

Evidence for Functional Protein Fragment Homology in Viral Genome Types

John R. Rose and Rishi Mukhopadhyay

Abstract— In this paper evidence is presented that supports the hypothesis that amino acid usage bias is a fundamental property of viral genome types. Clues to the biological basis for the observed differences in viral amino acid usage are examined. Capsid proteins are analyzed to evaluate the hypothesis that replication mechanism is exclusively responsible for observed amino acid usage profile bias. DSSP secondary structure characterization data is examined to evaluate the competing hypothesis that genome types universally use different sets of functionally equivalent amino acid fragments as building blocks.

Index Terms—Amino Acid Usage, Viral Genome Type, Functional Homology, Viral Classification

I. INTRODUCTION

PREVIOUS research has shown a significant correlation between amino acid usage and viral genome type in a data set of 236 mammalian viruses [12]. Surprisingly, no correlation was found between viral amino acid usage and host. While demonstrating the existence of amino acid preference, this earlier work did not address the deeper question of the biological basis for the observed bias. In this paper we investigate the basis for the observed amino acid bias. We consider several hypotheses that might explain the observed amino acid preferences. These investigations provide evidence to support the hypothesis that amino acid usage bias is a fundamental viral genomic type property.

In previous studies, it was demonstrated that amino acid usage can be used to classify mammalian viruses by viral genomic type [12]. What accounts for the observed differences in amino acid usage? The Baltimore viral genome type classification is based on viral replication mechanism. One hypothesis is that the differences in amino acid usage are tied to differences in replication mechanism. If this hypothesis were to fully account for the observations, then one might

expect structural proteins from different genome types to show little or no differences in amino acid usage. In other words, proteins that are only tangentially involved in replication should not be as strongly biased in amino acid usage as proteins directly involved in replication. A second hypothesis is that the different genome types universally use different complements of functionally equivalent amino acid fragments as building blocks. If this hypothesis were to fully account for the observed amino acids usage then one would expect structural proteins from different genome types to exhibit the same differences in amino acids usage as do their complete genomes. We examine both of these hypotheses in this paper.

II. DATA AND METHODOLOGY

A. Genome Data

The set of mammalian genomes investigated in this paper was downloaded from NCBI. For those viruses for which multiple examples of a given virus are available, only one example was included to avoid biasing the dataset towards overrepresented genomes. The breakdown of viruses by genome type is 60 *dsDNA*, 42 *retroid*, 16 *ssDNA*, 42 *ssRNA negative strand*, and 76 *ssRNA positive strand*. A list of the particular viruses included in the dataset can be found at <http://www.cse.sc.edu/~rose/aminoPreference/SupplementaryData/classificationFilesUsed.htm>.

B. Capsid Protein Data

The capsid protein data set was constructed by selecting all genes labeled as coding for capsid proteins from the set of 236 mammalian genomes described in the previous section (*Genome Data*). A total of 132 capsid genes were collected. The composition of the data set of genes labeled as coding for capsid proteins was: 80 *dsDNA*, 10 *ssDNA*, 23 *ssRNA negative strand*, 19 *ssRNA positive strand*. The *retroid* virus genomes in our original data set use a different nomenclature to indicate structural proteins. Consequently, our automatic script for extracting genomes was not able to recognize structural proteins in *retroid* viruses and none were included in the set of 132 genes. From this set of 132 of genes, capsid amino acid usage profiles were then calculated.

Manuscript received April 20, 2006. This work was supported in part by the U.S. Department of Agriculture under Cooperative Agreement No. 58-5438-2-341.

John R. Rose is with the University of South Carolina, Columbia, SC 29208 USA (phone: 803-777-2405; fax: 803-777-3767; e-mail: rose@cse.sc.edu).

Rishi Mukhopadhyay is with the University of South Carolina, Columbia, SC 29208 USA (e-mail: rishi@cse.sc.edu).

C. Amino Acid Usage Profiles

The amino acid usage profiles for the capsid genes for the four genome types (*dsDNA*, *ssDNA*, *ssRNA negative strand*, and *ssRNA positive strand*) were derived using the methodology described in our previous analysis of mammalian viruses [12]. The triple amino acid preference classification model (3-AAP) is based on the distribution of ordered triples of amino acids of single viral genome type. The basic methodology for creating the profile is to first extract coding sequences and translate into the corresponding amino acid representation to produce ordered triple amino acid preference distributions on a per genome basis. For capsid profiles, each capsid gene is profiled separately and treated as a separate data point. For 3-AAP models, overlapping triples are extracted from the coding sections of genomes. If $\langle a_1 a_2 a_3 \dots a_n \rangle$ is a contiguous sequence of n amino acids, there are $n-2$ ordered triples, *i.e.*, $\langle a_1 a_2 a_3 \rangle$, $\langle a_2 a_3 a_4 \rangle$, ..., $\langle a_{n-2} a_{n-1} a_n \rangle$. The 3-AAP data for a capsid protein is computed by tabulating the number of occurrences of each of the 8000 (20^3) possible ordered triples in the protein and then normalizing the resulting distribution. Similarly, 3-AAP data for an entire genome is computed by tabulating the number of occurrences of each of the 8000 (20^3) possible ordered triples represented by the set of genes in the genome and then normalizing the resulting distribution.

D. SVM Classification

For classification, we used a supervised learning method based on the support vector machine (SVM), which belongs to the class of kernel-based learning methods [14]. Since the focus of this study is to test the hypothesis that amino acid preference is tied to replication mechanism and not to determine which classification method is optimal for this class of problem, we did not investigate alternative classification methods.

The SVM package that was used, SVM^{light}, was developed by Thorsten Joachims [6], [7]. This package is available for download at <http://svmlight.joachims.org/>. The format of the training data for this package is the class membership, which can be 1 or -1 for positive or negative example, followed by the feature vector.

Since the SVMs supported by SVM^{light} are dichotomizers, multiple two-class problems were formulated such that each class is contrasted with the remaining classes. For example, the training file for an SVM model that is used to distinguish *ssRNA positive strand* genomes from other viral genome types based on the usage of ordered triples of amino acids contains the positive examples, *i.e.*, data from the *ssRNA positive*

strand genomes and the negative examples, *i.e.*, data from all non-*ssRNA positive strand* genomes. Each data point in the training set is an amino acid usage profile with 8000 features. A positive example in the training dataset is represented by a 1, denoting a positive example, followed by the 8000 features (percentages of each ordered amino acid triple). Likewise, each negative example in the training dataset is represented by a -1, denoting a negative example, followed by the 8000 features.

E. Secondary Structure of Viral Amino Acid Triples

The DSSP version [8] of the Protein Data Bank (PDB) [16] was searched for *dsDNA* and *ssRNA positive strand* virus data. The resulting data was stored in separate files for each viral class. Specifically, all PDB *dsDNA* and *ssRNA positive strand* viral entries that contained both “SEQUENCE” and “DSSP” tags but not the term “mutation” or “mutant” were collected. The data for each viral class was then analyzed separately.

DSSP summarizes secondary structure using the following table of structure codes:

- “H” = 4-helix (α -helix)
- “B” = residue in isolated β -bridge
- “E” = extended strand
- “G” = 3-helix
- “I” = 5-helix
- “T” = H-bonded turn
- “S” = bend

Since the entries were extracted from PDB, the symbol “C” is also present as a secondary structure code. It indicates a loop or irregular element. DSSP itself marks such positions with a blank. However, programs such as PDBFINDER replace this blank by a “C”.

The structural code “C” taken together with the 7 standard DSSP codes potentially results in 8 possible codes for structural characterization of each position. The list of structure codes is prioritized in the order they are listed above. When there are structural overlaps, priority is given to the structure first in this list. Thus not all combinations of possible combinations of triple codes appear.

For each viral class the secondary structure summary was parsed into triples. Next, the corresponding amino acid triples were collected. From these collections the amino acid distribution for each secondary structure triple was determined. This resulted in a *dsDNA* amino acid distribution and an *ssRNA positive strand* amino acid distribution for each secondary structure triple class. An example of a secondary structure triple class is shown in Table I.

TABLE I
SECONDARY STRUCTURE TRIPLE: HHH

Top <i>dsDNA</i> triples	<i>dsDNA</i> frequency	<i>ssRNA</i> + frequency	Top <i>ssRNA</i> + triples	<i>ssRNA</i> + frequency	<i>dsDNA</i> frequency
RYL	0.006335	0.000951	GEI	0.006891	0.000341
VLA	0.006062	0.000119	RAA	0.006891	0.000954
AVL	0.005722	0.000000	EAM	0.005465	0.000341
AAL	0.005517	0.001663	LPQ	0.005465	0.000000
AAR	0.005177	0.000238	RVE	0.005228	0.000000
TTQ	0.005109	0.000000	LRK	0.005109	0.000545
TTT	0.005041	0.000000	VEE	0.004871	0.000886
AAV	0.004496	0.000475	LLE	0.004752	0.001090
LAA	0.004496	0.002020	RKL	0.004515	0.000000
RTF	0.004155	0.000000	SKF	0.004396	0.000136

Secondary structure triple distribution for HHH. This table compares the top 10 most frequent amino acid triples with secondary structure code HHH from *dsDNA* viral proteins and *ssRNA positive strand* viral proteins in PDB.

The amount of secondary structure data in PDB for *dsDNA* and *ssRNA positive strand* viruses was not sufficient to populate adequately all of the secondary structure classes that appeared for triples. Consequently, we arbitrarily limited our focus to the top 20 most populated secondary structure classes. As will be discussed in the subsequent section, this set of 20 secondary structure classes provide adequate evidence for examining the hypothesis that viral genome types preferentially use functional equivalence classes of protein fragments as building blocks. In the future we expect to compile a larger data set that would allow us to consider additional secondary structure classes.

TABLE II
SUMMARY OF CAPSID PROTEIN CLASSIFICATION

	<i>dsDNA</i>	Retroid	<i>ssDNA</i>	<i>ssRNA</i> Negative	<i>ssRNA</i> Positive	Total	% Correct
<i>dsDNA</i>	65	0	0	4	11	80	81.25
<i>ssDNA</i>	0	0	10	0	0	10	100.00
<i>ssRNA</i> Negative	0	0	0	23	0	23	100.00
<i>ssRNA</i> Positive	0	4	1	3	11	19	57.89

Performance of the 3-AAP models on capsid protein data.

III. RESULTS AND DISCUSSION

A. Classification of Structural Proteins

We examined the hypothesis that the differences in amino acid usage are tied exclusively to differences in replication mechanism. Our approach was to examine the performance of the amino acid preference classifiers on structural proteins. If this hypothesis were to fully account for the observations, we would expect structural proteins from different genome types to show little or no difference in amino acid usage. We would expect the classification rate to fall significantly from the high rate (>95%) for whole genome data.

The results are shown below in Table II. These classification results were derived using SVM classifiers trained on whole-genome data. The classifiers were not tuned specifically for capsid data. The point of this study was to contrast structural-protein-only data with whole-genome data using the same classifiers. In the *dsDNA* case, there were 80 genes that were labeled as coding for capsid proteins. From Table II we see that 65 were correctly classified by the *dsDNA* classifier trained on whole-genome data. Eleven were misclassified as *ssRNA positive strand* and 4 were misclassified as *ssRNA negative strand*. In both the *ssDNA* and *ssRNA negative strand* cases, all capsid proteins were correctly classified. However, classification performance drops significantly for *ssRNA positive strand* capsid proteins. Only 11 out of 19 are correctly classified. Of the 9 misclassified *ssRNA positive strand* proteins, 4 are misclassified as *retroid*, 1 is misclassified as *ssDNA*, and 3 are misclassified as *ssRNA negative strand*. The overall classification rate is 83%.

In contrast, Table III shows the result for whole-genome classification from our previous study [12]. As noted in the table caption, cross-validation [9], [13] was used to evaluate the performance and robustness of the models for whole-genome classification. Cross-validation was carried out by taking the total available set of genome datasets and partitioning it into 10 approximately equal-sized sets. The genomes in each partition were randomly selected. Each partition

contained approximately one tenth of the available genomes of each viral genome type. We then used 9 partitions to train the models and tested with the remaining partition. This was repeated nine times, leaving in turn a different partition of the data out of the training set and using it to validate the resulting models. The classification results shown in Table III are thus the aggregate results of these 10 training/testing runs. As an aside, we have also used the bootstrap [3], [4] to evaluate the SVM classifiers for whole-genome and partial sequence data. Those results have been reported elsewhere [12].

A closer examination of the classification results for *ssRNA positive strand* capsid proteins reveals a

correlation between capsid protein length and classification accuracy. As shown in Figure 1, the misclassified capsid proteins tend to be shorter than the correctly classified proteins.

High classification rates for such short proteins is not a surprise. A statistical analysis of the triples shows that a small number of triples are statistically most significant for distinguishing between viral types.

TABLE III
SUMMARY OF WHOLE GENOME CROSS-VALIDATION

	dsDNA	Retroid	ssDNA	ssRNA Negative	ssRNA Positive	Total	% Correct
dsDNA	58	0	0	0	2	60	96.67
Retroid	0	42	0	0	0	42	100.00
ssDNA	1	0	14	0	1	16	87.50
ssRNA Negative	0	0	0	41	1	42	97.62
ssRNA Positive	1	0	0	0	75	76	98.68

Performance of cross-validation testing over 10 runs for the 3-AAP model on whole-genome data.

Comparing the results in these Tables II and III we see that indeed the classification rate for structural proteins does fall in comparison to that for whole genome data. However, the observed overall classification rate for 83% shown in Table II indicates that there are nevertheless significant differences in amino acid usage in structural proteins for different genome types. The drop off in classification rate for *ssRNA positive strand* capsid proteins can not be attributed solely to the short length of the misclassified proteins. Previous analysis of bootstrapped sequences of 300 amino acids in length has yielded a classification rate of 98% for *ssRNA positive strand* viruses [12]. This suggests that these capsid proteins may not have as strong an amino acid usage bias as non-structural proteins. Additional evidence that shortness of capsid protein length may not be the sole cause of classification error is provided by *dsDNA* capsid proteins. Out of 80 *dsDNA* capsid proteins, 32 are shorter than 400 amino acids in length. Yet, 25 out of 32 of these short proteins are correctly classified. In fact, 10 out of 14 *dsDNA* capsid proteins shorter than 200 amino acids in length are correctly classified.

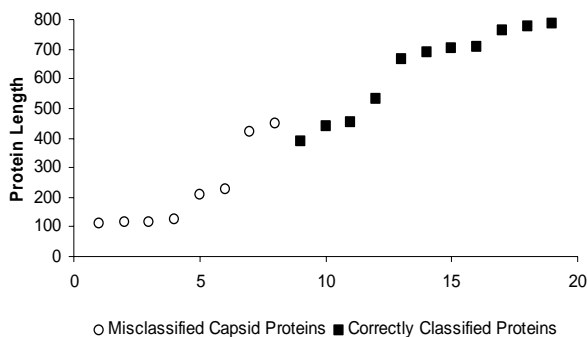


Fig. 1. Examination of the 19 *ssRNA positive strand* capsid proteins shows that shorter proteins tended to be misclassified.

Bootstrap sequence analysis indicates that these significant triples are spread throughout viral genomes. A preliminary analysis of the data using the R statistics package [2], [5] indicated that the triple amino acid distributions do not satisfy normality assumptions. We then performed a more rigorous distribution analysis for normality with the Kolmogorov-Smirnov test. The results confirmed the lack of normality. Of the 8000 triple amino acid distributions, only 8 are normal in all 5 viral genome types. Hence the Kruskal-Wallis test, a nonparametric analysis of variance [10], [11], [15] rather than ANOVA was performed. The Kruskal-Wallis test was used to establish which amino acids triples exhibit a statistical difference between viral genome types. Next, we used a multiple comparison procedure [1] based on Kruskal-Wallis rank sums to find where the differences occur. Table IV shows the number of significant amino acids triples for *p-values* from 10^{-5} down to 10^{-9} as determined by the multiple comparison procedure.

TABLE IV
SIGNIFICANCE ANALYSIS OF AMINO ACID TRIPLES

Significance	Amino Acid Triples
$p < 1$	8000
$p < 10^{-5}$	1530
$p < 10^{-6}$	927
$p < 10^{-7}$	529
$p < 10^{-8}$	310
$p < 10^{-9}$	197

Number of significant amino acid triples for selected *p-values*.

These classification results indicate that the observed differences in amino acid usage are not exclusively tied to differences in replication mechanism since the bias is also observed in structural proteins. On the other hand, this does not imply that differences are independent of replication mechanism. The drop in classification rate for structural proteins as compared with whole-genome data could be explained by significant differences in

amino acid usage tied to replication mechanism. Our interpretation is that replication mechanism is an important source of amino acid usage bias but it is not the only source.

B. Secondary Structure

As can be seen in Table I, there is no overlap between the top 10 most frequently occurring *dsDNA* amino acid triples with secondary structure code HHH and *ssRNA positive strand* amino acid triples. This suggests that these two genome types typically employ different amino acids triples to create this particular secondary structure.

In the top 20 most populated secondary structure classes there were only 4 cases where an amino acid triple appeared in the top 10 list for both *dsDNA* and *ssRNA positive strand* viruses. This is a 2% overlap for most frequently occurring secondary structure triples. In many cases an amino acid triple does not appear at all in one of the viral genomes types in this data set. For example, in Table I, the amino acid triples TTQ, TTT, and RTF which appear in the top 10 list for *dsDNA* viruses have a frequency of 0 for *ssRNA positive strand* viruses. Similarly, LPQ, RVE, and RKL which appear in the top 10 list for *ssRNA positive strand* viruses have a frequency of 0 for *dsDNA* viruses in this data set.

The distributions imply a functional protein fragment equivalence set. They can be interpreted to represent a type of functional protein fragment homology between these genome types. This evidence supports the hypothesis that different genome types use different complements of functionally equivalent amino acid fragments as building blocks.

IV. CONCLUSION AND FUTURE WORK

Previous research suggests that structure in amino acid preference can be used to predict viral genome type. Up until now, the mechanism that has resulted in the observed relation between amino acid preference and viral genome type has not been investigated. This paper examines several possible hypotheses that could account for the observed amino acid bias. Evidence presented in this paper indicates that viral genome type amino acid usage preferences are not limited to proteins directly involved in viral replication but also occur in structural proteins. This suggests that while some differences in amino acid usage may directly be tied to differences in replication mechanism, other differences may be only indirectly related or may have some other basis.

Second, the lack of overlap in secondary structure triples in the case of *dsDNA* and *ssRNA positive strand* viruses supports the hypothesis that viral genome types universally use different complements of functionally

equivalent amino acid fragments as building blocks.

The results presented in this paper provide evidence that helps to distinguish between competing explanations of the basis for amino acid bias. More extensive data sets must be compiled and analyzed in greater detail in order to substantiate the preliminary results presented in this paper that suggest that different viral genome types use functional equivalence classes of protein fragments. Furthermore, structural and nonstructural proteins must be analyzed more closely to determine the extent of possible amino acid bias directly tied to viral replication mechanisms.

ACKNOWLEDGMENT

We thank William Turkett, Jr., Will Laegreid, John Keele, and Jiangying Zhou for many fruitful discussions.

REFERENCES

- [1] W. J. Conover, *Practical Nonparametric Statistics*, 3rd ed., John Wiley & Sons, 1998.
- [2] P. Dalgaard, *Introductory Statistics with R*. Springer-Verlag, New York, 2002.
- [3] B. Efron, "Estimating the error rate of a prediction rule: Improvement on cross-validation," *J. Am. Stat. Assoc.*, vol.78, pp. 316 331, 1983.
- [4] B. Efron and R. Tibshirani, *An introduction to the bootstrap*. Chapman & Hall/CRC, 1994.
- [5] R. Ihaka and R. Gentleman, "R: A language for data analysis and graphics," *J. Comput. Graph. Stat.*, vol. 5, pp. 299 314, 1996.
- [6] T. Joachims, "Learning to Classify Text Using Support Vector Machines," *Dissertation*, Kluwer, 2002.
- [7] T. Joachims, (1999) "Making large-Scale SVM Learning Practical," in *Advances in Kernel Methods - Support Vector Learning*, Schölkopf,B., Burges,C., and Smola,A. Eds., MIT Press, pp. 169 185, 1999.
- [8] W. Kabsch and C. Sander, "Dictionary of Protein Secondary Structure: Pattern Recognition and Hydrogen-Bonded and Geometrical Features," *Biopolymers*, vol. 22, pp. 2577 2637, 1983.
- [9] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, Mellish,C.S. Ed., Morgan Kaufmann, San Mateo, CA, pp. 1137 1143, 1995.
- [10] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion analysis of variance," *J. Am. Stat. Assoc.*, vol. 47, pp. 583 621, 1952.
- [11] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied Linear statistical Models*, McGraw Hill, 1996.
- [12] J. R. Rose, W. H. Turkett, I. C. Oroian, W. W. Laegreid, and J. Keele, (2004) "Correlation of Amino Acid Preference and Mammalian Viral Genome Type," *Bioinformatics*.vol. 21, no. 8, pp. 1349 1357, 2005.
- [13] M. Stone, "Cross-validation choice and assessment of statistical predictions," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 36, pp. 111 147, 1974.
- [14] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.
- [15] J. H. Zar, *Biostatistical Analysis*. Prentice Hall, New Jersey, 1998.
- [16] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I. N. Shindalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235 242, 2000.