LA-UR-00-2303

Approved for public release; distribution is unlimited.

Title:	Linguistic Analysis of the Nucleoprotein Gene of Influenza A Virus
Author(s):	Alexei Skourikhine and Tom Burr
Submitted to:	IEEE Conference on Bio-Informatic & Biomedical Engineering (BIBE 2000), Washington, D.C., November 5-8, 2000

Los Alamos
NATIONAL LABORATORY

Form 836 (10/96)

DISCLAIMER

Linguistic analysis of the nucleoprotein gene of influenza A virus

Alexei N. Skourikhine[a], Tom Burr[b]
Safeguards Systems Group, MS E541, Los Alamos National Laboratory
Los Alamos, NM 87545, USA
E-mails: [a] alexei@lanl.gov
[b] tburr@lanl.gov

Abstract

We applied linguistic analysis approach, specifically N-grams, to classify nucleotide and amino acids sequences of nucleoprotein (NP) gene of the Influenza A virus isolated from a range of hosts and geographic regions. We considered letter frequency (1-grams), letter pairs frequency (2-grams) and triplets' frequency (3-grams). Classification trees based on 1,2,3-grams variables were constructed for the same NP nucleotide and amino acids strains and their classification efficiency were compared with the clustering obtained using phylogenetic analysis. The results have shown that disregarding positional information for a NP gene can provide the same level of recognition accuracy like alternative more complex classification techniques.

Problem Context

Statistics has been used in linguistic analysis for many years. One of the approaches to uniquely identify and recognize a string of characters is to use a string itself. However, for long genetic sequences it leads to inefficiency of sequence characterization due to lack of capability of extraction of aggregate significant features reflecting underlying sequence structure. The other approach used in linguistic analysis is N-grams that are applied to classify documents. N-grams are sequences of N consecutive letters from a document. In a given problem of the classification of NP genes these letters are either nucleotides or amino acids. We applied N-grams approach to differentiate hosts of the influenza virus. 1-grams are the nucleotide bases and amino acids frequencies for nucleotide and protein sequences correspondingly; 2-grams are frequencies of the nucleotide and amino acids pairs; 3-grams are frequencies of their triplets. Thus, in the case of 1-gram for the set of amino acid strains we calculated vectors consisting of 20 entries each; for the set of nucleotide sequences we calculated vectors of 4 entries, where every entry corresponds either to amino acid or nucleotide. The advantage of using N-grams in our context is its ability to ignore positional information and represent global descriptive measure about a strain as a whole regardless of what amino acids and nucleotides take specific position within the NP strain. A method can be used with different kind of information also. E.g., instead of counting letter frequency we can replace the character string by physic-chemical properties corresponding to each base within a genetic sequence and count number of occurrences of these properties. In any case the input for a classification is a set of vectors corresponding to original sequences and containing aggregate statistical variables, such as frequency of specific properties (characters, chemical properties).

Results

128 sequences (nucleotide and amino acid) of nucleoprotein (NP) gene of the influenza A virus isolated from a range of hosts and geographic regions were analyzed to determine the most significant features which are potentially useful for classification of NP strains between different hosts. We consider species groupings (human, swine, or avian) of 1565 base pairs. The data is available from the influenza database maintained at Los Alamos National Laboratory (http://linker.lanl.gov/flu/search_frame.html).

The research employed contingency table analysis, specifically chi-square statistics and mutual information. We also applied N-grams approach to the classification of NP strains. We considered only letter frequency (1-grams). Subsets of 201 positions out of 498 amino acids positions and 768 positions out of 1565 nucleotide positions were identified as features that might be useful for separation between different hosts. Classification trees based on

The research employed contingency table analysis, specifically chi-square statistics and mutual information. We also applied N-grams approach to the classification of NP strains. We considered only letter frequency (1-grams). Subsets of 201 positions out of 498 amino acids positions and 768 positions out of 1565 nucleotide positions were identified as features that might be useful for separation between different hosts. Classification trees based on 1,2,3-grams variables were constructed for nucleotide and amino acids strains and their classification efficiency was compared.

	Number of entries
Avian	43
Human	57
Swine	28
Total	128

Table 1. Number of data entries for different hosts.

	Number of entries
Avian	43
Human	57
Swine	28
Total	128

Table. 2. Number of data entries used for testing.

The data used in the study is shown in the table shown in the Table 1. Number of data entries used for testing the classification accuracy is shown in the table 2, the rest of sequences was used for building the classification trees.

The median values of the calculated 1-grams of the NP amino acid sequences is shown in the Fig. 1 and 1,2,3-grams median values of the NP nucleotide sequences are shown in Figs. 2-4. Based on the 1,2-grams for amino acids and 1,2,3-grams for nucleotide, classification trees identifying virus hosts were created. The resulted misclassification rates for the extracted trees are shown in the Table 3. Two best trees corresponding to amino acids and nucleotide sequences are shown in Figs. 5 and 6.

The resulted misclassification rates for the nucleotide classification tree is 0.2074 = 28/135, the resulted misclassification rate for the amino acid classification tree is 0.1242 = 20/161.

Explain how classification was done and what plots mean..., include 14 wrong host cases.

	Nucleotides	Amino acids
1-gram		
2-gram		
3-gram		

Table 3. Misclassification rates.

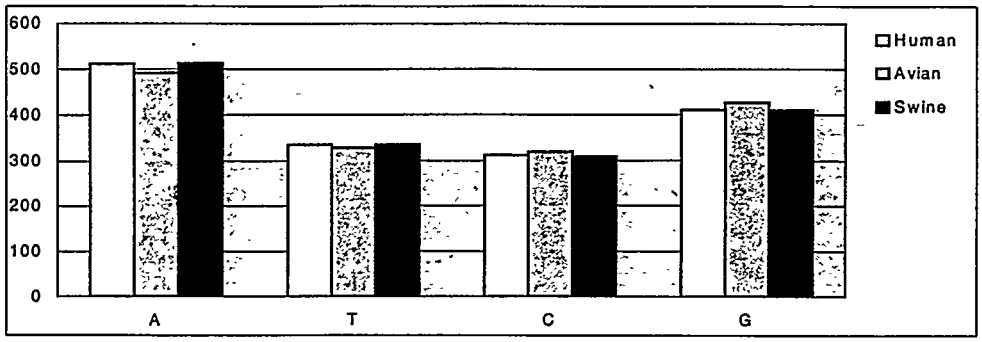


Fig. 1. Median values of 1-grams of NP nucleotide sequences.

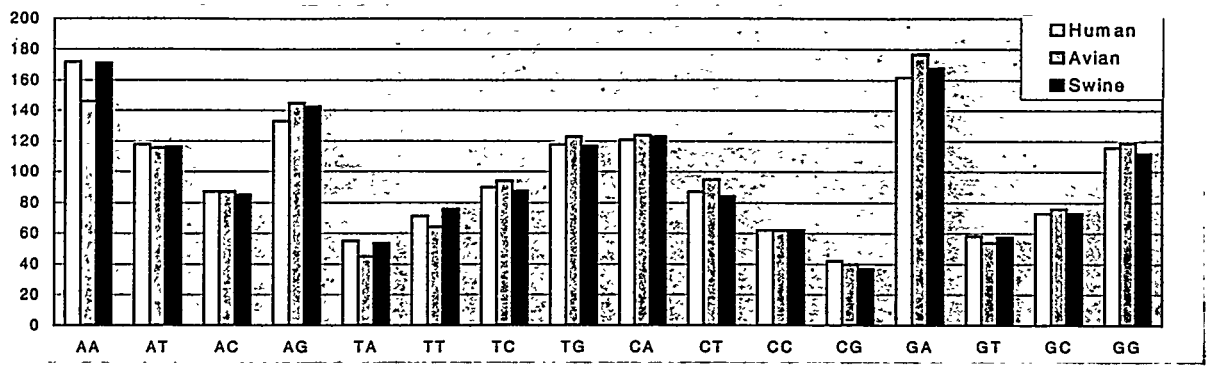


Fig. 2. Median values of 2-grams of NP nucleotide sequences.

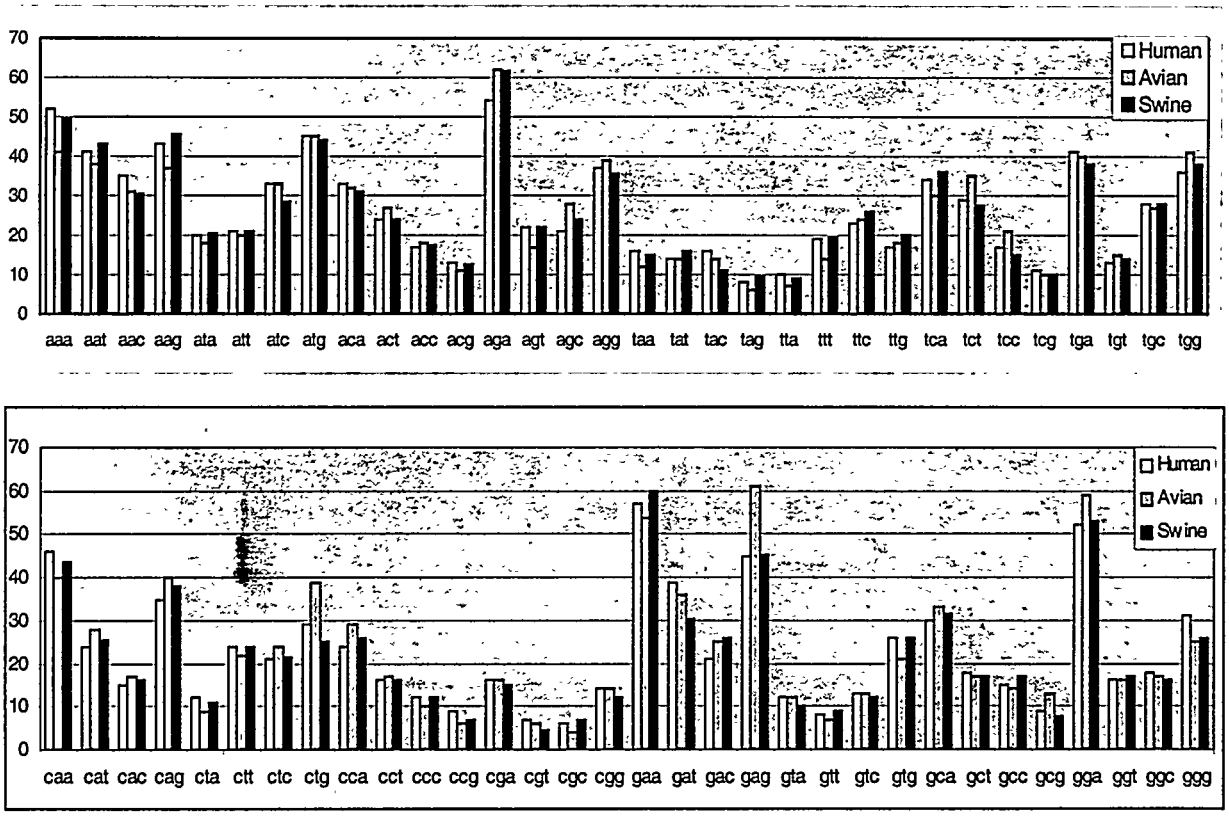
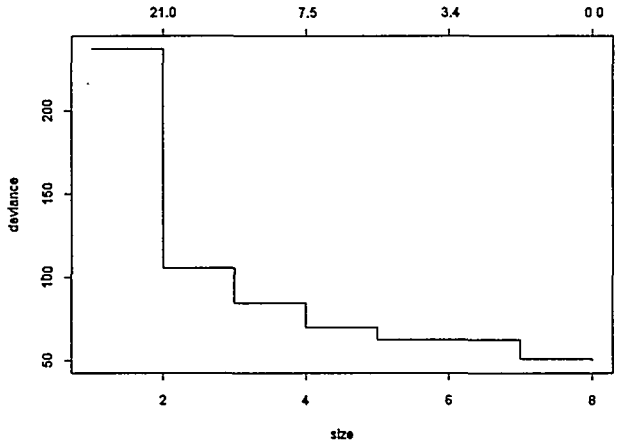


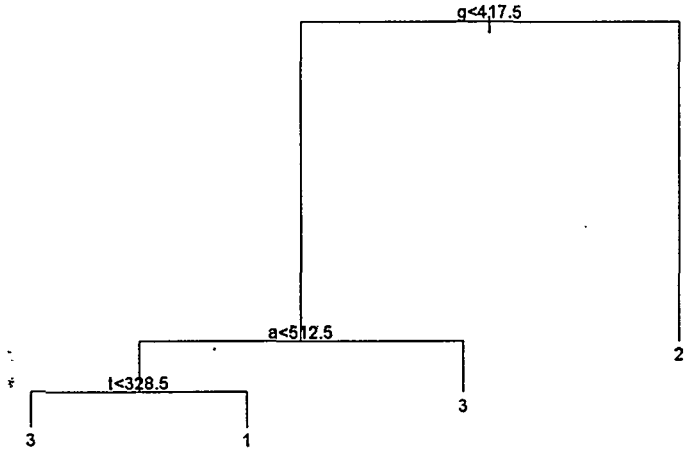
Fig. 3. Median values of 3-grams of NP nucleotide sequences.

3 figures here:
1: cross validation to choose good number of terminal nodes (size of tree) for 1-grams (128 by 4 with 14 cross-species cases OMITTED)
2: tree associated with same data as (1)
3: tree for 3-grams

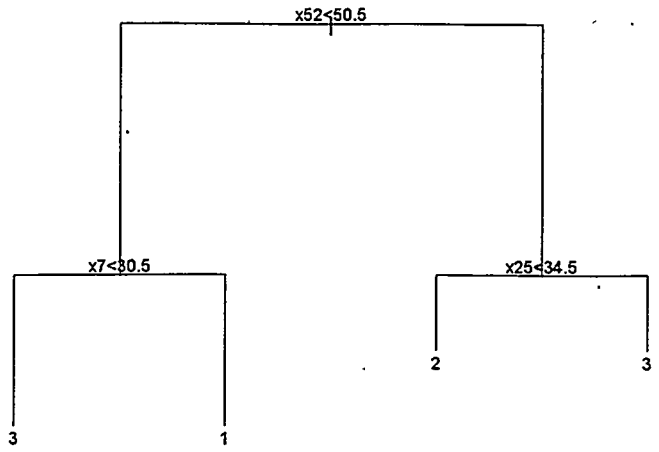
Example cross validation (to choose the number of terminal nodes) result applied to 1-grams, nucleotide sequence.



Example decision tree with 13/114 misclassified:



Example tree with 64 classifiers: (variable x1, x2, ... are aa, at, etc. in the order specified by first author).



Conclusions

We have shown that the classification accuracy achieved by utilizing N-grams information is comparable to the grouping obtained by phylogenetic tree reconstruction methods based on maximum likelihood approach. Thus, for NP gene of influenza A utilizing global characterization of a genetic sequence in the form of base frequencies leads to acceptable good classification. Future research will be concerned with the employment of information theory to extract global sequence features and possibly to extract significant sites within the sequence. The analysis is also expected to be applied to HA gene of the influenza.

References

[1] Dunn, O.J., Clark, V.A., 1974, *Applied Statistics: Analysis of variance and regression*, 1974.

[2] Swofford et. al., "Phylogeny Inference," Molecular Systematics, second edition, eds Hillis et. al., 1996.

[3] Felsenstein, J., X "Phylogenies from Molecular Sequences: Inference and Reliability," Annual Review of Genetics, **22**, 521-565, 1997.

[4] Burr, T., Skourikhine, A.N., Macken, C., Bruno, W., "Confidence Measures for Evolutionary Trees: Applications to Molecular Epidemiology," *Proc. of the 1999 IEEE Intern. Conference on Information, Intelligence and Systems*, pp.107-114, 1999.