LOYOLA UNIVERSITY CHICAGO

Preparing people to lead extraordinary lives

Information Convergence and Divergence of Protein Primary Structures Encoded by HIV-1

By Mohammed B. Syed and Daniel J. Graham
Department of Chemistry and Biochemistry
Loyola University Chicago, Chicago, Illinois, USA

Abstract

The operations of the proteins encoded by the HIV-1 virus are highly coordinated within the human host. The coordination derives from mutual information expressed in the primary structures. These are the amino acid sequences encoded by the viral genome that carry the requisite information about folding pathways and chemical functions. This project examines the information-convergent properties of HIV-1 proteins that underpin the overlap and reinforcement of functions. The goal is to identify proteins of the virus that express the highest degree of information coordination. These proteins pose compelling targets for small-molecule drug and cocktail therapy and vaccine engineering.

Introduction

The primary structures of viral proteins typically present as random sequences. They are represented by alphabetic letters. The letters refer to individual amino acids linked by peptide bonds in the order encoded by the genome from N- to C-terminal. The alphabet is familiar to molecular biologists and virologists. Within HIV-1, the proteins Asp and Vpu manifest different lengths and offer no appreciable overlap of sequences. Clearly the molecules are endowed with different biological functions critical to the viral operations. Asp has been shown to be recognized and targeted by CD8+ T host cells during infection phases. In contrast, Vpu has been demonstrated to enhance virion budding by targeting host CD4 and Tetherin/BST2 and by enhancing proteasome degradation. There are eight other proteins encoded by the HIV-1 virus whose structures and functions have been researched internationally over decades.

Methodology

Our research pursues how the HIV-1 encoded primary structures can be re-expressed as integer quantities with zero loss of information. The methodology directs a variant of Gödel-type coding while taking the lexicographic order of the amino acids to be a random variable. Each order possibility offers a unique, alternative way of viewing the primary structure information. To demonstrate how this works, consider a lexicographic string variable of AVLIPFWMGSTCYNQDEKRH while revisiting the sequence for Asp:

	M	P	Q	T	V	S	C	.....
k	1	2	3	4	5	6	7	.....
p	2	3	5	7	11	13	17	.....

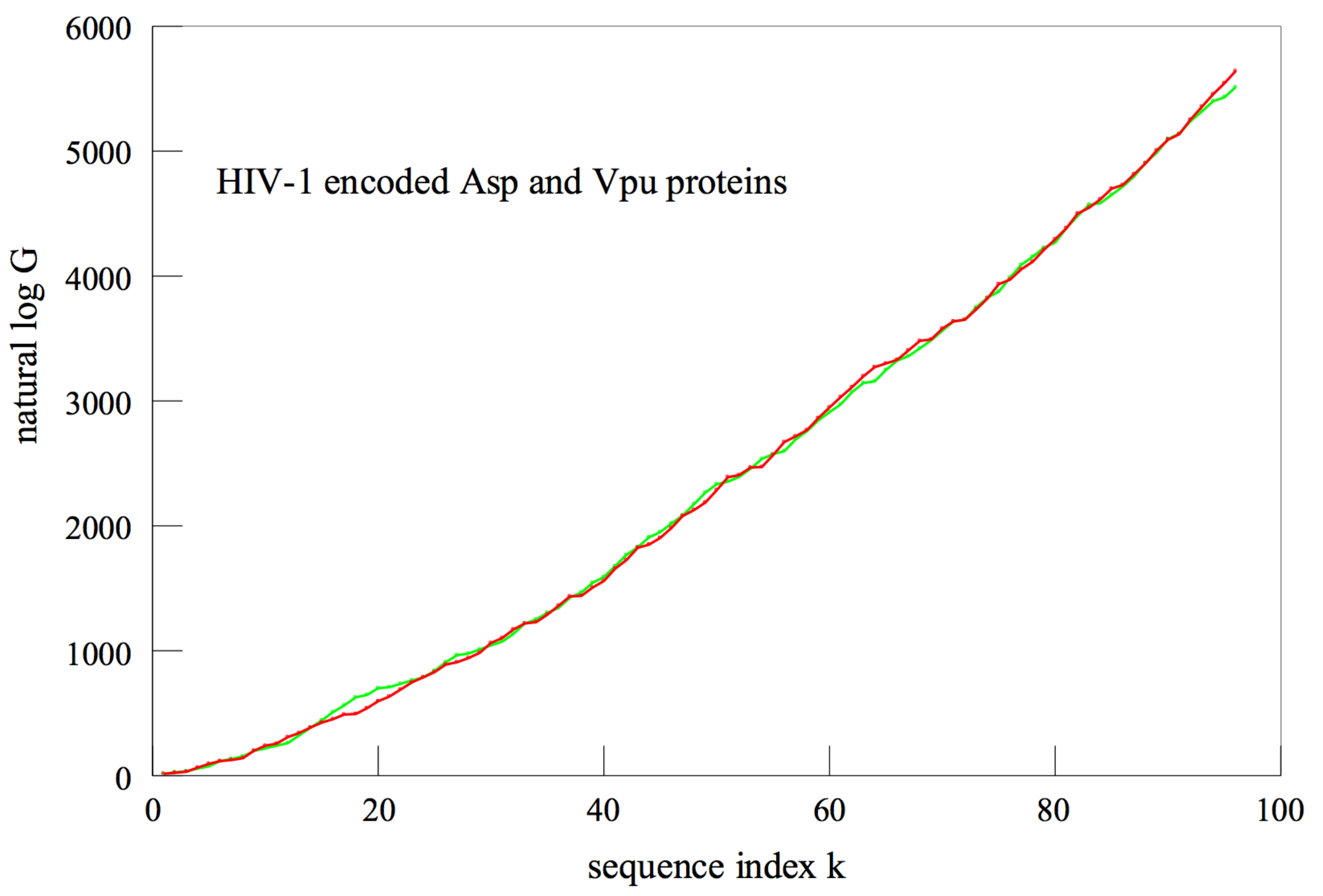
$$G(k) \quad 2^{8} \ \mathrm{x} \ 3^{5} \ \mathrm{x} \ 5^{15} \ \mathrm{x} \ 7^{11} \ \mathrm{x} \ 11^{2} \ \mathrm{x} \ 13^{10} \ \mathrm{x} \ 17^{12} \ \mathrm{x} \ \ldots..$$

The above communicates that the prime numbers *p* are deployed in ascending order, each raised to a power determined by the alphabetic position *j* of the particular amino acid. For the alphabetic string AVLIPFWMGSTCYNQDEKRH, we have:

	A	V	L	I	P	F	W	.....	K	R	H
j	1	2	3	4	5	6	7	.....	18	19	20

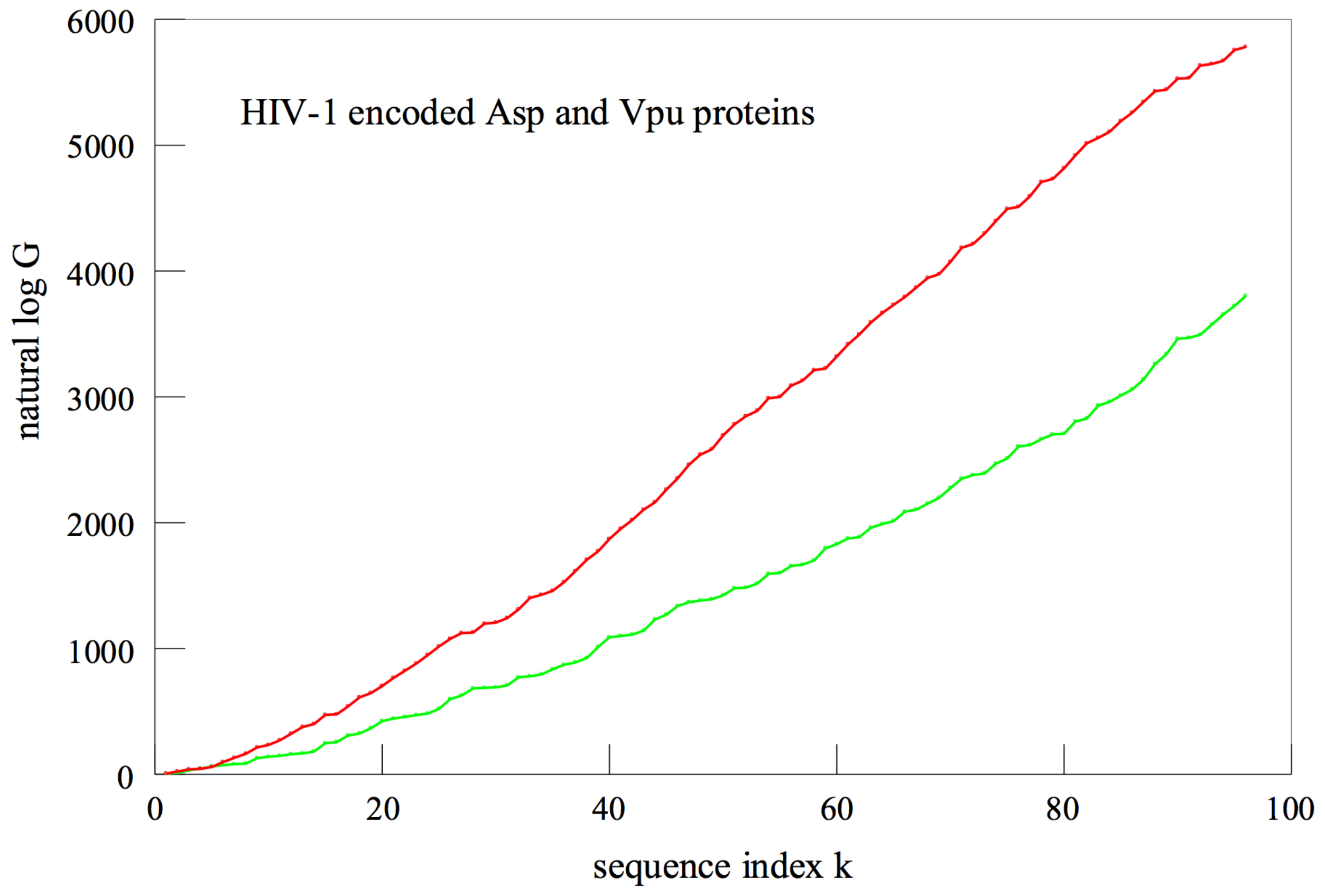
The function *G*(*k*) delivers an integer specific to the sequence and lexicographic order. It is straightforward to retrace steps to recover the alphabetic representation of the primary structure. What is most critical is that *G*(*k*) offers multiple and diverse perspectives of a protein structure beyond a single letter string.

The experiments in the research are computational searches across the space of lexicographic orders that *align* the *G*(*k*) for different proteins. Such orders draw out the information convergence of proteins that is not otherwise apparent in the letter representations of primary structures. The experiments go further to identify orderings that push *G*(*k*) for proteins apart. These draw out the information divergence of primary structures beyond that apparent in the letter sequences. The figure below illustrates *G*(*k*) data for Asp and Vpu proteins encoded by HIV-1. The trace colors correspond to the font colors used above for the letter sequences. The traces represent *G*(*k*) based on the lexicographic order WFDKQCNVEHTPLRISAYGM. Note the strikingly close alignment. It testifies that if were were to reconstruct the primary structure of Asp from its *G*(*k*), the *G*(*k*) for Vpu would offer an alternative and indeed powerful vehicle for doing so.



Methodology (continued)

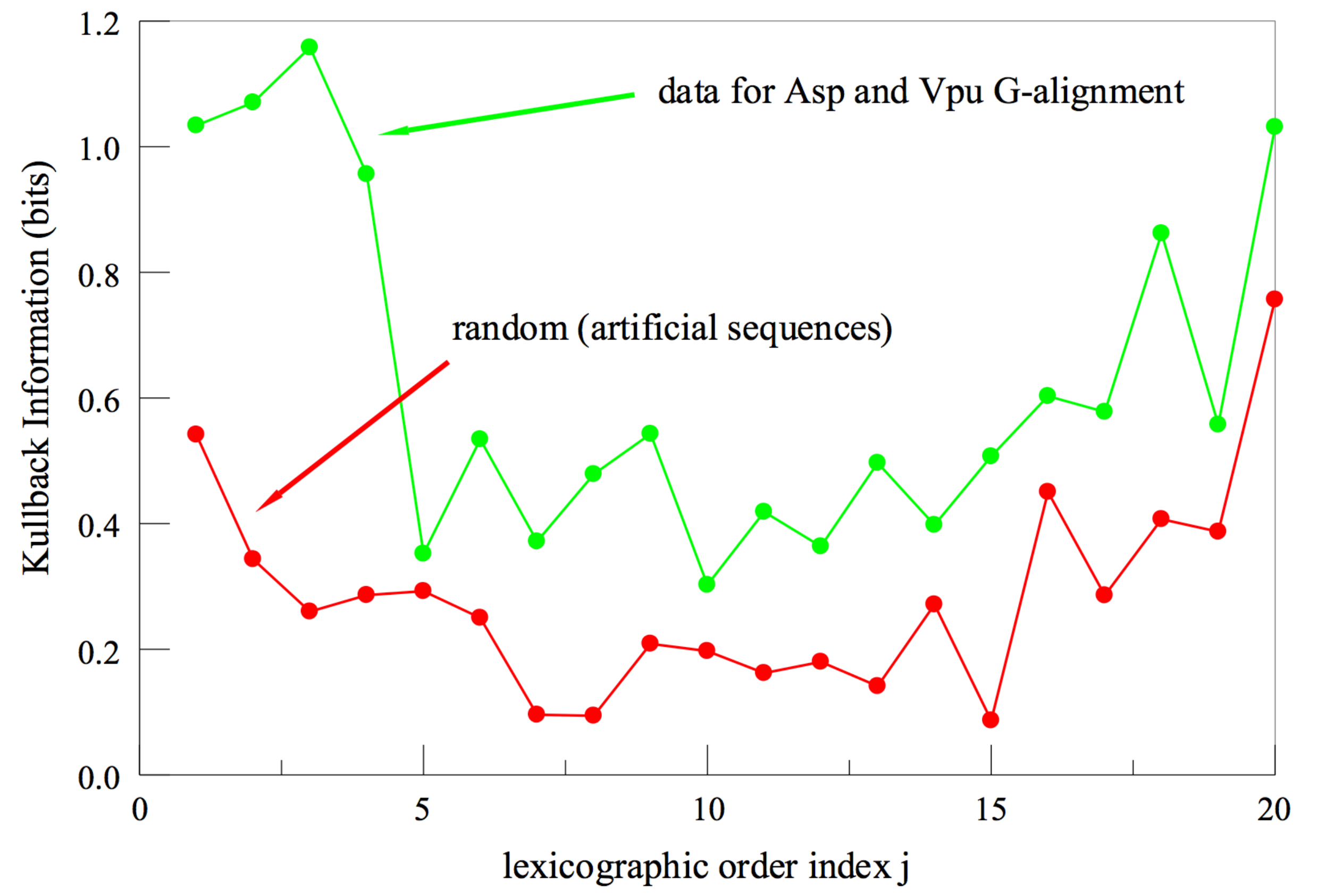
In contrast, the figure below shows *G*(*k*) for Asp and Vpu proteins using NACSFPVTMKQDLREGIWHY as a lexicographic order. Note the marked divergence of the traces. Here the lexicographic order draws out the contrasts in information expression. Such contrasts underpin the complementary functions carried by the proteins.



Results

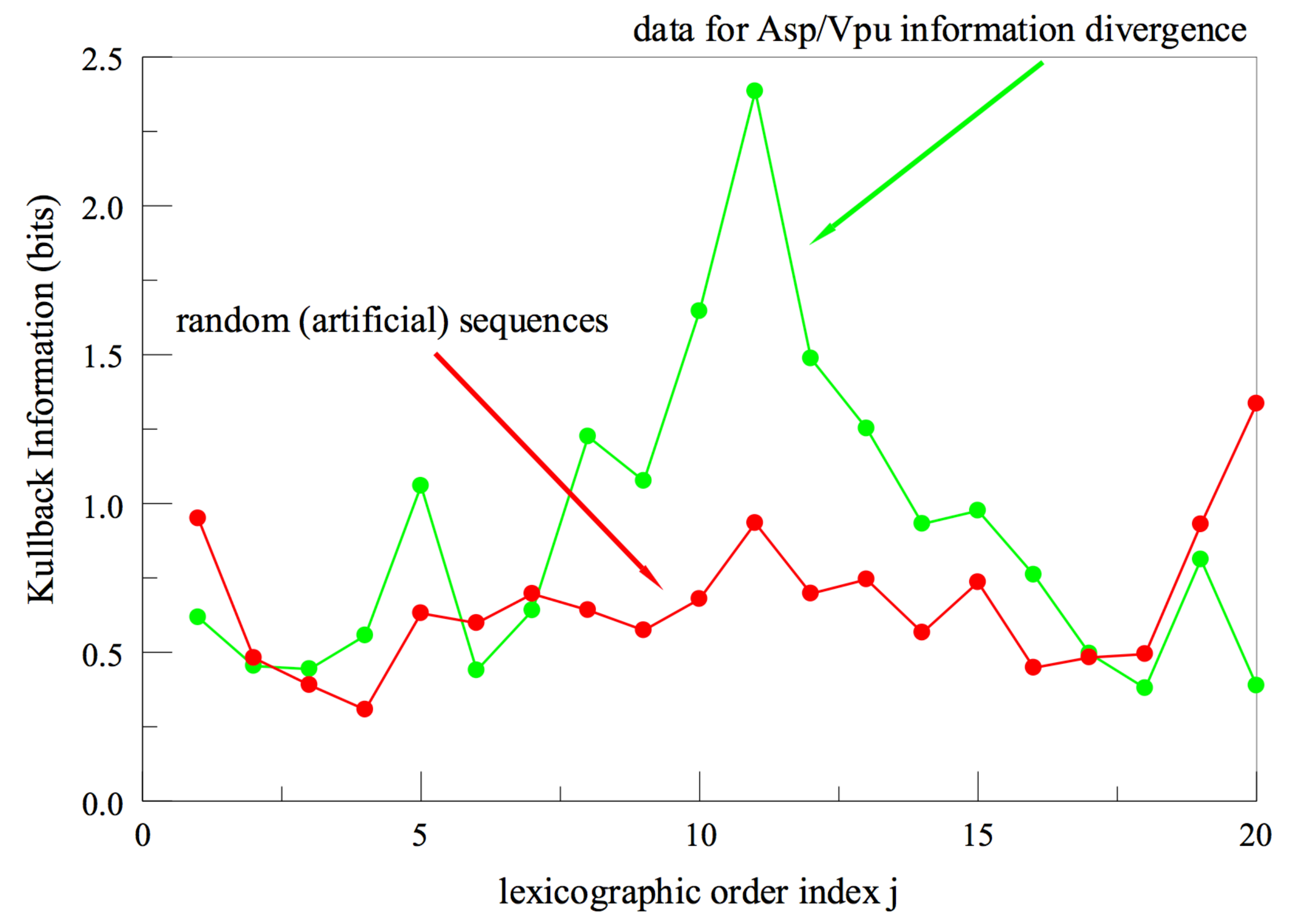
Ten proteins are encoded by the HIV-1 genome labeled and are traditionally labeled using acronyms: Asp, Gag-pol, Nef, Vpr, Vpu, Gp160, Vif, Gag-Pr55, Rev, Tat. They were examined in pairs to establish the information convergence and divergence. Given ten primary structures, there were 10 x 9 / 2 = 45 pairs to examine over the course of research. The *G*(*k*) were determined as in the above figures to establish *distributions* of alphabetic orderings that minimized and maximized the information differences.

Given the distributions of lexicographic orders, the Kullback information *j*-labeled position. In words, the Kullback information measures the departure from randomness in the lexicographic orders in relation to the protein sequences. There is a measure in bits for each of *j* = 1, 2, 3, …, 20 sites in an order such as NACSFPVTMKQDLREGIWHY. An example of the Kullback information for Asp-Vpu information convergence (green trace) is shown in the plot below:



The plot above shows that the amino acids at the beginning and end of the lexicographic order distribution contribute most significantly to the information convergence. For control experiments, the computations were extended to sequences assembled by a random number generator that matched Asp and Vpu in size. We observe the Kullback information to be suppressed significantly as expected. This attests to the dearth of information coordination of randomly assembled proteins.

In contrast, the green trace in the plot below shows the Kullback information for the information divergence of the Asp and Vpu proteins. As in the previous figure, the data are shown for sequences assembled by random number generation.



Results (continued)

Here, we observe how the amino acids in the middle of a lexicographic order contribute most to the information divergence. We further witness that artificial sequences lack the information divergence conferred by natural proteins.

The standard deviation σ_I of the Kullback information across the lexicographic order index gauges the intensity of *G*(*k*) alignment for two proteins. The corresponding statement applies to the *G*(*k*) divergence. The computational experiments showed four results of greatest significance. First was that the strength of information *convergence* of the proteins encoded by HIV-1 considerably exceeds the strength of divergence. To wit, of the top twenty σ_I values, fourteen were allied with information convergence. If there were no convergence bias expressed by the proteins, there would be less than four percent chance of observing such a result. Second, the most intense information divergence was expressed by Asp and Vpu proteins used in the foregoing figures. Third, the most intense information convergence was expressed by the Vpr and Rev proteins. Lastly, the most intense convergence and divergence involved the Asp and Rev proteins. In other words, Asp and Rev expressed the strongest coordination of functions across the HIV-1 proteome.

Discussion/Conclusion

Based on the results above, it is highly improbable that such information convergence was due to chance. This improbability indicates that the HIV-1 proteins indeed coordinate among on another. But which proteins express the highest degree of information coordination? The latter three results answer this question. The intense divergence between Asp and Vpu proteins can be explained by their respective functions. Research indicates that Asp expression is heightened during infection. CD8+ T cells of infected individuals target Asp. With regards to the function of Vpu, the protein is responsible for binding to CD4 which enhances virion budding. Like CD8, CD4 is also a glycoprotein found on the surface of host T cells. A possible explanation for the divergence between Asp and Vpu is that Asp is targeted by a glycoprotein whereas Vpu targets a glycoprotein.

The convergence between Vpr and Rev can also be explained by their functions. Vpr plays a role in host cell infection and specifically acts by transporting the viral pre-integration complex to the nucleus. Rev escorts unspliced or incompletely spliced viral pre-mRNAs out of the nucleus. Most HIV-1 proteins are translated from these pre-mRNAs. The escort protein is needed because pre-mRNAs cannot readily exit the nucleus like fully processed mRNAs. The entry and exit of viral fragments into and out of the host nucleus may be tightly regulated and may explain the convergence of Vpr and Rev. Perhaps the amount of Vpr that transports the viral pre-integration complex into the nucleus is synonymous to the amount of Rev that escorts pre-mRNAs out of the nucleus.

The functions of Vpr and Rev are vital to the life cycle of HIV-1 as are the functions of the other proteins. The fact that Vpr and Rev represent the most intense convergence makes sense because without Vpr and Rev, the DNA replication mechanisms of the host cell cannot be utilized in the first place. The results indicate that the deactivation of Vpr or Rev would potentially cause a significant halt in HIV-1 infection of the host. However, further studies are needed to strengthen the claim.

References

Bet, A., Maze, E., Bansal, A., Sterrett, S., Gross, A., Graff-Dubois, S., … Cardinaud, S. (2015). The HIV-1 Antisense Protein (ASP) induces CD8 T cell responses during chronic infection. *Retrovirology*, *12*(1), 15. doi: 10.1186/s12977-015-0135-y

UniProt ConsortiumEuropean Bioinformatics InstituteProtein Information ResourceSIB Swiss Institute of Bioinformatics. (n.d.). UniProt Consortium. Retrieved from https://www.uniprot.org/