Comparing Relative Frequency Distributions of Amino Acids and Nucleic Acids over DNA, RNA, and Protein sequences for *Vitis vinifera* and *Nicotiana tabacum*

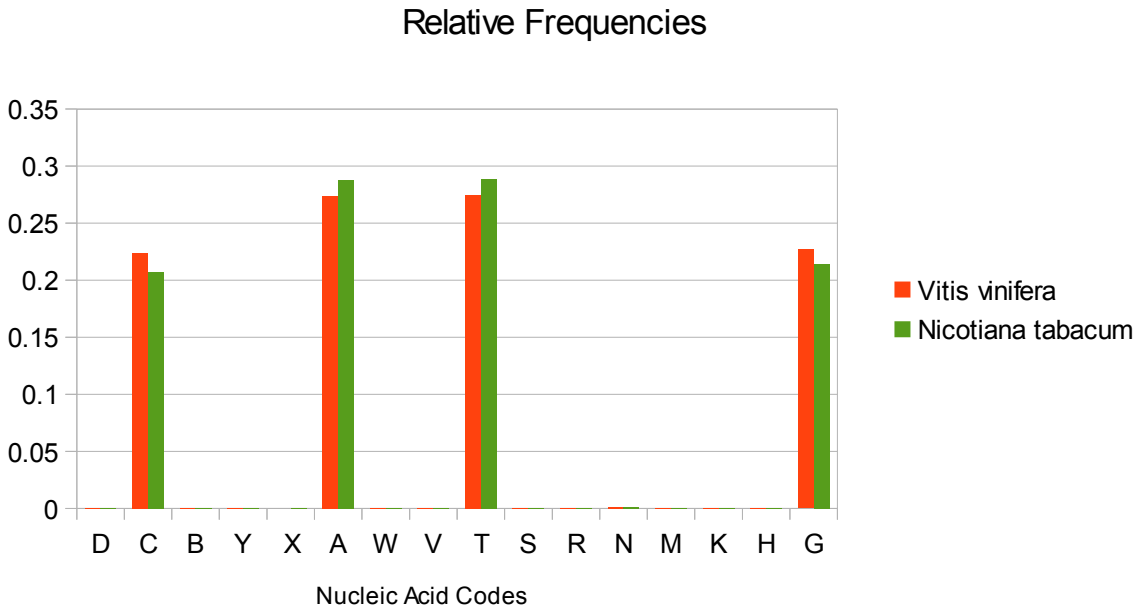
Brian J. Walters

Methodology

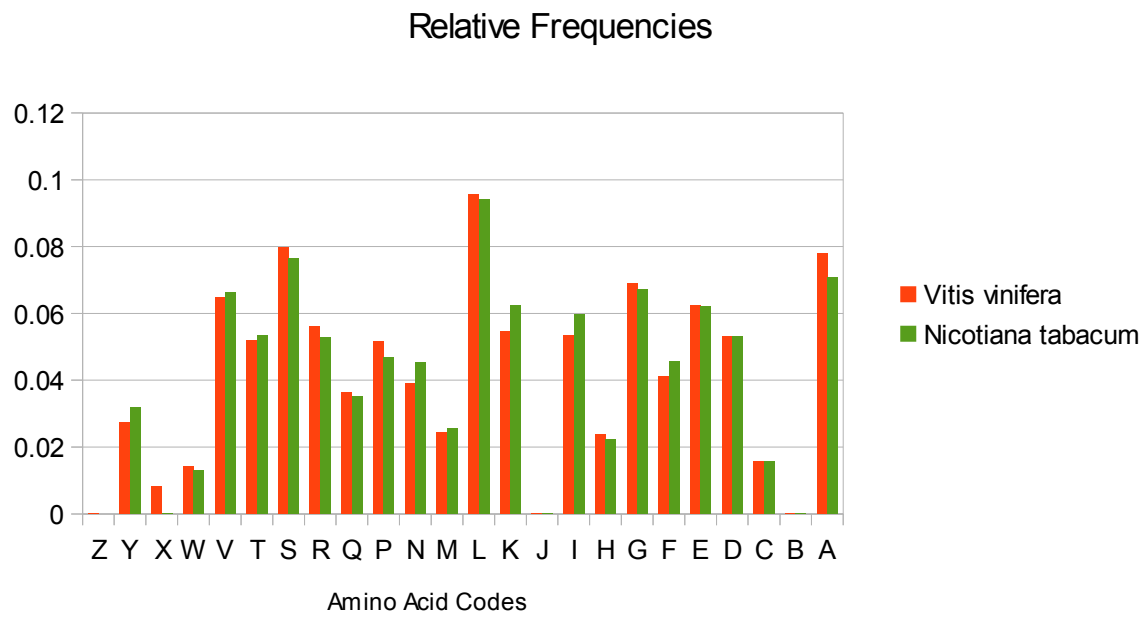
I selected common tobacco (*Nicotianan tabacum*) and the wine grape (*Vitis vinifera*) as two species on which to run relative frequency distributions. I acquired FASTA files for *Vitis vinifera* and *Nicotiana tabacum* from NCBI (http://www.ncbi.nlm.nih.gov/). The FASTA files were run through a Java program designed to load sequences from FASTA files, determine whether they were DNA, RNA, or Protein sequences, aggregate the total number of nucleic or amino acid codes (respectively), and calculate the relative frequency for each acid code.

Data

Nucleic Acid Codes	Vitis vinifera		Nicotiana tabacum	
	Total	Relative Frequency	Total	Relative Frequency
D	12	0.000004222%	5	0.000002585%
C	63612620	22.381567383%	40045392	20.706496361%
B	3	0.000001056%	3	0.000001551%
Y	61	0.000021462%	51	0.000026371%
X	0	0	1	0.000000517%
A	77892373	27.405778837%	55727655	28.815412407%
W	72	0.000025333%	26	0.000013444%
V	4	0.000001407%	7	0.000003620%
T	77952590	27.426965684%	55934119	28.922169910%
S	34	0.000011963%	42	0.000021717%
R	88	0.000030962%	55	0.000028439%
N	274966	0.096744483%	244801	0.126580632%
M	68	0.000023925%	39	0.000020166%
K	55	0.000019351%	37	0.000019132%
H	6	0.000002111%	4	0.000002068%
G	64485838	22.688801821%	41443069	21.429201079%
	284218790	100.000000000%	193395306	100.000000000%



Amino Acid Codes	Vitis vinifera Total	Vitis vinifera Relative Frequency	Nicotiana tabacum Total	Nicotiana tabacum Relative Frequency
Z	498	0.000890218%	0	0.000000000%
Y	1537892	2.749114479%	113393	3.176343909%
X	456966	0.816866104%	377	0.010560455%
W	791383	1.414665311%	46172	1.293361592%
V	3625897	6.481603352%	236391	6.621741315%
T	2897859	5.180172688%	190814	5.345046755%
S	4452926	7.959988959%	272648	7.637365746%
R	3142104	5.616781673%	188665	5.284849361%
Q	2030267	3.629277222%	125755	3.522625985%
P	2885628	5.158308721%	166884	4.674723986%
N	2186527	3.908605438%	161254	4.517017459%
M	1355048	2.422265072%	91525	2.563781506%
L	5351133	9.565611375%	336375	9.422474777%
K	3062377	5.474262790%	222275	6.226326514%
J	604	0.001079702%	1	0.000028012%
I	2983617	5.333472503%	213339	5.976012921%
H	1321907	2.363022679%	79799	2.235314945%
G	3860906	6.901702191%	240215	6.728858502%
F	2304089	4.118757644%	163185	4.571108276%
E	3483020	6.226198402%	222066	6.220472044%
D	2972388	5.313399698%	189744	5.315074111%
C	875326	1.564720657%	56253	1.575748714%
B	1673	0.002990632%	2	0.000056024%
A	4361324	7.796242490%	252790	7.081107094%
	55941359	100.00%	3569922	100.00%



Results

Both species primarily contain large amounts of nucleic acid codes C (Cytosine), A (Adenine), T (Thymine), and G (Guanine). Also for both species, A and T appear slightly more frequently than C and G. Nucleic acids C and G appear in *Vitis vinifera* with a slightly greater frequency, while nucleic acids A and T are more frequent in *Nicotiana tabacum*.

Both species contained relatively low amounts of amino acid codes for Y, W, M, H, and C. There were very low amounts of Z, J, and B. Encoding X ("any") had almost no occurrences in *Nicotiana tabacum*, but a slightly higher amount in *Vitis vinifera*.

For amino acids, the relative frequency distributions are also comparable. The relative frequencies for *Vitis vinifera* is notably higher for amino acid codes S (Serine), P (Proline), and A (Alanine) than *Nicotiana tabacum*, while the latter has a notably higher frequency of amino acid codes N (Asparagine), K (Lysine), I (Isoleucine), and F (Phenylalanine). There are a number of X ("any") encodings for *Vitis vinifera*, while almost none are present in *Nicotiana tabacum*.