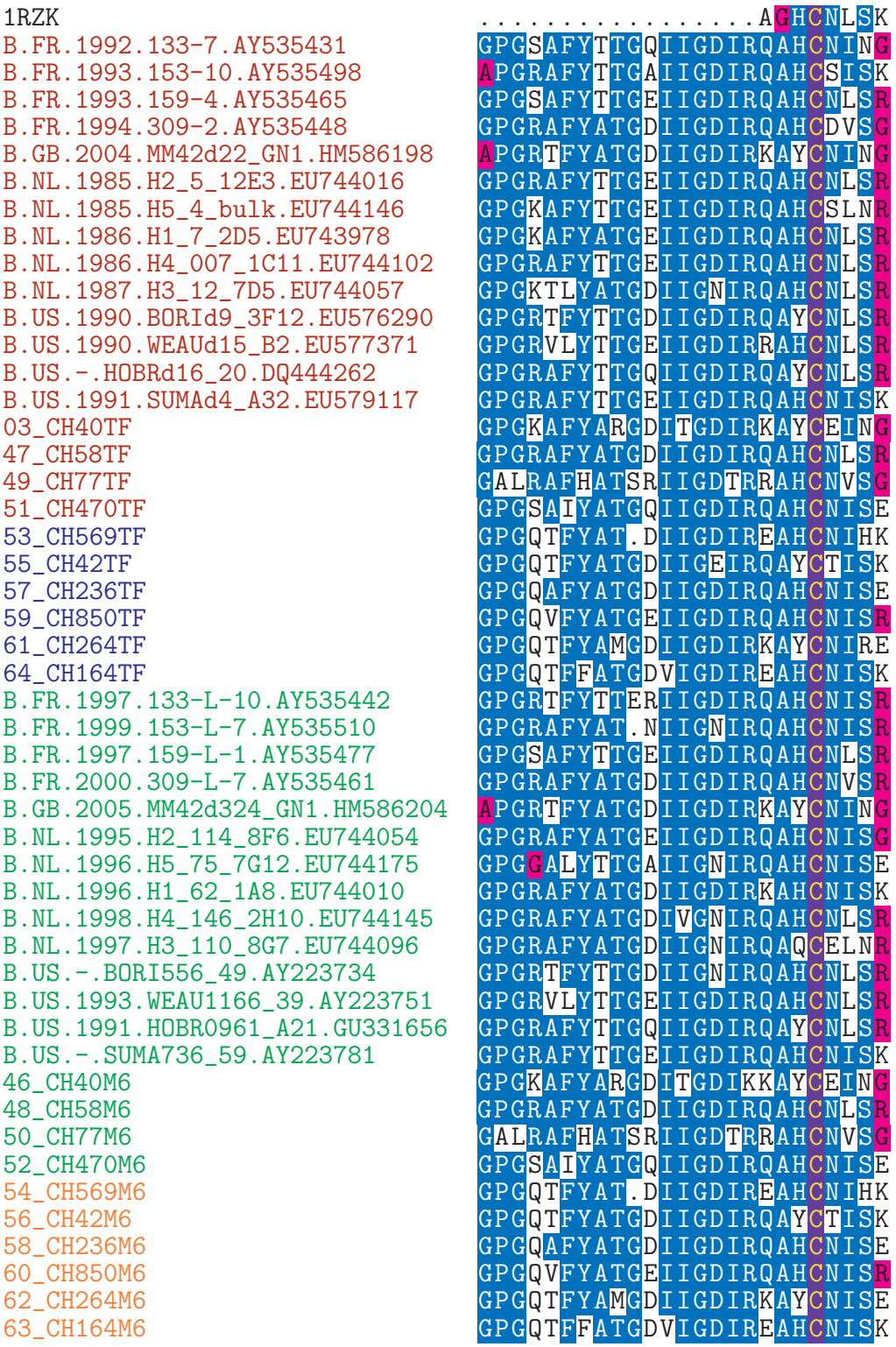

1RZK .....GSCNTSVITQACPKVS
B.FR.1992.133-7.AY535431 ....INYRLIGCNTSVITQACPKVS
B.FR.1993.153-10.AY535498 ....TSYRLINCNTSVITQACPKVT
B.FR.1993.159-4.AY535465 ....TSYRLISCNTSVITQACPKIS
B.FR.1994.309-2.AY535448 ..DNTSYRLISCNTSVITQACPKVT
B.GB.2004.MM42d22_GN1.HM586198 ....TNYRLTSCNTSVITQACPKVS
B.NL.1985.H2_5_12E3.EU744016 ....TSYRLISCNTSVITQACPKVS
B.NL.1985.H5_4_bulk.EU744146 ....TSYRLISCNTSVITQACPKVS
B.NL.1986.H1_7_2D5.EU743978 ....TSYRLISCNTSVITQACPKVS
B.NL.1986.H4_007_1C11.EU744102 ....TSYRLISCNTSVITQACPKVS
B.NL.1987.H3_12_7D5.EU744057 ....NSYRLISCNTSVITQACPKVT
B.US.1990.BORId9_3F12.EU576290 ....TSYRLINCNTSVITQACPKVS
B.US.1990.WEAUd15_B2.EU577371 ....TSYTLINCNSSTITQACPKVS
B.US.-.HOBRd16_20.DQ444262 TDSNTTYRLRSCNTSVITQACPKVS
B.US.1991.SUMAd4_A32.EU579117 ....NSYRLISCNTSVITQACPKVS
03_CH40TF ....TRYRLVSCNTSVITQACPKVS
47_CH58TF .TNNSTYRLISCNTSVVTQACPKVS
49_CH77TF ....SKYRLISCNTSVITQACPKVS
51_CH470TF ....TSYMLVHCNSSVVTQACPKIS
53_CH569TF YENSSTYRLINCNASTITQACPKVS
55_CH42TF GGSSSQYRLINCNTSAITQTCPKVS
57_CH236TF ASNYSDYRLINCNTSAIKQACPKVS
59_CH850TF GNNHSDYILINCNTSVITQACPKVT
61_CH264TF ....SSYILINCNSSTITQACPKVS
64_CH164TF NNSFTEYRLINCNTSAITQACPKIS
B.FR.1997.133-L-10.AY535442 ....INYTLIGCNTSVITQACPKVS
B.FR.1999.153-L-7.AY535510 ....VSYRLINCNTSVITQACPKVT
B.FR.1997.159-L-1.AY535477 ....TSYRLISCNTSVITQACPKIS
B.FR.2000.309-L-7.AY535461 SKDNTSYRLISCNTSVITQACPKVT
B.GB.2005.MM42d324_GN1.HM586204 ....TNYRLTSCNTSVITQACPKVS
B.NL.1995.H2_114_8F6.EU744054 NTNYTNYRLISCNTSVITQACPKVS
B.NL.1996.H5_75_7G12.EU744175 ....TSYKLTSCNTSVITQACPKVS
B.NL.1996.H1_62_1A8.EU744010 TQDTTSYRLISCNTSVITSACPKIS
B.NL.1998.H4_146_2H10.EU744145 ....TSYRLISCNTSVITQACPKVS
B.NL.1997.H3_110_8G7.EU744096 ....TSYRLISCNTSVITQACPKVT
B.US.-.BORI556_49.AY223734 ....TSYRLISCNTSVITQACPKVT
B.US.1993.WEAU1166_39.AY223751 ....TSYTLRNCNSSTITQACPKVS
B.US.1991.HOBR0961_A21.GU331656 TDSNTTYRLRSCNTSVITQACPKVS
B.US.-.SUMA736_59.AY223781 ....NSYRLISCNTSVITQACPKVS
46_CH40M6 ....TRYRLVSCNTSVITQACPKVS
48_CH58M6 .TNNSTYRLISCNTSVVTQACPKVS
50_CH77M6 ....SEYRLISCNTSVITQACPKVS
52_CH470M6 ....TSYMLVHCNSSVVTQACPKIS
54_CH569M6 YENSSTYRLINCNTSTITQACPKVS
56_CH42M6 GGSSSQYRLINCNTSAITQTCPKVS
58_CH236M6 ASNYSDYRLINCNTSAIKQACPKVS
60_CH850M6 GNNHSDYILINCNTSVITQACPKVT
62_CH264M6 ....SSYILINCNSSTITQACPKVS
63_CH164M6 NNSFTEYRLINCNTSAITQACPKIS

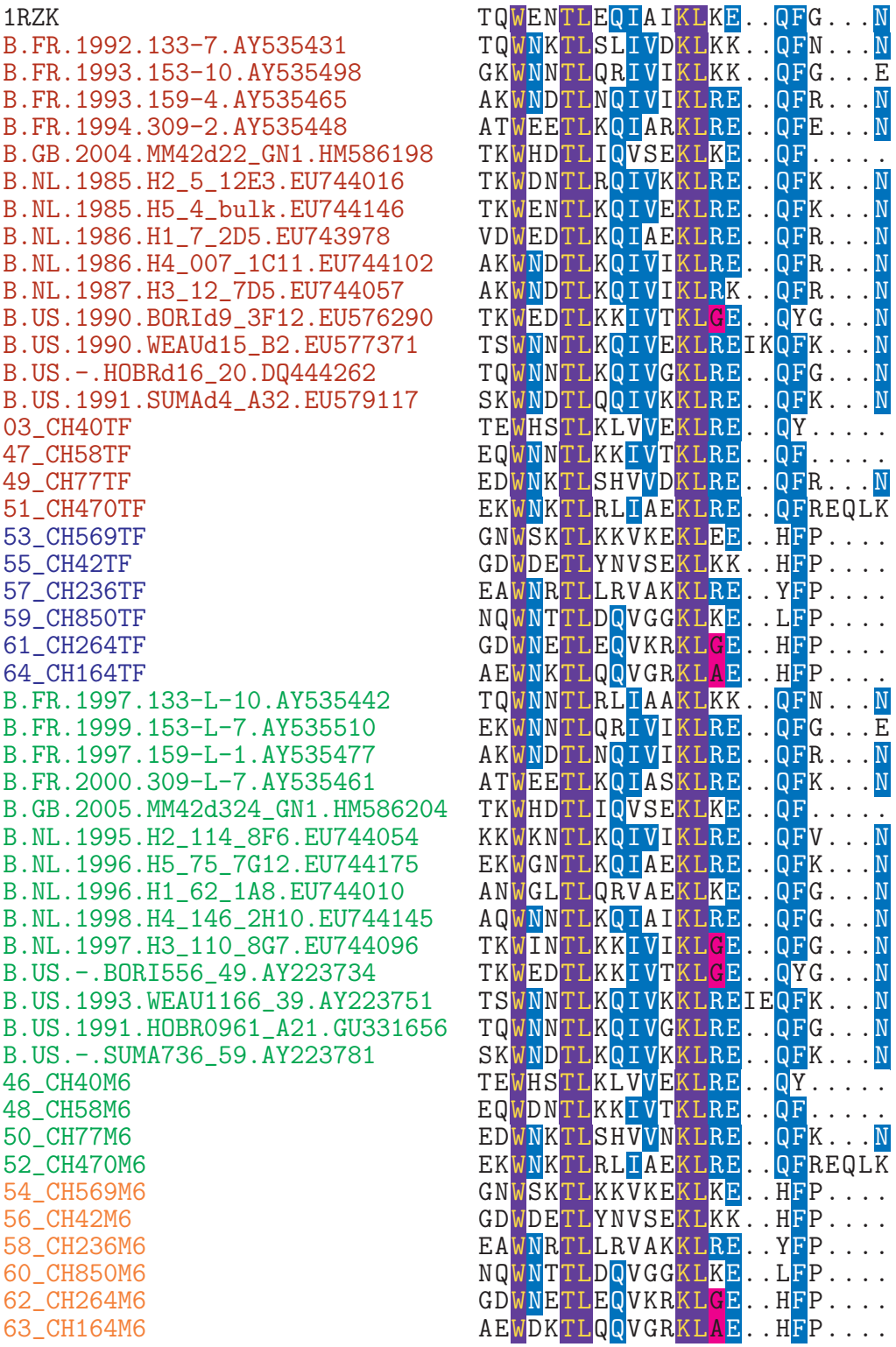

1RZK
B.FR.1992.133-7.AY535431
B.FR.1993.153-10.AY535498
B.FR.1993.159-4.AY535465
B.FR.1994.309-2.AY535448
B.GB.2004.MM42d22_GN1.HM586198
B.NL.1985.H2_5_12E3.EU744016
B.NL.1985.H5_4_bulk.EU744146
B.NL.1986.H1_7_2D5.EU743978
B.NL.1986.H4_007_1C11.EU744102
B.NL.1987.H3_12_7D5.EU744057
B.US.1990.BORId9_3F12.EU576290
B.US.1990.WEAUd15_B2.EU577371
B.US.-.HOBRd16_20.DQ444262
B.US.1991.SUMAd4_A32.EU579117
03_CH40TF
47_CH58TF
49_CH77TF
51_CH470TF
53_CH569TF
55_CH42TF
57_CH236TF
59_CH850TF
61_CH264TF
64_CH164TF
B.FR.1997.133-L-10.AY535442
B.FR.1999.153-L-7.AY535510
B.FR.1997.159-L-1.AY535477
B.FR.2000.309-L-7.AY535461
B.GB.2005.MM42d324_GN1.HM586204
B.NL.1995.H2_114_8F6.EU744054
B.NL.1996.H5_75_7G12.EU744175
B.NL.1996.H1_62_1A8.EU744010
B.NL.1998.H4_146_2H10.EU744145
B.NL.1997.H3_110_8G7.EU744096
B.US.-.BORI556_49.AY223734
B.US.1993.WEAU1166_39.AY223751
B.US.1991.HOBR0961_A21.GU331656
B.US.-.SUMA736_59.AY223781
46_CH40M6
48_CH58M6
50_CH77M6
52_CH470M6
54_CH569M6
56_CH42M6
58_CH236M6
60_CH850M6
62_CH264M6
63_CH164M6

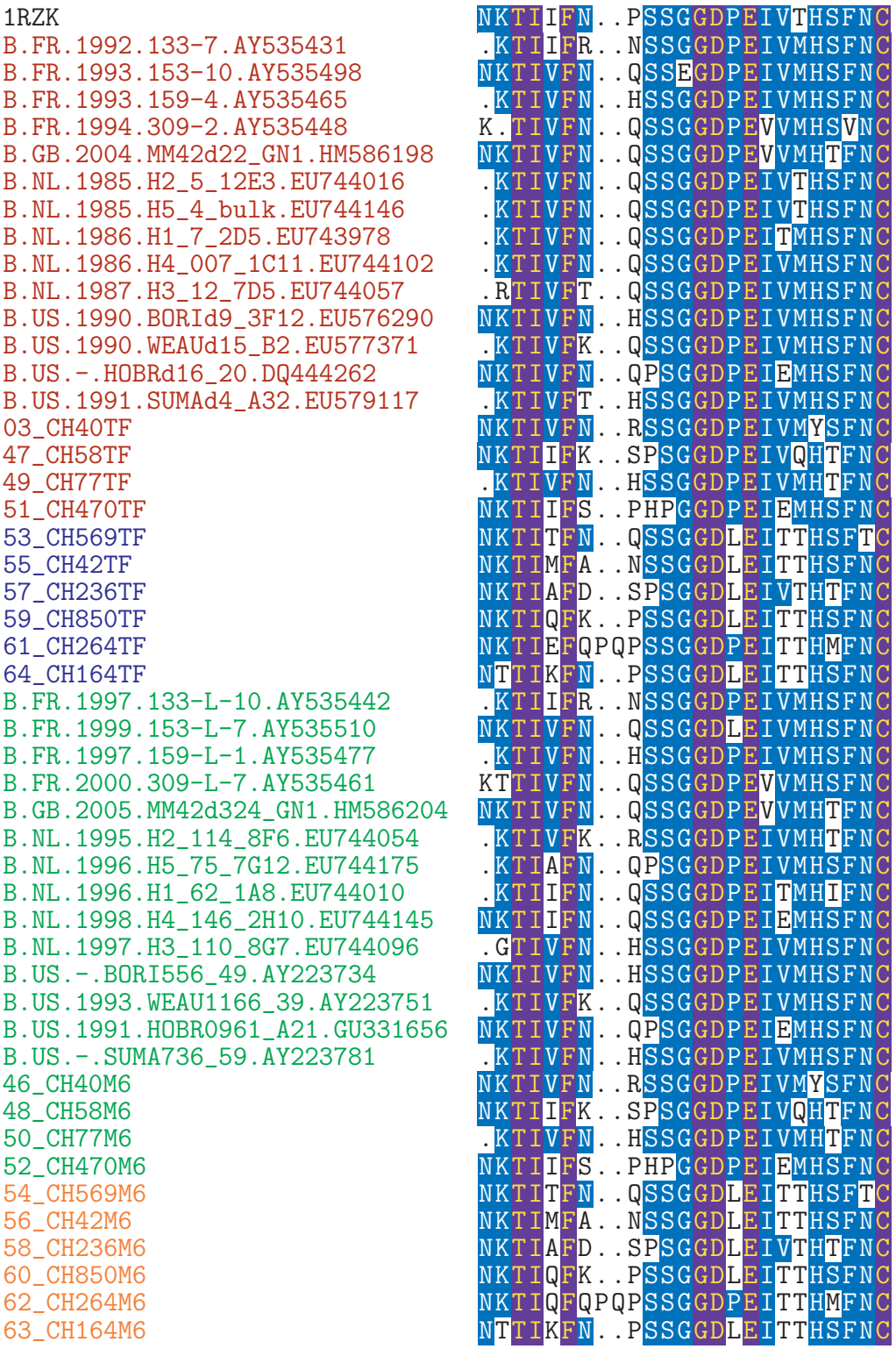

1RZK
B.FR.1992.133-7.AY535431
B.FR.1993.153-10.AY535498
B.FR.1993.159-4.AY535465
B.FR.1994.309-2.AY535448
B.GB.2004.MM42d22_GN1.HM586198
B.NL.1985.H2_5_12E3.EU744016
B.NL.1985.H5_4_bulk.EU744146
B.NL.1986.H1_7_2D5.EU743978
B.NL.1986.H4_007_1C11.EU744102
B.NL.1987.H3_12_7D5.EU744057
B.US.1990.BORId9_3F12.EU576290
B.US.1990.WEAUd15_B2.EU577371
B.US.-.HOBRd16_20.DQ444262
B.US.1991.SUMAd4_A32.EU579117
03_CH40TF
47_CH58TF
49_CH77TF
51_CH470TF
53_CH569TF
55_CH42TF
57_CH236TF
59_CH850TF
61_CH264TF
64_CH164TF
B.FR.1997.133-L-10.AY535442
B.FR.1999.153-L-7.AY535510
B.FR.1997.159-L-1.AY535477
B.FR.2000.309-L-7.AY535461
B.GB.2005.MM42d324_GN1.HM586204
B.NL.1995.H2_114_8F6.EU744054
B.NL.1996.H5_75_7G12.EU744175
B.NL.1996.H1_62_1A8.EU744010
B.NL.1998.H4_146_2H10.EU744145
B.NL.1997.H3_110_8G7.EU744096
B.US.-.BORI556_49.AY223734
B.US.1993.WEAU1166_39.AY223751
B.US.1991.HOBR0961_A21.GU331656
B.US.-.SUMA736_59.AY223781
46_CH40M6
48_CH58M6
50_CH77M6
52_CH470M6
54_CH569M6
56_CH42M6
58_CH236M6
60_CH850M6
62_CH264M6
63_CH164M6

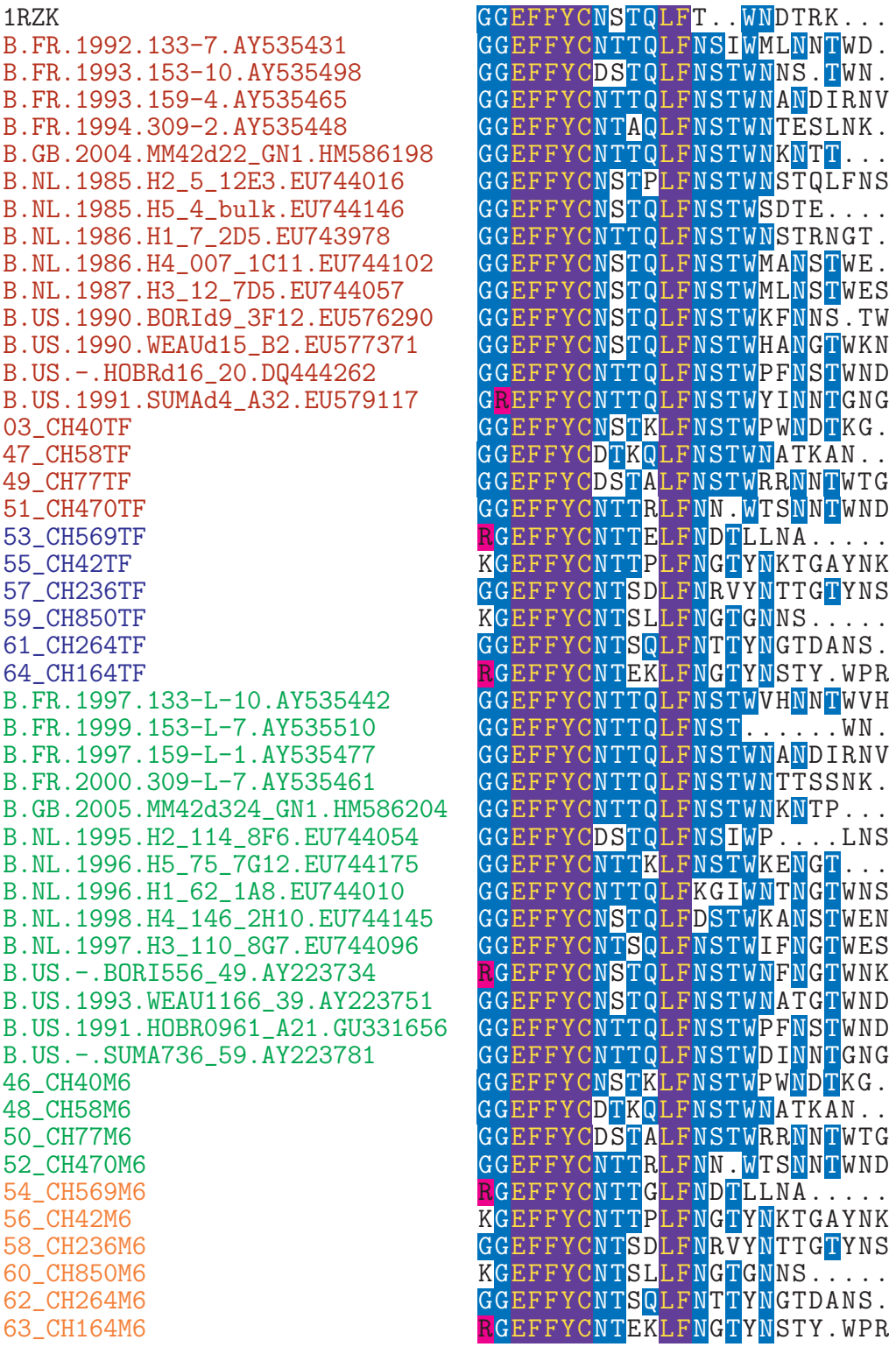

1RZK
B.FR.1992.133-7.AY535431
B.FR.1993.153-10.AY535498
B.FR.1993.159-4.AY535465
B.FR.1994.309-2.AY535448
B.GB.2004.MM42d22_GN1.HM586198
B.NL.1985.H2_5_12E3.EU744016
B.NL.1985.H5_4_bulk.EU744146
B.NL.1986.H1_7_2D5.EU743978
B.NL.1986.H4_007_1C11.EU744102
B.NL.1987.H3_12_7D5.EU744057
B.US.1990.BORId9_3F12.EU576290
B.US.1990.WEAUd15_B2.EU577371
B.US.-.HOBRd16_20.DQ444262
B.US.1991.SUMAd4_A32.EU579117
03_CH40TF
47_CH58TF
49_CH77TF
51_CH470TF
53_CH569TF
55_CH42TF
57_CH236TF
59_CH850TF
61_CH264TF
64_CH164TF
B.FR.1997.133-L-10.AY535442
B.FR.1999.153-L-7.AY535510
B.FR.1997.159-L-1.AY535477
B.FR.2000.309-L-7.AY535461
B.GB.2005.MM42d324_GN1.HM586204
B.NL.1995.H2_114_8F6.EU744054
B.NL.1996.H5_75_7G12.EU744175
B.NL.1996.H1_62_1A8.EU744010
B.NL.1998.H4_146_2H10.EU744145
B.NL.1997.H3_110_8G7.EU744096
B.US.-.BORI556_49.AY223734
B.US.1993.WEAU1166_39.AY223751
B.US.1991.HOBR0961_A21.GU331656
B.US.-.SUMA736_59.AY223781
46_CH40M6
48_CH58M6
50_CH77M6
52_CH470M6
54_CH569M6
56_CH42M6
58_CH236M6
60_CH850M6
62_CH264M6
63_CH164M6

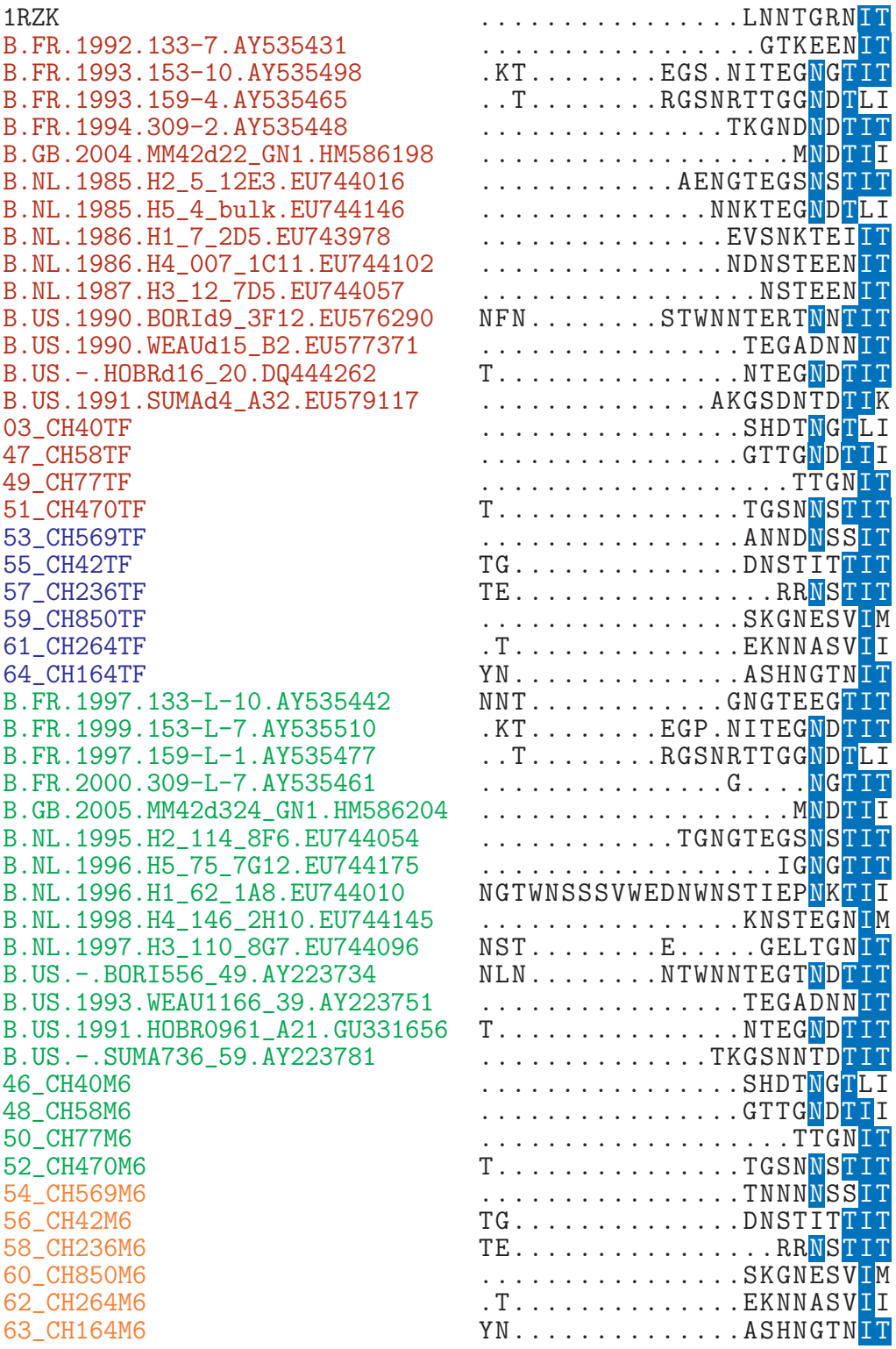

1RZK
B.FR.1992.133-7.AY535431
B.FR.1993.153-10.AY535498
B.FR.1993.159-4.AY535465
B.FR.1994.309-2.AY535448
B.GB.2004.MM42d22_GN1.HM586198
B.NL.1985.H2_5_12E3.EU744016
B.NL.1985.H5_4_bulk.EU744146
B.NL.1986.H1_7_2D5.EU743978
B.NL.1986.H4_007_1C11.EU744102
B.NL.1987.H3_12_7D5.EU744057
B.US.1990.BORId9_3F12.EU576290
B.US.1990.WEAUd15_B2.EU577371
B.US.-.HOBRd16_20.DQ444262
B.US.1991.SUMAd4_A32.EU579117
03_CH40TF
47_CH58TF
49_CH77TF
51_CH470TF
53_CH569TF
55_CH42TF
57_CH236TF
59_CH850TF
61_CH264TF
64_CH164TF
B.FR.1997.133-L-10.AY535442
B.FR.1999.153-L-7.AY535510
B.FR.1997.159-L-1.AY535477
B.FR.2000.309-L-7.AY535461
B.GB.2005.MM42d324_GN1.HM586204
B.NL.1995.H2_114_8F6.EU744054
B.NL.1996.H5_75_7G12.EU744175
B.NL.1996.H1_62_1A8.EU744010
B.NL.1998.H4_146_2H10.EU744145
B.NL.1997.H3_110_8G7.EU744096
B.US.-.BORI556_49.AY223734
B.US.1993.WEAU1166_39.AY223751
B.US.1991.HOBR0961_A21.GU331656
B.US.-.SUMA736_59.AY223781
46_CH40M6
48_CH58M6
50_CH77M6
52_CH470M6
54_CH569M6
56_CH42M6
58_CH236M6
60_CH850M6
62_CH264M6
63_CH164M6

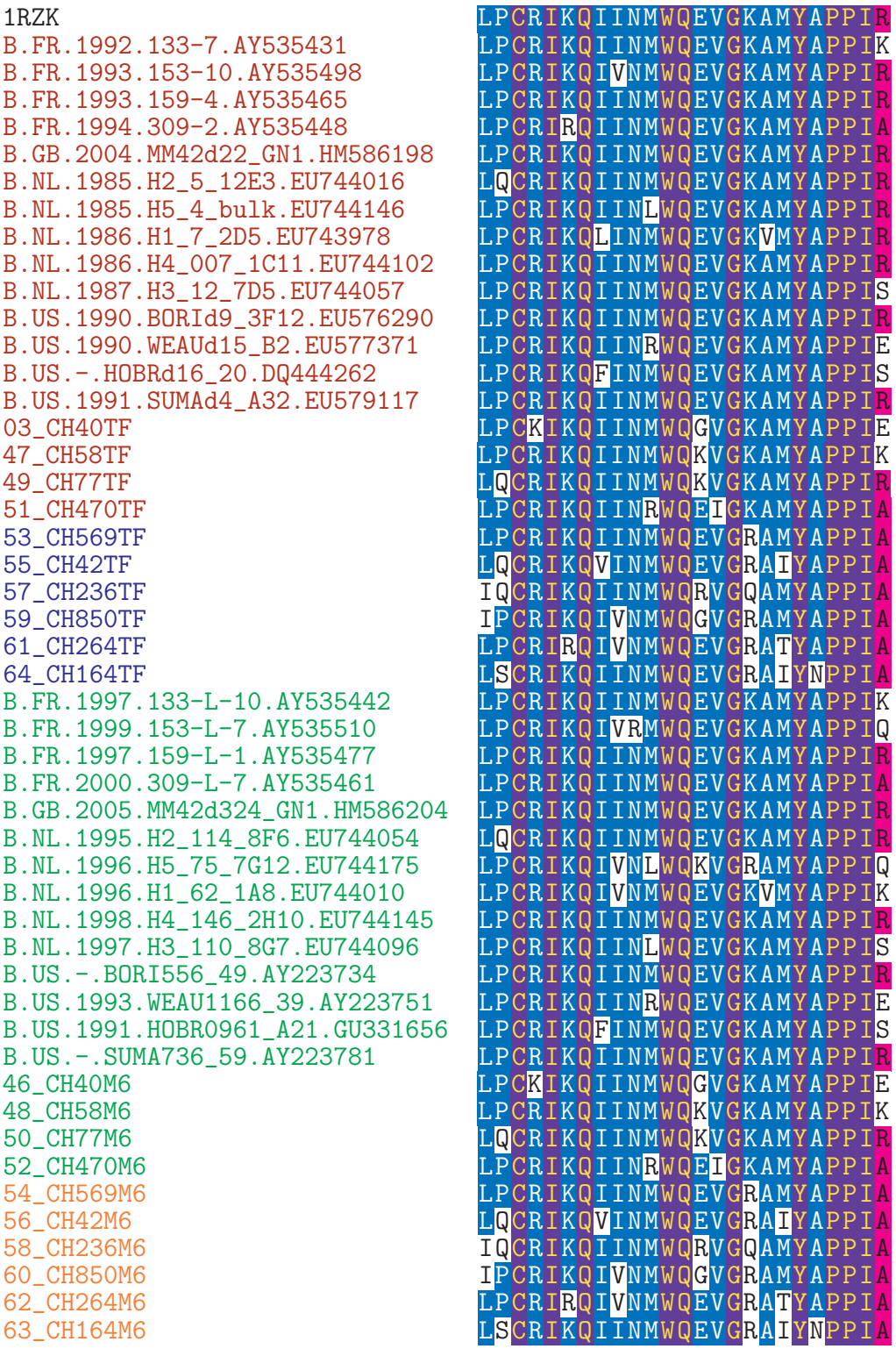

1RZK
B.FR.1992.133-7.AY535431
B.FR.1993.153-10.AY535498
B.FR.1993.159-4.AY535465
B.FR.1994.309-2.AY535448
B.GB.2004.MM42d22_GN1.HM586198
B.NL.1985.H2_5_12E3.EU744016
B.NL.1985.H5_4_bulk.EU744146
B.NL.1986.H1_7_2D5.EU743978
B.NL.1986.H4_007_1C11.EU744102
B.NL.1987.H3_12_7D5.EU744057
B.US.1990.BORId9_3F12.EU576290
B.US.1990.WEAUd15_B2.EU577371
B.US.-.HOBRd16_20.DQ444262
B.US.1991.SUMAd4_A32.EU579117
03_CH40TF
47_CH58TF
49_CH77TF
51_CH470TF
53_CH569TF
55_CH42TF
57_CH236TF
59_CH850TF
61_CH264TF
64_CH164TF
B.FR.1997.133-L-10.AY535442
B.FR.1999.153-L-7.AY535510
B.FR.1997.159-L-1.AY535477
B.FR.2000.309-L-7.AY535461
B.GB.2005.MM42d324_GN1.HM586204
B.NL.1995.H2_114_8F6.EU744054
B.NL.1996.H5_75_7G12.EU744175
B.NL.1996.H1_62_1A8.EU744010
B.NL.1998.H4_146_2H10.EU744145
B.NL.1997.H3_110_8G7.EU744096
B.US.-.BORI556_49.AY223734
B.US.1993.WEAU1166_39.AY223751
B.US.1991.HOBR0961_A21.GU331656
B.US.-.SUMA736_59.AY223781
46_CH40M6
48_CH58M6
50_CH77M6
52_CH470M6
54_CH569M6
56_CH42M6
58_CH236M6
60_CH850M6
62_CH264M6
63_CH164M6

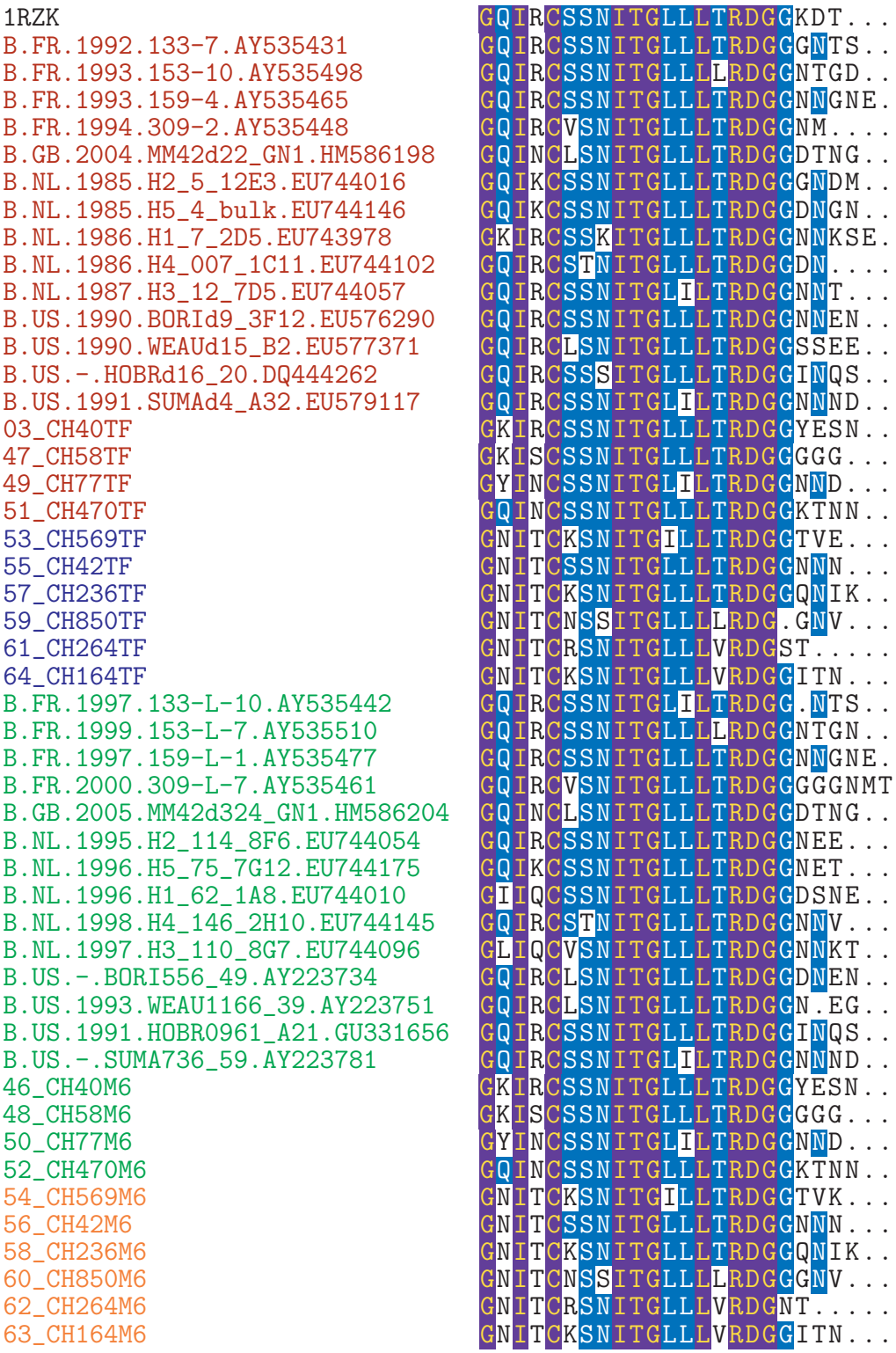

1RZK TQWENTLEQIAIKLKE..QFG...N
B.FR.1992.133-7.AY535431 TQWNKTLSLIVDKLKK..QFN...N
B.FR.1993.153-10.AY535498 GKWNNTLQRIVIKLKK..QFG...E
B.FR.1993.159-4.AY535465 AKWNDTLNQIVIKLRE..QFR...N
B.FR.1994.309-2.AY535448 ATWEETLKQIARKLRE..QFE...N
B.GB.2004.MM42d22_GN1.HM586198 TKWHDTLIQVSEKLKE..QF.....
B.NL.1985.H2_5_12E3.EU744016 TKWDNTLRQIVKKLRE..QFK...N
B.NL.1985.H5_4_bulk.EU744146 TKWENTLKQIVEKLRE..QFK...N
B.NL.1986.H1_7_2D5.EU743978 VDWEDTLKQIAEKLRE..QFR...N
B.NL.1986.H4_007_1C11.EU744102 AKWNDTLKQIVIKLRE..QFR...N
B.NL.1987.H3_12_7D5.EU744057 AKWNDTLKQIVIKLRK..QFR...N
B.US.1990.BORId9_3F12.EU576290 TKWEDTLKKIVTKLGE..QYG...N
B.US.1990.WEAUd15_B2.EU577371 TSWNNTLKQIVEKLREIKQFK...N
B.US.-.HOBRd16_20.DQ444262 TQWNNTLKQIVGKLRE..QFG...N
B.US.1991.SUMAd4_A32.EU579117 SKWNDTLQQIVKKLRE..QFK...N
03_CH40TF TEWHSTLKLVVEKLRE..QY.....
47_CH58TF EQWNNTLKKIVTKLRE..QF.....
49_CH77TF EDWNKTLSHVVDKLRE..QFR...N
51_CH470TF EKWNKTLRLIAEKLRE..QFREQLK
53_CH569TF GNWSKTLKKVKEKLEE..HFP....
55_CH42TF GDWDETLYNVSEKLKK..HFP....
57_CH236TF EAWNRTLLRVAKKLRE..YFP....
59_CH850TF NQWNTTLDQVGGKLKE..LFP....
61_CH264TF GDWNETLEQVKRKLGE..HFP....
64_CH164TF AEWNKTLQQVGRKLAE..HFP....
B.FR.1997.133-L-10.AY535442 TQWNNTLRLIAAKLKK..QFN...N
B.FR.1999.153-L-7.AY535510 EKWNNTLQRIVIKLRE..QFG...E
B.FR.1997.159-L-1.AY535477 AKWNDTLNQIVIKLRE..QFR...N
B.FR.2000.309-L-7.AY535461 ATWEETLKQIASKLRE..QFK...N
B.GB.2005.MM42d324_GN1.HM586204 TKWHDTLIQVSEKLKE..QF.....
B.NL.1995.H2_114_8F6.EU744054 KKWKNTLKQIVIKLRE..QFV...N
B.NL.1996.H5_75_7G12.EU744175 EKWGNTLKQIAEKLRE..QFK...N
B.NL.1996.H1_62_1A8.EU744010 ANWGLTLQRVAEKLKE..QFG...N
B.NL.1998.H4_146_2H10.EU744145 AQWNNTLKQIAIKLRE..QFG...N
B.NL.1997.H3_110_8G7.EU744096 TKWINTLKKIVIKLGE..QFG...N
B.US.-.BORI556_49.AY223734 TKWEDTLKKIVTKLGE..QYG...N
B.US.1993.WEAU1166_39.AY223751 TSWNNTLKQIVKKLREIEQFK...N
B.US.1991.HOBR0961_A21.GU331656 TQWNNTLKQIVGKLRE..QFG...N
B.US.-.SUMA736_59.AY223781 SKWNDTLKQIVKKLRE..QFK...N
46_CH40M6 TEWHSTLKLVVEKLRE..QY.....
48_CH58M6 EQWDNTLKKIVTKLRE..QF.....
50_CH77M6 EDWNKTLSHVVNKLRE..QFK...N
52_CH470M6 EKWNKTLRLIAEKLRE..QFREQLK
54_CH569M6 GNWSKTLKKVKEKLKE..HFP....
56_CH42M6 GDWDETLYNVSEKLKK..HFP....
58_CH236M6 EAWNRTLLRVAKKLRE..YFP....
60_CH850M6 NQWNTTLDQVGGKLKE..LFP....
62_CH264M6 GDWNETLEQVKRKLGE..HFP....
63_CH164M6 AEWDKTLQQVGRKLAE..HFP....

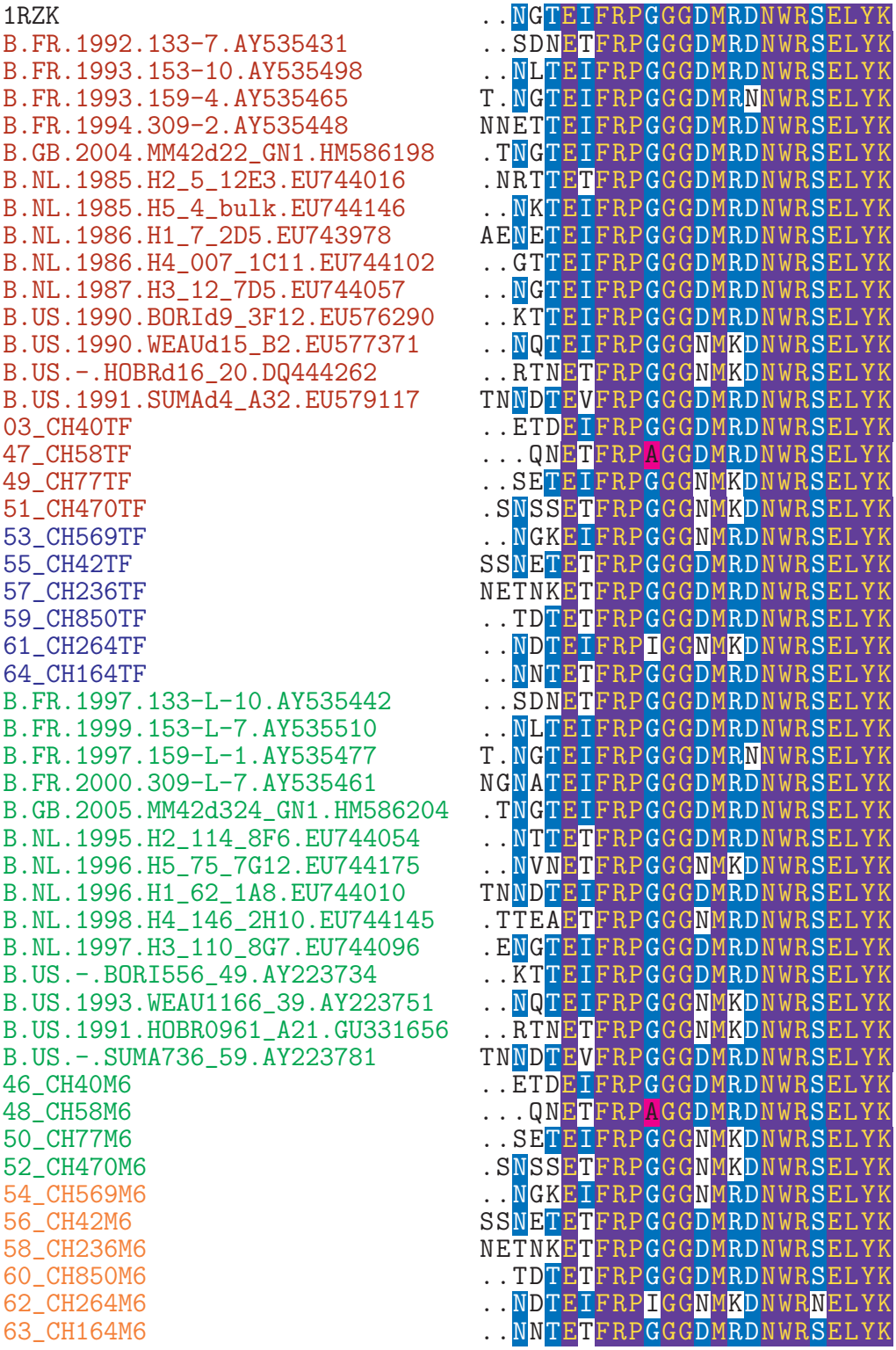

1RZK NKTIIFN..PSSGGDPEIVTHSFNC
B.FR.1992.133-7.AY535431 .KTIIFR..NSSGGDPEIVMHSFNC
B.FR.1993.153-10.AY535498 NKTIVFN..QSSEGDPEIVMHSFNC
B.FR.1993.159-4.AY535465 .KTIVFN..HSSGGDPEIVMHSFNC
B.FR.1994.309-2.AY535448 K.TIVFN..QSSGGDPEVVMHSVNC
B.GB.2004.MM42d22_GN1.HM586198 NKTIVFN..QSSGGDPEVVMHTFNC
B.NL.1985.H2_5_12E3.EU744016 .KTIVFN..QSSGGDPEIVTHSFNC
B.NL.1985.H5_4_bulk.EU744146 .KTIVFN..QSSGGDPEIVTHSFNC
B.NL.1986.H1_7_2D5.EU743978 .KTIVFN..QSSGGDPEITMHSFNC
B.NL.1986.H4_007_1C11.EU744102 .KTIVFN..QSSGGDPEIVMHSFNC
B.NL.1987.H3_12_7D5.EU744057 .RTIVFT..QSSGGDPEIVMHSFNC
B.US.1990.BORId9_3F12.EU576290 NKTIVFN..HSSGGDPEIVMHSFNC
B.US.1990.WEAUd15_B2.EU577371 .KTIVFK..QSSGGDPEIVMHSFNC
B.US.-.HOBRd16_20.DQ444262 NKTIVFN..QPSGGDPEIEMHSFNC
B.US.1991.SUMAd4_A32.EU579117 .KTIVFT..HSSGGDPEIVMHSFNC
03_CH40TF NKTIVFN..RSSGGDPEIVMYSFNC
47_CH58TF NKTIIFK..SPSGGDPEIVQHTFNC
49_CH77TF .KTIVFN..HSSGGDPEIVMHTFNC
51_CH470TF NKTIIFS..PHPGGDPEIEMHSFNC
53_CH569TF NKTITFN..QSSGGDLEITTHSFTC
55_CH42TF NKTIMFA..NSSGGDLEITTHSFNC
57_CH236TF NKTIAFD..SPSGGDLEIVTHTFNC
59_CH850TF NKTIQFK..PSSGGDLEITTHSFNC
61_CH264TF NKTIEFQPQPSSGGDPEITTHMFNC
64_CH164TF NTTIKFN..PSSGGDLEITTHSFNC
B.FR.1997.133-L-10.AY535442 .KTIIFR..NSSGGDPEIVMHSFNC
B.FR.1999.153-L-7.AY535510 NKTIVFN..QSSGGDLEIVMHSFNC
B.FR.1997.159-L-1.AY535477 .KTIVFN..HSSGGDPEIVMHSFNC
B.FR.2000.309-L-7.AY535461 KTTIVFN..QSSGGDPEVVMHSFNC
B.GB.2005.MM42d324_GN1.HM586204 NKTIVFN..QSSGGDPEVVMHTFNC
B.NL.1995.H2_114_8F6.EU744054 .KTIVFK..RSSGGDPEIVMHTFNC
B.NL.1996.H5_75_7G12.EU744175 .KTIAFN..QPSGGDPEIVMHSFNC
B.NL.1996.H1_62_1A8.EU744010 .KTIIFN..QSSGGDPEITMHIFNC
B.NL.1998.H4_146_2H10.EU744145 NKTIIFN..QSSGGDPEIEMHSFNC
B.NL.1997.H3_110_8G7.EU744096 .GTIVFN..HSSGGDPEIVMHSFNC
B.US.-.BORI556_49.AY223734 NKTIVFN..HSSGGDPEIVMHSFNC
B.US.1993.WEAU1166_39.AY223751 .KTIVFK..QSSGGDPEIVMHSFNC
B.US.1991.HOBR0961_A21.GU331656 NKTIVFN..QPSGGDPEIEMHSFNC
B.US.-.SUMA736_59.AY223781 .KTIVFN..HSSGGDPEIVMHSFNC
46_CH40M6 NKTIVFN..RSSGGDPEIVMYSFNC
48_CH58M6 NKTIIFK..SPSGGDPEIVQHTFNC
50_CH77M6 .KTIVFN..HSSGGDPEIVMHTFNC
52_CH470M6 NKTIIFS..PHPGGDPEIEMHSFNC
54_CH569M6 NKTITFN..QSSGGDLEITTHSFTC
56_CH42M6 NKTIMFA..NSSGGDLEITTHSFNC
58_CH236M6 NKTIAFD..SPSGGDLEIVTHTFNC
60_CH850M6 NKTIQFK..PSSGGDLEITTHSFNC
62_CH264M6 NKTIQFQPQPSSGGDPEITTHMFNC
63_CH164M6 NTTIKFN..PSSGGDLEITTHSFNC


1RZK GGEFFYCNSTQLFT..WNDTRK...
B.FR.1992.133-7.AY535431 GGEFFYCNTTQLFNSIWMLNNTWD.
B.FR.1993.153-10.AY535498 GGEFFYCDSTQLFNSTWNNS.TWN.
B.FR.1993.159-4.AY535465 GGEFFYCNTTQLFNSTWNANDIRNV
B.FR.1994.309-2.AY535448 GGEFFYCNTAQLFNSTWNTESLNK.
B.GB.2004.MM42d22_GN1.HM586198 GGEFFYCNTTQLFNSTWNKNTT...
B.NL.1985.H2_5_12E3.EU744016 GGEFFYCNSTPLFNSTWNSTQLFNS
B.NL.1985.H5_4_bulk.EU744146 GGEFFYCNSTQLFNSTWSDTE....
B.NL.1986.H1_7_2D5.EU743978 GGEFFYCNTTQLFNSTWNSTRNGT.
B.NL.1986.H4_007_1C11.EU744102 GGEFFYCNSTQLFNSTWMANSTWE.
B.NL.1987.H3_12_7D5.EU744057 GGEFFYCNSTQLFNSTWMLNSTWES
B.US.1990.BORId9_3F12.EU576290 GGEFFYCNSTQLFNSTWKFNNS.TW
B.US.1990.WEAUd15_B2.EU577371 GGEFFYCNSTQLFNSTWHANGTWKN
B.US.-.HOBRd16_20.DQ444262 GGEFFYCNTTQLFNSTWPFNSTWND
B.US.1991.SUMAd4_A32.EU579117 GREFFYCNTTQLFNSTWYINNTGNG
03_CH40TF GGEFFYCNSTKLFNSTWPWNDTKG.
47_CH58TF GGEFFYCDTKQLFNSTWNATKAN..
49_CH77TF GGEFFYCDSTALFNSTWRRNNTWTG
51_CH470TF GGEFFYCNTTRLFNN.WTSNNTWND
53_CH569TF RGEFFYCNTTELFNDTLLNA.....
55_CH42TF KGEFFYCNTTPLFNGTYNKTGAYNK
57_CH236TF GGEFFYCNTSDLFNRVYNTTGTYNS
59_CH850TF KGEFFYCNTSLLFNGTGNNS.....
61_CH264TF GGEFFYCNTSQLFNTTYNGTDANS.
64_CH164TF RGEFFYCNTEKLFNGTYNSTY.WPR
B.FR.1997.133-L-10.AY535442 GGEFFYCNTTQLFNSTWVHNNTWVH
B.FR.1999.153-L-7.AY535510 GGEFFYCNTTQLFNST......WN.
B.FR.1997.159-L-1.AY535477 GGEFFYCNTTQLFNSTWNANDIRNV
B.FR.2000.309-L-7.AY535461 GGEFFYCNTTQLFNSTWNTTSSNK.
B.GB.2005.MM42d324_GN1.HM586204 GGEFFYCNTTQLFNSTWNKNTP...
B.NL.1995.H2_114_8F6.EU744054 GGEFFYCDSTQLFNSIWP....LNS
B.NL.1996.H5_75_7G12.EU744175 GGEFFYCNTTKLFNSTWKENGT...
B.NL.1996.H1_62_1A8.EU744010 GGEFFYCNTTQLFKGIWNTNGTWNS
B.NL.1998.H4_146_2H10.EU744145 GGEFFYCNSTQLFDSTWKANSTWEN
B.NL.1997.H3_110_8G7.EU744096 GGEFFYCNTSQLFNSTWIFNGTWES
B.US.-.BORI556_49.AY223734 RGEFFYCNSTQLFNSTWNFNGTWNK
B.US.1993.WEAU1166_39.AY223751 GGEFFYCNSTQLFNSTWNATGTWND
B.US.1991.HOBR0961_A21.GU331656 GGEFFYCNTTQLFNSTWPFNSTWND
B.US.-.SUMA736_59.AY223781 GGEFFYCNTTQLFNSTWDINNTGNG
46_CH40M6 GGEFFYCNSTKLFNSTWPWNDTKG.
48_CH58M6 GGEFFYCDTKQLFNSTWNATKAN..
50_CH77M6 GGEFFYCDSTALFNSTWRRNNTWTG
52_CH470M6 GGEFFYCNTTRLFNN.WTSNNTWND
54_CH569M6 RGEFFYCNTTGLFNDTLLNA.....
56_CH42M6 KGEFFYCNTTPLFNGTYNKTGAYNK
58_CH236M6 GGEFFYCNTSDLFNRVYNTTGTYNS
60_CH850M6 KGEFFYCNTSLLFNGTGNNS.....
62_CH264M6 GGEFFYCNTSQLFNTTYNGTDANS.
63_CH164M6 RGEFFYCNTEKLFNGTYNSTY.WPR


1RZK
B.FR.1992.133-7.AY535431
B.FR.1993.153-10.AY535498
B.FR.1993.159-4.AY535465
B.FR.1994.309-2.AY535448
B.GB.2004.MM42d22_GN1.HM586198
B.NL.1985.H2_5_12E3.EU744016
B.NL.1985.H5_4_bulk.EU744146
B.NL.1986.H1_7_2D5.EU743978
B.NL.1986.H4_007_1C11.EU744102
B.NL.1987.H3_12_7D5.EU744057
B.US.1990.BORId9_3F12.EU576290
B.US.1990.WEAUd15_B2.EU577371
B.US.-.HOBRd16_20.DQ444262
B.US.1991.SUMAd4_A32.EU579117
03_CH40TF
47_CH58TF
49_CH77TF
51_CH470TF
53_CH569TF
55_CH42TF
57_CH236TF
59_CH850TF
61_CH264TF
64_CH164TF
B.FR.1997.133-L-10.AY535442
B.FR.1999.153-L-7.AY535510
B.FR.1997.159-L-1.AY535477
B.FR.2000.309-L-7.AY535461
B.GB.2005.MM42d324_GN1.HM586204
B.NL.1995.H2_114_8F6.EU744054
B.NL.1996.H5_75_7G12.EU744175
B.NL.1996.H1_62_1A8.EU744010
B.NL.1998.H4_146_2H10.EU744145
B.NL.1997.H3_110_8G7.EU744096
B.US.-.BORI556_49.AY223734
B.US.1993.WEAU1166_39.AY223751
B.US.1991.HOBR0961_A21.GU331656
B.US.-.SUMA736_59.AY223781
46_CH40M6
48_CH58M6
50_CH77M6
52_CH470M6
54_CH569M6
56_CH42M6
58_CH236M6
60_CH850M6
62_CH264M6
63_CH164M6


1RZK
B.FR.1992.133-7.AY535431
B.FR.1993.153-10.AY535498
B.FR.1993.159-4.AY535465
B.FR.1994.309-2.AY535448
B.GB.2004.MM42d22_GN1.HM586198
B.NL.1985.H2_5_12E3.EU744016
B.NL.1985.H5_4_bulk.EU744146
B.NL.1986.H1_7_2D5.EU743978
B.NL.1986.H4_007_1C11.EU744102
B.NL.1987.H3_12_7D5.EU744057
B.US.1990.BORId9_3F12.EU576290
B.US.1990.WEAUd15_B2.EU577371
B.US.-.HOBRd16_20.DQ444262
B.US.1991.SUMAd4_A32.EU579117
03_CH40TF
47_CH58TF
49_CH77TF
51_CH470TF
53_CH569TF
55_CH42TF
57_CH236TF
59_CH850TF
61_CH264TF
64_CH164TF
B.FR.1997.133-L-10.AY535442
B.FR.1999.153-L-7.AY535510
B.FR.1997.159-L-1.AY535477
B.FR.2000.309-L-7.AY535461
B.GB.2005.MM42d324_GN1.HM586204
B.NL.1995.H2_114_8F6.EU744054
B.NL.1996.H5_75_7G12.EU744175
B.NL.1996.H1_62_1A8.EU744010
B.NL.1998.H4_146_2H10.EU744145
B.NL.1997.H3_110_8G7.EU744096
B.US.-.BORI556_49.AY223734
B.US.1993.WEAU1166_39.AY223751
B.US.1991.HOBR0961_A21.GU331656
B.US.-.SUMA736_59.AY223781
46_CH40M6
48_CH58M6
50_CH77M6
52_CH470M6
54_CH569M6
56_CH42M6
58_CH236M6
60_CH850M6
62_CH264M6
63_CH164M6


1RZK GQIRCSSNITGLLLTRDGGKDT...
B.FR.1992.133-7.AY535431 GQIRCSSNITGLLLTRDGGGNTS..
B.FR.1993.153-10.AY535498 GQIRCSSNITGLLLLRDGGNTGD..
B.FR.1993.159-4.AY535465 GQIRCSSNITGLLLTRDGGNNGNE.
B.FR.1994.309-2.AY535448 GQIRCVSNITGLLLTRDGGNM....
B.GB.2004.MM42d22_GN1.HM586198 GQINCLSNITGLLLTRDGGDTNG..
B.NL.1985.H2_5_12E3.EU744016 GQIKCSSNITGLLLTRDGGGNDM..
B.NL.1985.H5_4_bulk.EU744146 GQIKCSSNITGLLLTRDGGDNGN..
B.NL.1986.H1_7_2D5.EU743978 GKIRCSSKITGLLLTRDGGNNKSE.
B.NL.1986.H4_007_1C11.EU744102 GQIRCSTNITGLLLTRDGGDN....
B.NL.1987.H3_12_7D5.EU744057 GQIRCSSNITGLILTRDGGNNT...
B.US.1990.BORId9_3F12.EU576290 GQIRCSSNITGLLLTRDGGNNEN..
B.US.1990.WEAUd15_B2.EU577371 GQIRCLSNITGLLLTRDGGSSEE..
B.US.-.HOBRd16_20.DQ444262 GQIRCSSSITGLLLTRDGGINQS..
B.US.1991.SUMAd4_A32.EU579117 GQIRCSSNITGLILTRDGGNNND..
03_CH40TF GKIRCSSNITGLLLTRDGGYESN..
47_CH58TF GKISCSSNITGLLLTRDGGGGG...
49_CH77TF GYINCSSNITGLILTRDGGNND...
51_CH470TF GQINCSSNITGLLLTRDGGKTNN..
53_CH569TF GNITCKSNITGILLTRDGGTVE...
55_CH42TF GNITCSSNITGLLLTRDGGNNN...
57_CH236TF GNITCKSNITGLLLTRDGGQNIK..
59_CH850TF GNITCNSSITGLLLLRDG.GNV...
61_CH264TF GNITCRSNITGLLLVRDGST.....
64_CH164TF GNITCKSNITGLLLVRDGGITN...
B.FR.1997.133-L-10.AY535442 GQIRCSSNITGLILTRDGG.NTS..
B.FR.1999.153-L-7.AY535510 GQIRCSSNITGLLLLRDGGNTGN..
B.FR.1997.159-L-1.AY535477 GQIRCSSNITGLLLTRDGGNNGNE.
B.FR.2000.309-L-7.AY535461 GQIRCVSNITGLLLTRDGGGGGNMT
B.GB.2005.MM42d324_GN1.HM586204 GQINCLSNITGLLLTRDGGDTNG..
B.NL.1995.H2_114_8F6.EU744054 GQIRCSSNITGLLLTRDGGNEE...
B.NL.1996.H5_75_7G12.EU744175 GQIKCSSNITGLLLTRDGGNET...
B.NL.1996.H1_62_1A8.EU744010 GIIQCSSNITGLLLTRDGGDSNE..
B.NL.1998.H4_146_2H10.EU744145 GQIRCSTNITGLLLTRDGGNNV...
B.NL.1997.H3_110_8G7.EU744096 GLIQCVSNITGLLLTRDGGNNKT..
B.US.-.BORI556_49.AY223734 GQIRCLSNITGLLLTRDGGDNEN..
B.US.1993.WEAU1166_39.AY223751 GQIRCLSNITGLLLTRDGGN.EG..
B.US.1991.HOBR0961_A21.GU331656 GQIRCSSNITGLLLTRDGGINQS..
B.US.-.SUMA736_59.AY223781 GQIRCSSNITGLILTRDGGNNND..
46_CH40M6 GKIRCSSNITGLLLTRDGGYESN..
48_CH58M6 GKISCSSNITGLLLTRDGGGGG...
50_CH77M6 GYINCSSNITGLILTRDGGNND...
52_CH470M6 GQINCSSNITGLLLTRDGGKTNN..
54_CH569M6 GNITCKSNITGILLTRDGGTVK...
56_CH42M6 GNITCSSNITGLLLTRDGGNNN...
58_CH236M6 GNITCKSNITGLLLTRDGGQNIK..
60_CH850M6 GNITCNSSITGLLLLRDGGGNV...
62_CH264M6 GNITCRSNITGLLLVRDGNT.....
63_CH164M6 GNITCKSNITGLLLVRDGGITN...

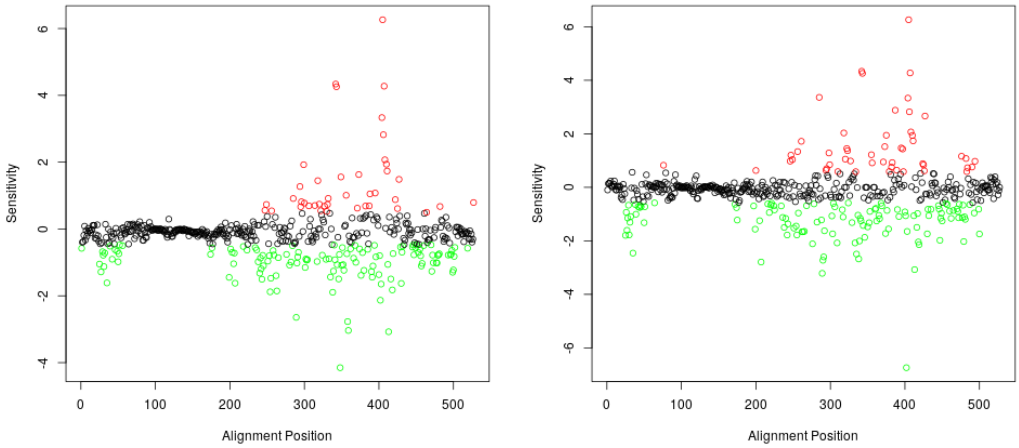

1RZK ..NGTEIFRPGGGDMRDNWRSELYK
B.FR.1992.133-7.AY535431 ..SDNETFRPGGGDMRDNWRSELYK
B.FR.1993.153-10.AY535498 ..NLTEIFRPGGGDMRDNWRSELYK
B.FR.1993.159-4.AY535465 T.NGTEIFRPGGGDMRNNWRSELYK
B.FR.1994.309-2.AY535448 NNETTEIFRPGGGDMRDNWRSELYK
B.GB.2004.MM42d22_GN1.HM586198 .TNGTEIFRPGGGDMRDNWRSELYK
B.NL.1985.H2_5_12E3.EU744016 .NRTTETFRPGGGDMRDNWRSELYK
B.NL.1985.H5_4_bulk.EU744146 ..NKTEIFRPGGGDMRDNWRSELYK
B.NL.1986.H1_7_2D5.EU743978 AENETEIFRPGGGDMRDNWRSELYK
B.NL.1986.H4_007_1C11.EU744102 ..GTTEIFRPGGGDMRDNWRSELYK
B.NL.1987.H3_12_7D5.EU744057 ..NGTEIFRPGGGDMRDNWRSELYK
B.US.1990.BORId9_3F12.EU576290 ..KTTEIFRPGGGDMRDNWRSELYK
B.US.1990.WEAUd15_B2.EU577371 ..NQTEIFRPGGGNMKDNWRSELYK
B.US.-.HOBRd16_20.DQ444262 ..RTNETFRPGGGNMKDNWRSELYK
B.US.1991.SUMAd4_A32.EU579117 TNNDTEVFRPGGGDMRDNWRSELYK
03_CH40TF ..ETDEIFRPGGGDMRDNWRSELYK
47_CH58TF ...QNETFRPAGGDMRDNWRSELYK
49_CH77TF ..SETEIFRPGGGNMKDNWRSELYK
51_CH470TF .SNSSETFRPGGGNMKDNWRSELYK
53_CH569TF ..NGKEIFRPGGGNMRDNWRSELYK
55_CH42TF SSNETETFRPGGGDMRDNWRSELYK
57_CH236TF NETNKETFRPGGGDMRDNWRSELYK
59_CH850TF ..TDTETFRPGGGDMRDNWRSELYK
61_CH264TF ..NDTEIFRPIGGNMKDNWRSELYK
64_CH164TF ..NNTETFRPGGGDMRDNWRSELYK
B.FR.1997.133-L-10.AY535442 ..SDNETFRPGGGDMRDNWRSELYK
B.FR.1999.153-L-7.AY535510 ..NLTEIFRPGGGDMRDNWRSELYK
B.FR.1997.159-L-1.AY535477 T.NGTEIFRPGGGDMRNNWRSELYK
B.FR.2000.309-L-7.AY535461 NGNATEIFRPGGGDMRDNWRSELYK
B.GB.2005.MM42d324_GN1.HM586204 .TNGTEIFRPGGGDMRDNWRSELYK
B.NL.1995.H2_114_8F6.EU744054 ..NTTETFRPGGGDMRDNWRSELYK
B.NL.1996.H5_75_7G12.EU744175 ..NVNETFRPGGGNMKDNWRSELYK
B.NL.1996.H1_62_1A8.EU744010 TNNDTEIFRPGGGDMRDNWRSELYK
B.NL.1998.H4_146_2H10.EU744145 .TTEAETFRPGGGNMRDNWRSELYK
B.NL.1997.H3_110_8G7.EU744096 .ENGTEIFRPGGGDMRDNWRSELYK
B.US.-.BORI556_49.AY223734 ..KTTEIFRPGGGDMRDNWRSELYK
B.US.1993.WEAU1166_39.AY223751 ..NQTEIFRPGGGNMKDNWRSELYK
B.US.1991.HOBR0961_A21.GU331656 ..RTNETFRPGGGNMKDNWRSELYK
B.US.-.SUMA736_59.AY223781 TNNDTEVFRPGGGDMRDNWRSELYK
46_CH40M6 ..ETDEIFRPGGGDMRDNWRSELYK
48_CH58M6 ...QNETFRPAGGDMRDNWRSELYK
50_CH77M6 ..SETEIFRPGGGNMKDNWRSELYK
52_CH470M6 .SNSSETFRPGGGNMKDNWRSELYK
54_CH569M6 ..NGKEIFRPGGGNMRDNWRSELYK
56_CH42M6 SSNETETFRPGGGDMRDNWRSELYK
58_CH236M6 NETNKETFRPGGGDMRDNWRSELYK
60_CH850M6 ..TDTETFRPGGGDMRDNWRSELYK
62_CH264M6 ..NDTEIFRPIGGNMKDNWRNELYK
63_CH164M6 ..NNTETFRPGGGDMRDNWRSELYK


1RZK
B.FR.1992.133-7.AY535431
B.FR.1993.153-10.AY535498
B.FR.1993.159-4.AY535465
B.FR.1994.309-2.AY535448
B.GB.2004.MM42d22_GN1.HM586198
B.NL.1985.H2_5_12E3.EU744016
B.NL.1985.H5_4_bulk.EU744146
B.NL.1986.H1_7_2D5.EU743978
B.NL.1986.H4_007_1C11.EU744102
B.NL.1987.H3_12_7D5.EU744057
B.US.1990.BORId9_3F12.EU576290
B.US.1990.WEAUd15_B2.EU577371
B.US.-.HOBRd16_20.DQ444262
B.US.1991.SUMAd4_A32.EU579117
03_CH40TF
47_CH58TF
49_CH77TF
51_CH470TF
53_CH569TF
55_CH42TF
57_CH236TF
59_CH850TF
61_CH264TF
64_CH164TF
B.FR.1997.133-L-10.AY535442
B.FR.1999.153-L-7.AY535510
B.FR.1997.159-L-1.AY535477
B.FR.2000.309-L-7.AY535461
B.GB.2005.MM42d324_GN1.HM586204
B.NL.1995.H2_114_8F6.EU744054
B.NL.1996.H5_75_7G12.EU744175
B.NL.1996.H1_62_1A8.EU744010
B.NL.1998.H4_146_2H10.EU744145
B.NL.1997.H3_110_8G7.EU744096
B.US.-.BORI556_49.AY223734
B.US.1993.WEAU1166_39.AY223751
B.US.1991.HOBR0961_A21.GU331656
B.US.-.SUMA736_59.AY223781
46_CH40M6
48_CH58M6
50_CH77M6
52_CH470M6
54_CH569M6
56_CH42M6
58_CH236M6
60_CH850M6
62_CH264M6
63_CH164M6

```
1RZK                                  ..
B.FR.1992.133-7.AY535431              R.
B.FR.1993.153-10.AY535498             R.
B.FR.1993.159-4.AY535465              R.
B.FR.1994.309-2.AY535448              R.
B.GB.2004.MM42d22_GN1.HM586198        R.
B.NL.1985.H2_5_12E3.EU744016          R.
B.NL.1985.H5_4_bulk.EU744146          R.
B.NL.1986.H1_7_2D5.EU743978           R.
B.NL.1986.H4_007_1C11.EU744102        R.
B.NL.1987.H3_12_7D5.EU744057          R.
B.US.1990.BORId9_3F12.EU576290        R.
B.US.1990.WEAUd15_B2.EU577371         R.
B.US.-.HOBRd16_20.DQ444262            R.
B.US.1991.SUMAd4_A32.EU579117         R.
03_CH40TF                             R.
47_CH58TF                             R.
49_CH77TF                             R.
51_CH470TF                            R.
53_CH569TF                            R.
55_CH42TF                             R.
57_CH236TF                            R.
59_CH850TF                            R.
61_CH264TF                            KR
64_CH164TF                            R.
B.FR.1997.133-L-10.AY535442           R.
B.FR.1999.153-L-7.AY535510            R.
B.FR.1997.159-L-1.AY535477            R.
B.FR.2000.309-L-7.AY535461            R.
B.GB.2005.MM42d324_GN1.HM586204       R.
B.NL.1995.H2_114_8F6.EU744054         R.
B.NL.1996.H5_75_7G12.EU744175         R.
B.NL.1996.H1_62_1A8.EU744010          R.
B.NL.1998.H4_146_2H10.EU744145        R.
B.NL.1997.H3_110_8G7.EU744096         R.
B.US.-.BORI556_49.AY223734            R.
B.US.1993.WEAU1166_39.AY223751        R.
B.US.1991.HOBR0961_A21.GU331656       R.
B.US.-.SUMA736_59.AY223781            R.
46_CH40M6                             R.
48_CH58M6                             R.
50_CH77M6                             R.
52_CH470M6                            R.
54_CH569M6                            R.
56_CH42M6                             R.
58_CH236M6                            R.
60_CH850M6                            R.
62_CH264M6                            KR
63_CH164M6                            R.
```

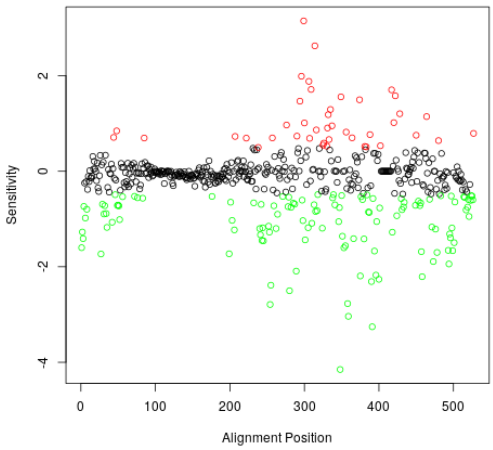
APPENDIX C

Additional Residue Specific Sensitivity



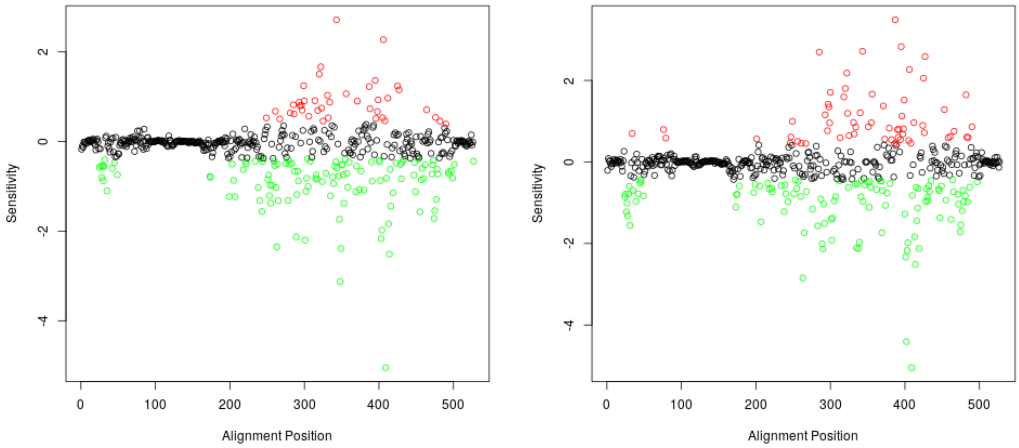
(A) Overall

(B) Within B



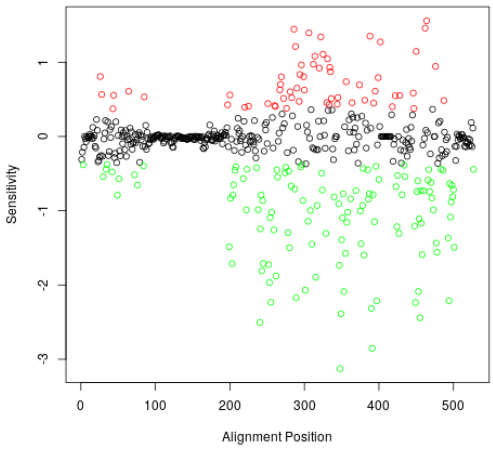
(C) Within C

Figure C.1: TF vs CC Relative Residue Specific pH Sensitivity Using pH 4 and 7. Median charges of residues at each position in the alignment were computed. Sensitivity was calculated as the charge at pH 4 subtracted from the charge at pH 7. CC charges were subtracted from TF charges. Values above zero indicate greater sensitivity in TF strains, while values below zero indicate greater sensitivity in CC strains. Red points are greater than one interquartile range above zero, and green points are below one interquartile range below zero.



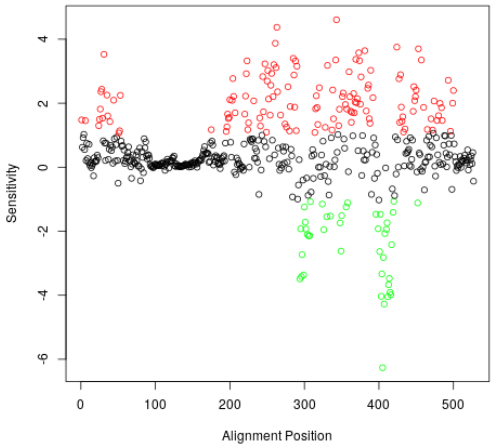
(A) Overall

(B) Within B

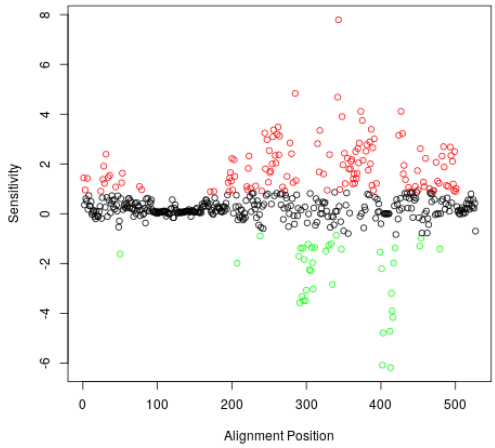


(C) Within C

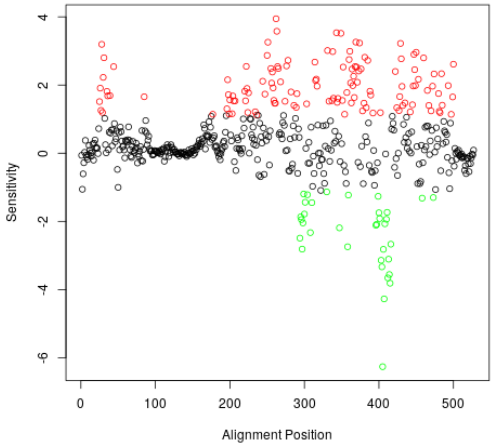
Figure C.2: TF vs CC Relative Residue Specific pH Sensitivity Using pH 5 and 8. Median charges of residues at each position in the alignment were computed. Sensitivity was calculated as the charge at pH 5 subtracted from the charge at pH 8. CC charges were subtracted from TF charges. Values above zero indicate greater sensitivity in TF strains, while values below zero indicate greater sensitivity in CC strains. Red points are greater than one interquartile range above zero, and green points are below one interquartile range below zero.



(A) Overall

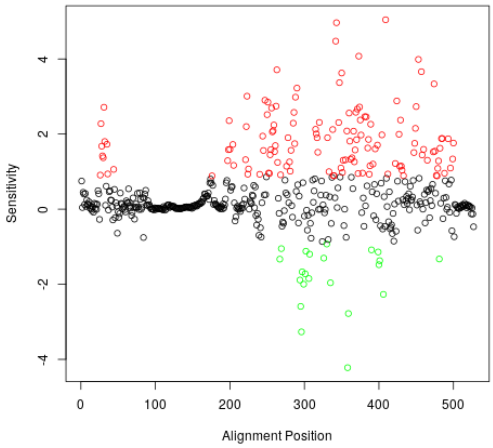


(B) Within TF

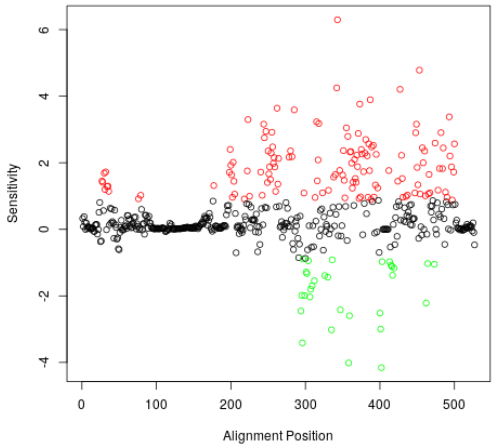


(C) Within CC

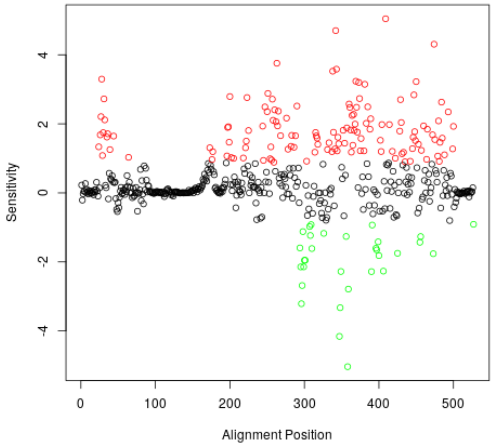
Figure C.3: B vs C Relative Residue Specific pH Sensitivity Using pH 4 and 7. Median charges of residues at each position in the alignment were computed. Sensitivity was calculated as the charge at pH 4 subtracted from the charge at pH 7. B charges were subtracted from B charges. Values above zero indicate greater sensitivity in B strains, while values below zero indicate greater sensitivity in C strains. Red points are greater than one interquartile range above zero, and green points are below one interquartile range below zero.



(A) Overall



(B) Within TF

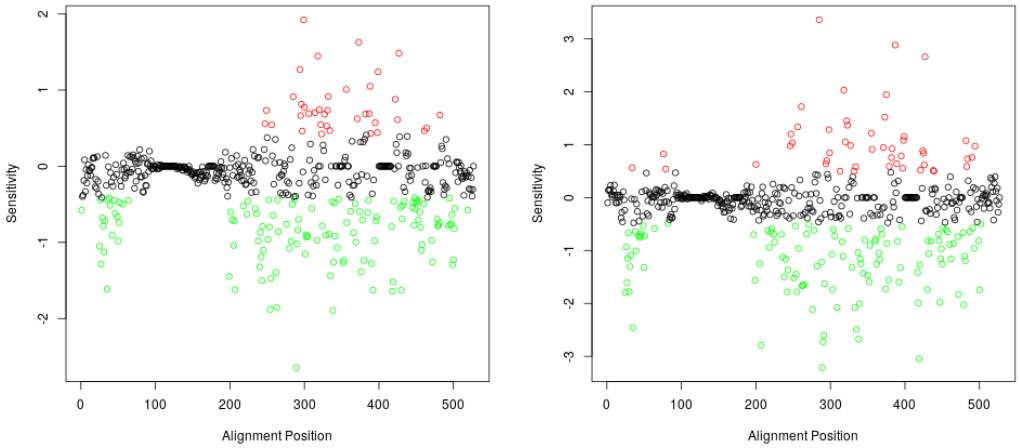


(C) Within CC

Figure C.4: B vs C Relative Residue Specific pH Sensitivity Using pH 5 and 8. Median charges of residues at each position in the alignment were computed. Sensitivity was calculated as the charge at pH 5 subtracted from the charge at pH 8. B charges were subtracted from B charges. Values above zero indicate greater sensitivity in B strains, while values below zero indicate greater sensitivity in C strains. Red points are greater than one interquartile range above zero, and green points are below one interquartile range below zero.

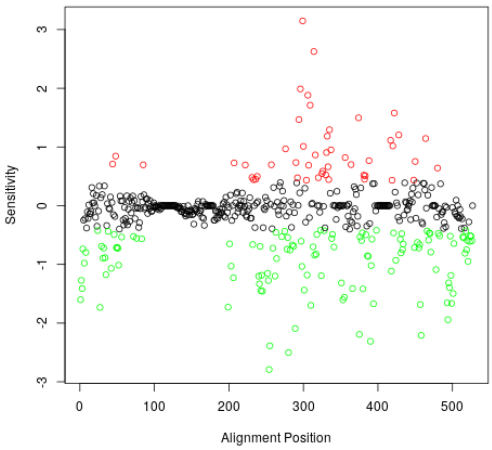
APPENDIX D

Additional Residue Specific Sensitivity Considering Gaps



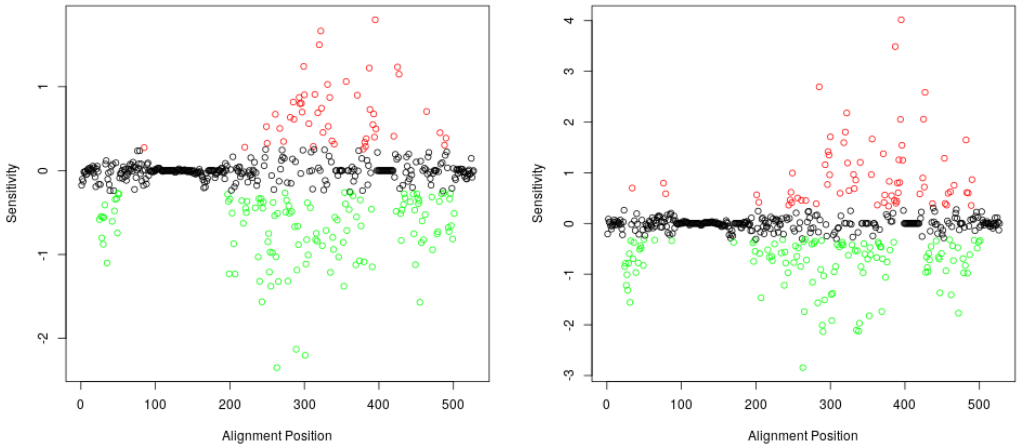
(A) Overall

(B) Within B



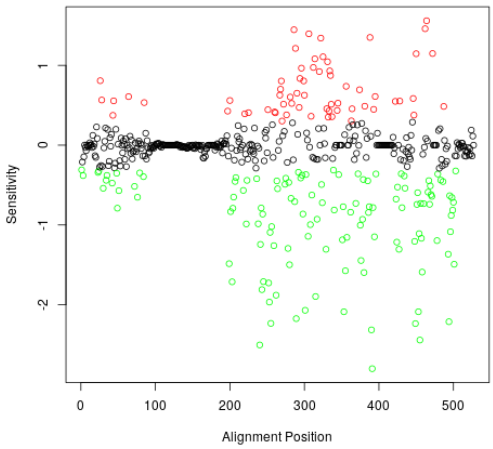
(C) Within C

Figure D.1: TF vs CC Relative Residue Specific pH Sensitivity Using pH 4 and 7 Considering Gaps. Median charges of residues at each position in the alignment were computed with gaps being considered a charge value of zero. Sensitivity was calculated as the charge at pH 4 subtracted from the charge at pH 7. CC charges were subtracted from TF charges. Values above zero indicate greater sensitivity in TF strains, while values below zero indicate greater sensitivity in CC strains. Red points are greater than one interquartile range above zero, and green points are below one interquartile range below zero.



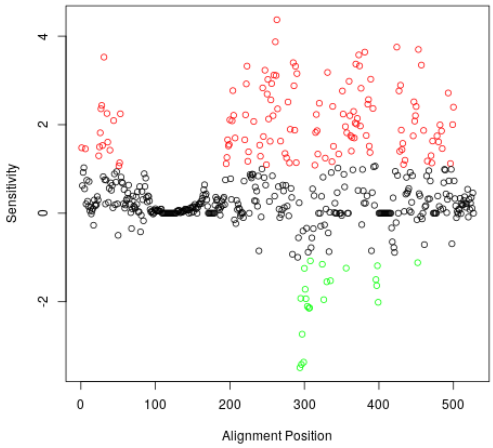
(A) Overall

(B) Within B

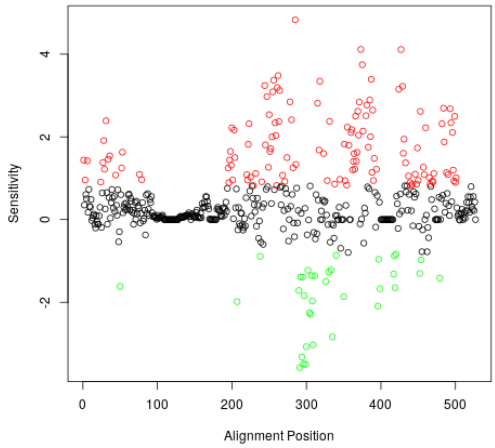


(C) Within C

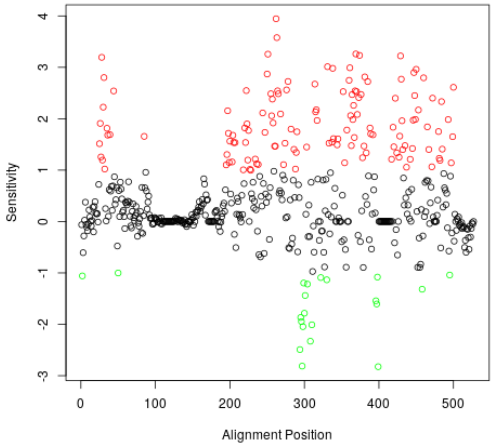
Figure D.2: TF vs CC Relative Residue Specific pH Sensitivity Using pH 5 and 8 Considering Gaps. Median charges of residues at each position in the alignment were computed with gaps being considered a charge value of zero. Sensitivity was calculated as the charge at pH 5 subtracted from the charge at pH 8. CC charges were subtracted from TF charges. Values above zero indicate greater sensitivity in TF strains, while values below zero indicate greater sensitivity in CC strains. Red points are greater than one interquartile range above zero, and green points are below one interquartile range below zero.



(A) Overall

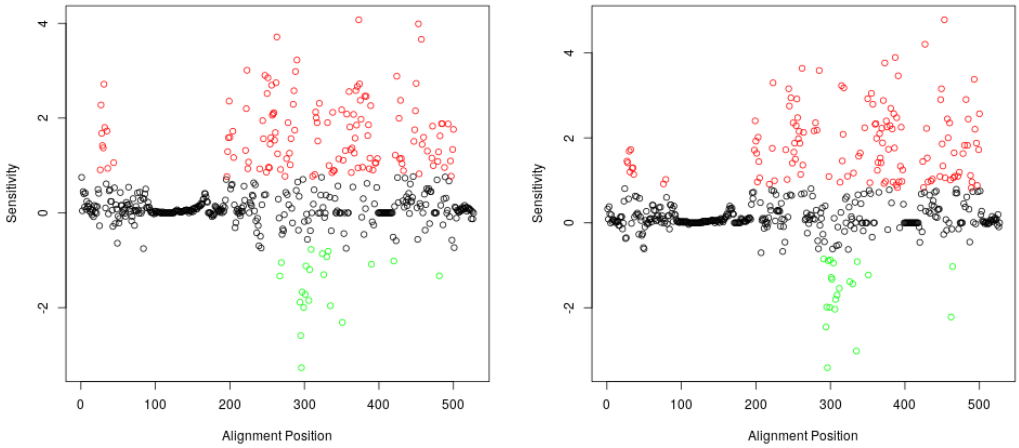


(B) Within TF



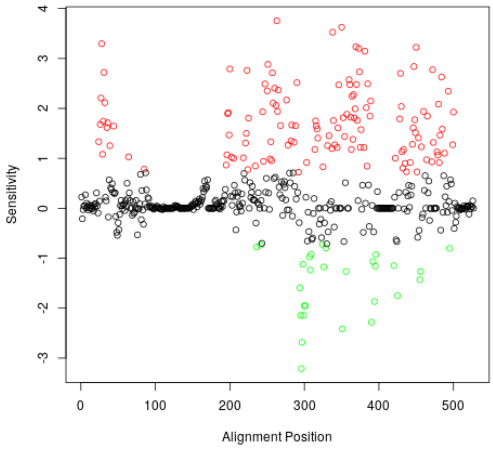
(C) Within CC

Figure D.3: B vs C Relative Residue Specific pH Sensitivity Using pH 4 and 7 Considering Gaps. Median charges of residues at each position in the alignment were computed with gaps being considered a charge value of zero. Sensitivity was calculated as the charge at pH 4 subtracted from the charge at pH 7. B charges were subtracted from B charges. Values above zero indicate greater sensitivity in B strains, while values below zero indicate greater sensitivity in C strains. Red points are greater than one interquartile range above zero, and green points are below one interquartile range below zero.



(A) Overall

(B) Within TF

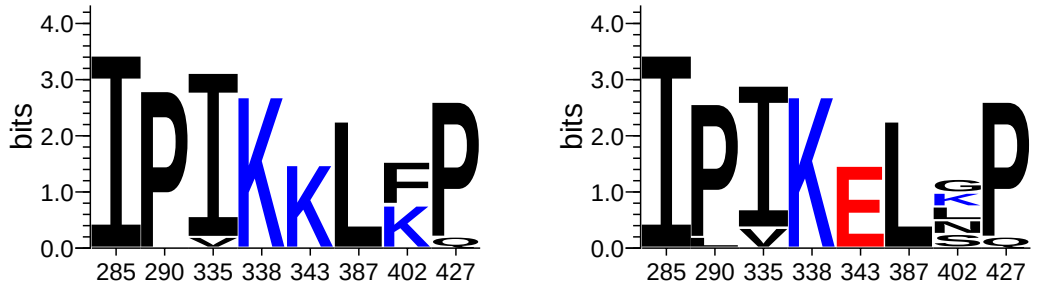


(C) Within CC

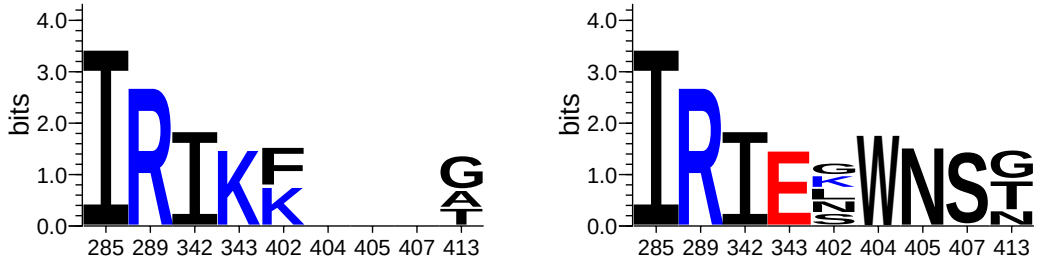
Figure D.4: B vs C Relative Residue Specific pH Sensitivity Using pH 5 and 8 Considering Gaps. Median charges of residues at each position in the alignment were computed with gaps being considered a charge value of zero. Sensitivity was calculated as the charge at pH 5 subtracted from the charge at pH 8. B charges were subtracted from B charges. Values above zero indicate greater sensitivity in B strains, while values below zero indicate greater sensitivity in C strains. Red points are greater than one interquartile range above zero, and green points are below one interquartile range below zero.

APPENDIX E

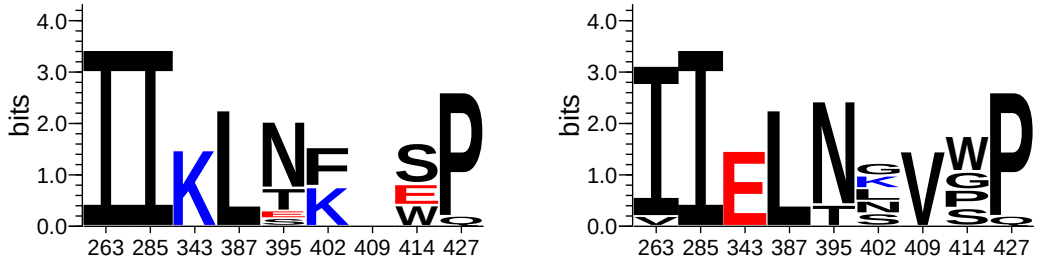
Sequence Logos Within Groups



(A) TF vs CC - Average pH

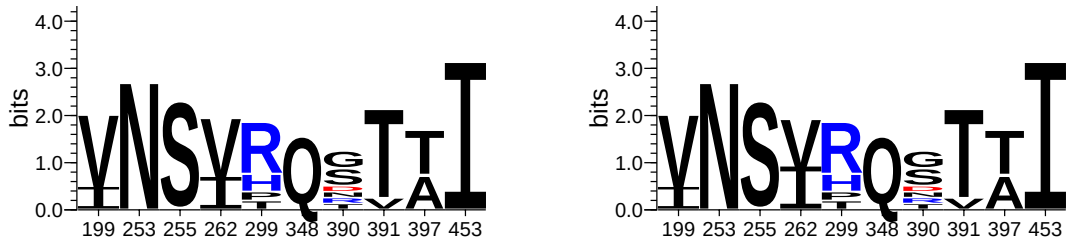


(B) TF vs CC - pH 4 & 7

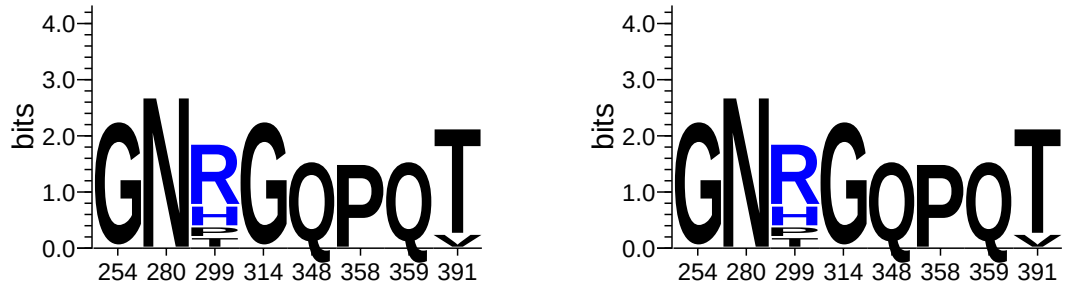


(C) TF vs CC - pH 5 & 8

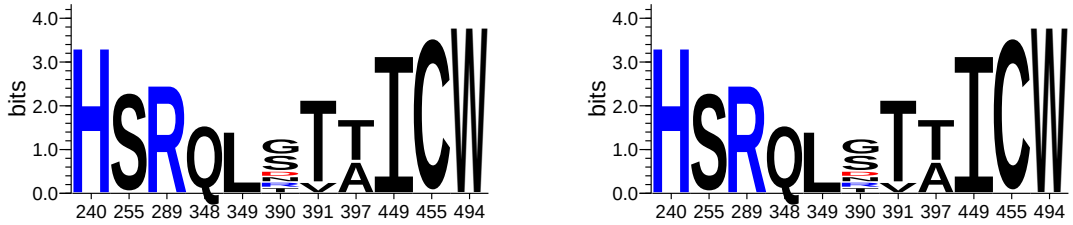
Figure E.1: TF vs CC Sensitive Residue Composition Within B Clade. A) Comparison of the top 1% of residues identified identified from Figure 27B. B) Comparison of the top 1% of residues identified in Figure C.1B. C) Comparison of the top 1% of residues identified in Figure C.2B.



(A) TF vs CC - Average pH

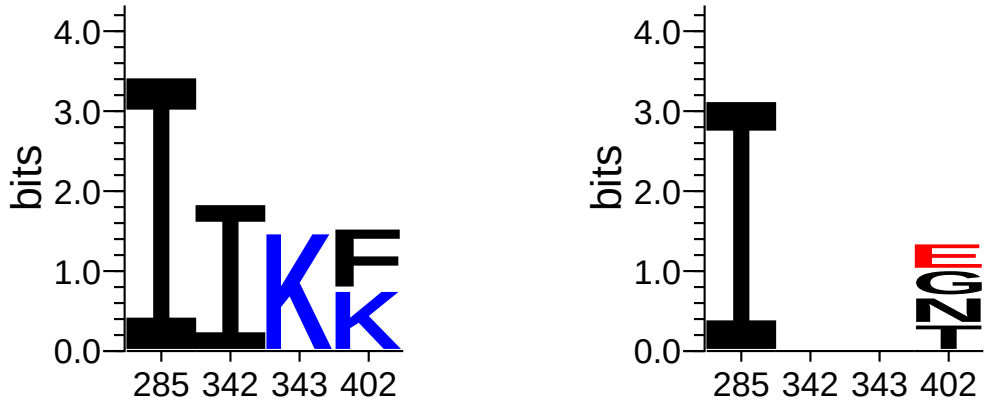


(B) TF vs CC - pH 4 & 7

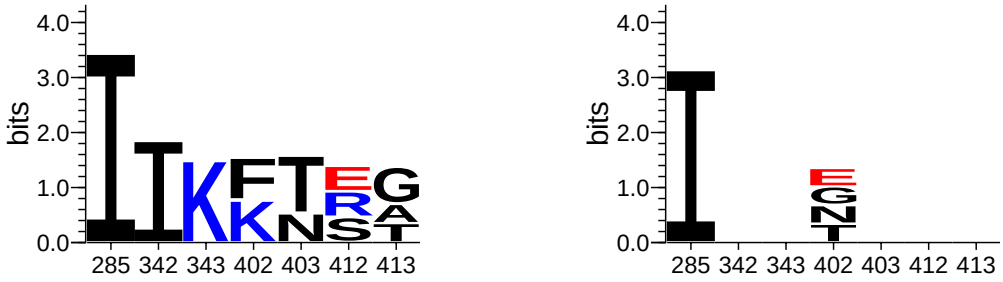


(C) TF vs CC - pH 5 & 8

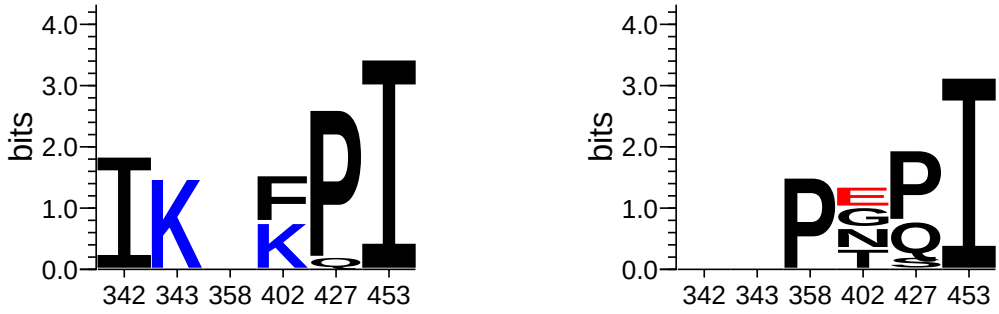
Figure E.2: TF vs CC Sensitive Residue Composition Within C Clade. A) Comparison of the top 1% of residues identified identified from Figure 27C. B) Comparison of the top 1% of residues identified in Figure C.1C. C) Comparison of the top 1% of residues identified in Figure C.2C.



(A) B vs C - Average pH



(B) B vs C - pH 4 & 7



(C) B vs C - pH 5 & 8

Figure E.3: B vs C Sensitive Residue Composition Within the TF Class. A) Comparison of the top 1% of residues identified identified from Figure 28B. B) Comparison of the top 1% of residues identified in Figure C.3B. C) Comparison of the top 1% of residues identified in Figure C.4B.

(A) B vs C - Average pH

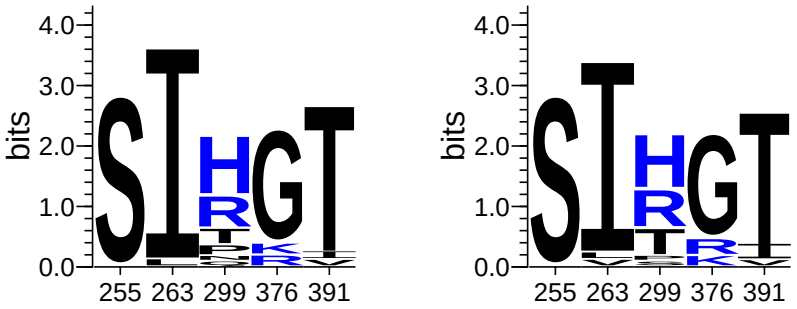
(B) B vs C - pH 4 & 7

(C) B vs C - pH 5 & 8

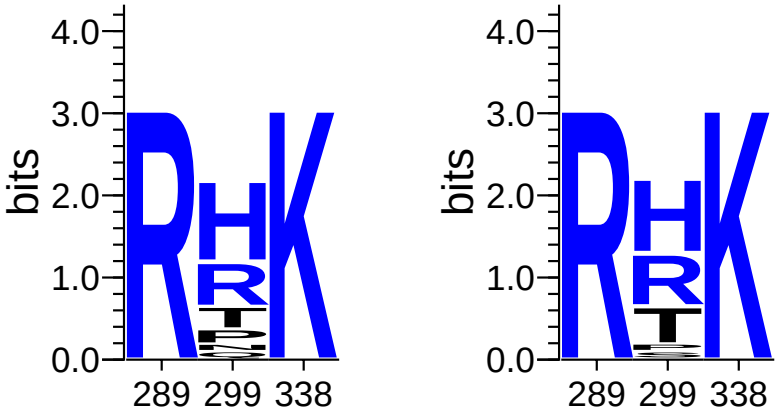
Figure E.4: B vs C Sensitive Residue Composition Within the CC Class. A) Comparison of the top 1% of residues identified identified from Figure 28C. B) Comparison of the top 1% of residues identified in Figure C.3C. C) Comparison of the top 1% of residues identified in Figure C.4C.

APPENDIX F

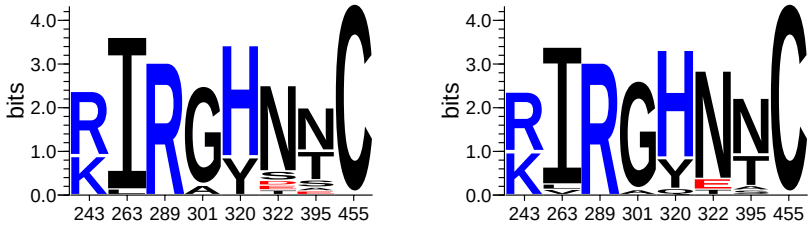
Sequence Logos Considering Gaps



(A) TF vs CC - Average pH



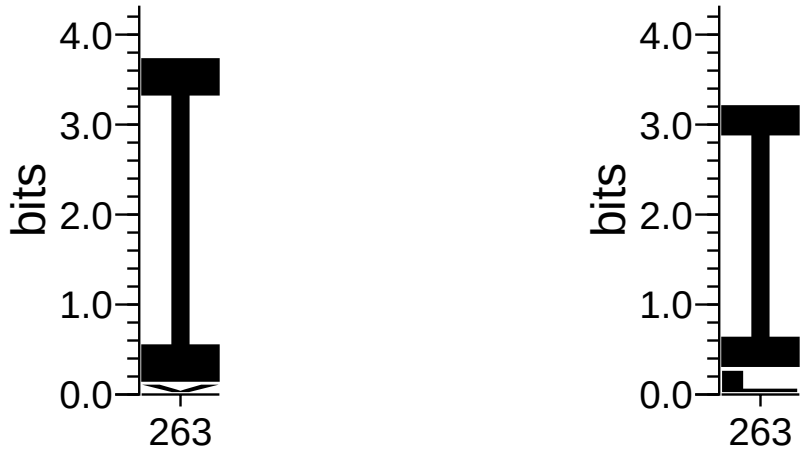
(B) TF vs CC - pH 4 & 7



(C) TF vs CC - pH 5 & 8

Figure F.1: TF vs CC Sensitive Residue Composition Considering Gaps. A) Comparison of the top 1% of residues identified identified from Figure 29A. B) Comparison of the top 1% of residues identified in Figure D.1A. C) Comparison of the top 1% of residues identified in Figure D.2A.

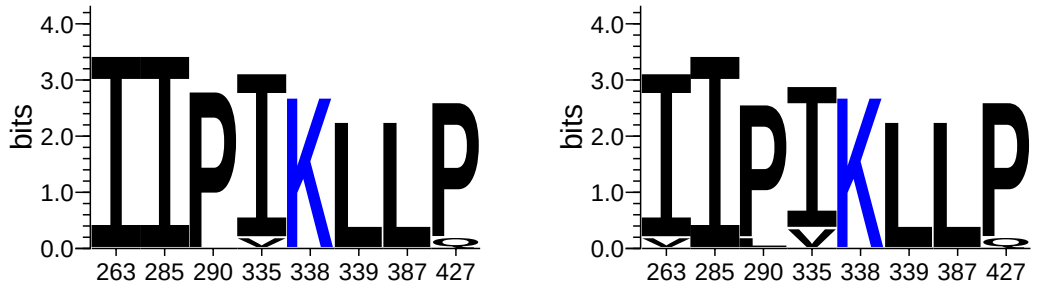
(A) B vs C - Average pH



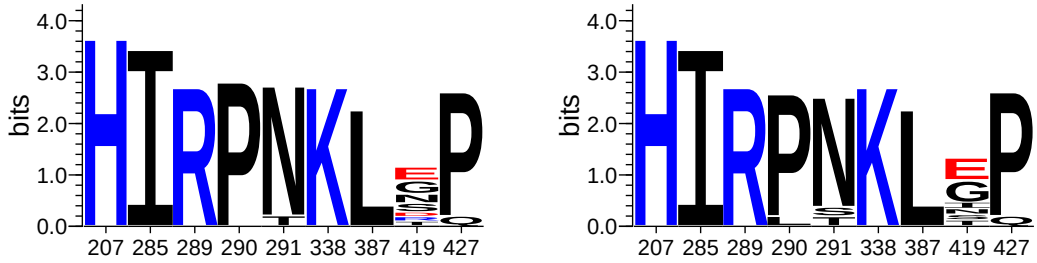
(B) B vs C - pH 4 & 7

(C) B vs C - pH 5 & 8

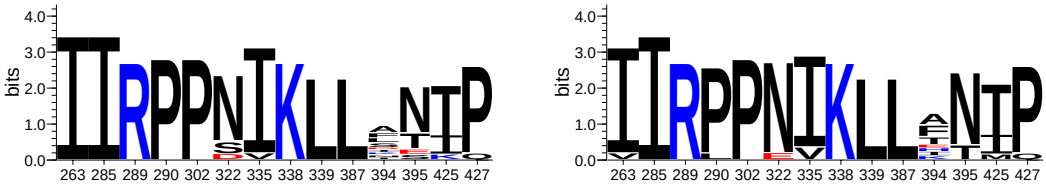
Figure F.2: B vs C Sensitive Residue Composition Considering Gaps. A) Comparison of the top 1% of residues identified identified from Figure 30A. B) Comparison of the top 1% of residues identified in Figure D.3A. C) Comparison of the top 1% of residues identified in Figure D.4A.



(A) TF vs CC - Average pH

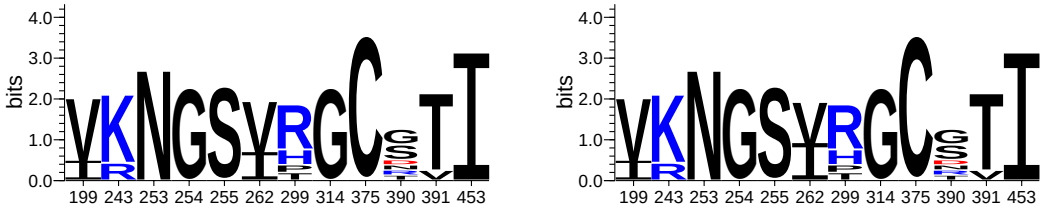


(B) TF vs CC - pH 4 & 7

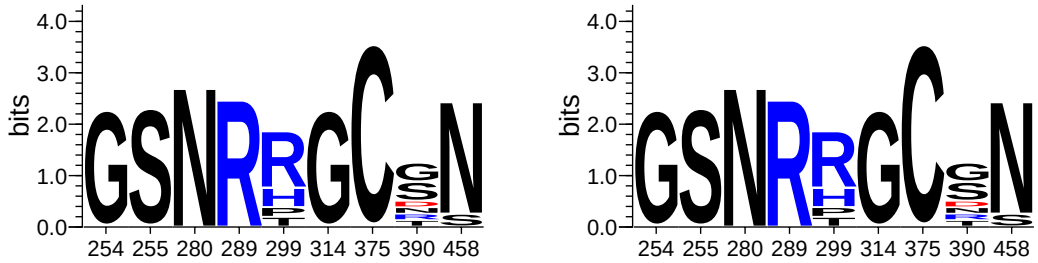


(C) TF vs CC - pH 5 & 8

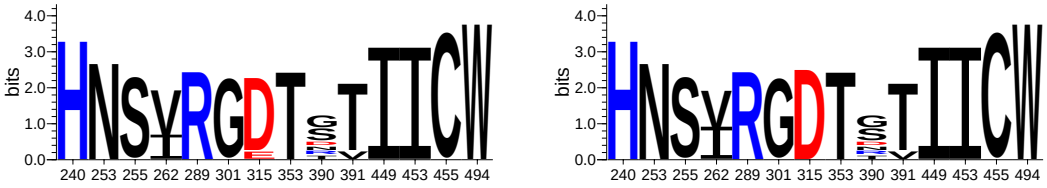
Figure F.3: TF vs CC Sensitive Residue Composition Considering Gaps Within B Clade. A) Comparison of the top 1% of residues identified identified from Figure 29B. B) Comparison of the top 1% of residues identified in Figure D.1B. C) Comparison of the top 1% of residues identified in Figure D.2B.



(A) TF vs CC - Average pH

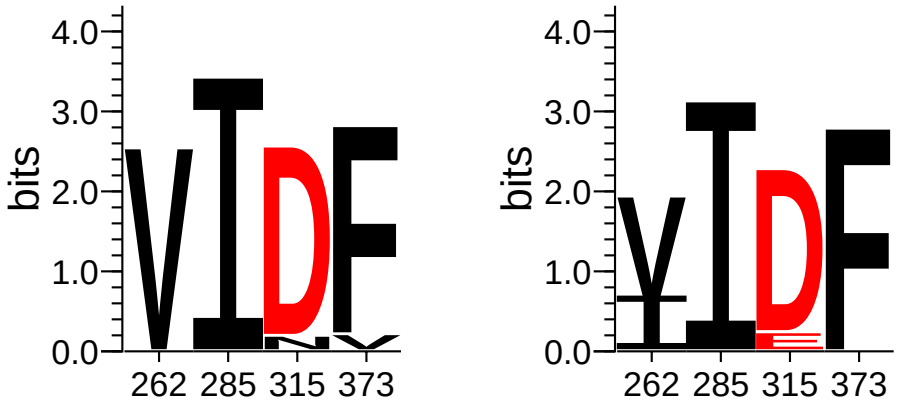


(B) TF vs CC - pH 4 & 7

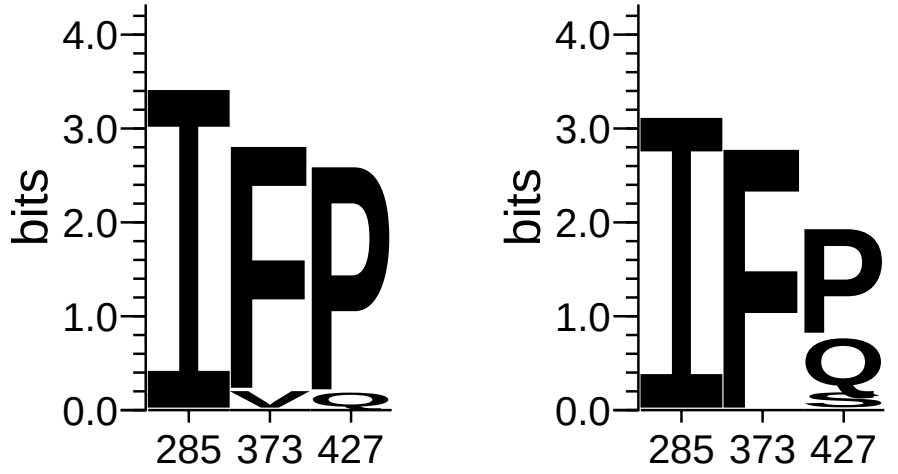


(C) TF vs CC - pH 5 & 8

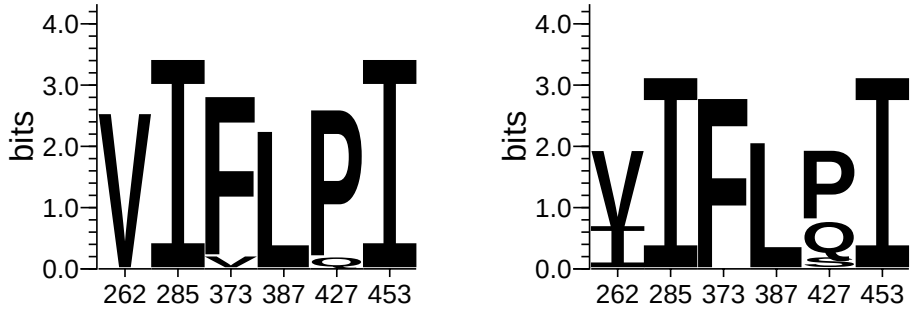
Figure F.4: TF vs CC Sensitive Residue Composition Considering Gaps Within C Clade. A) Comparison of the top 1% of residues identified identified from Figure 29C. B) Comparison of the top 1% of residues identified in Figure D.1C. C) Comparison of the top 1% of residues identified in Figure D.2C.



(A) B vs C - Average pH

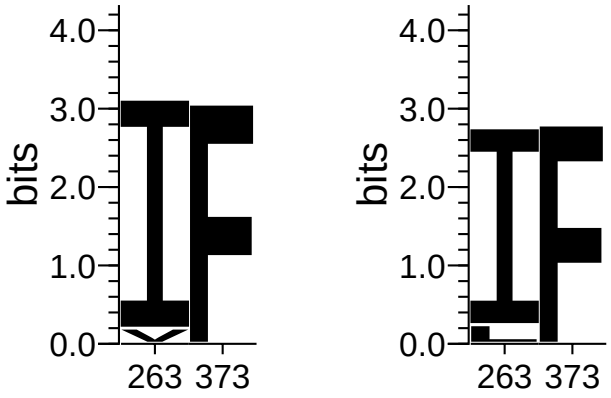


(B) B vs C - pH 4 & 7

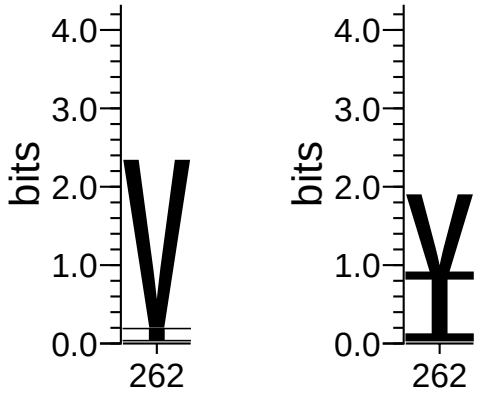


(C) B vs C - pH 5 & 8

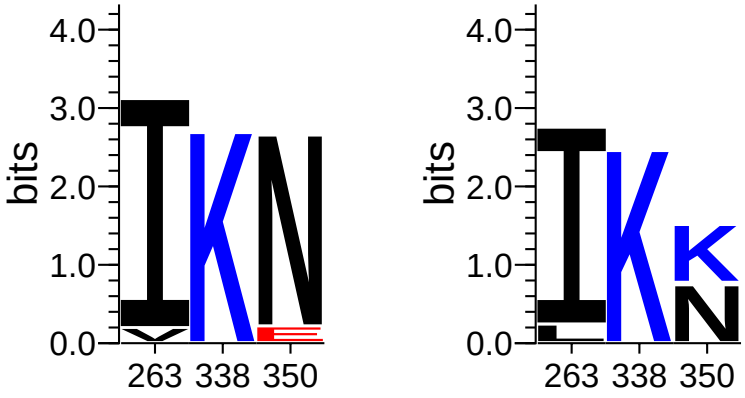
Figure F.5: B vs C Sensitive Residue Composition Considering Gaps Within TF Class. A) Comparison of the top 1% of residues identified identified from Figure 30B. B) Comparison of the top 1% of residues identified in Figure D.3B. C) Comparison of the top 1% of residues identified in Figure D.4B.



(A) B vs C - Average pH



(B) B vs C - pH 4 & 7



(C) B vs C - pH 5 & 8

Figure F.6: B vs C Sensitive Residue Composition Considering Gaps Within CC Class. A) Comparison of the top 1% of residues identified identified from Figure 30C. B) Comparison of the top 1% of residues identified in Figure D.3C. C) Comparison of the top 1% of residues identified in Figure D.4C.