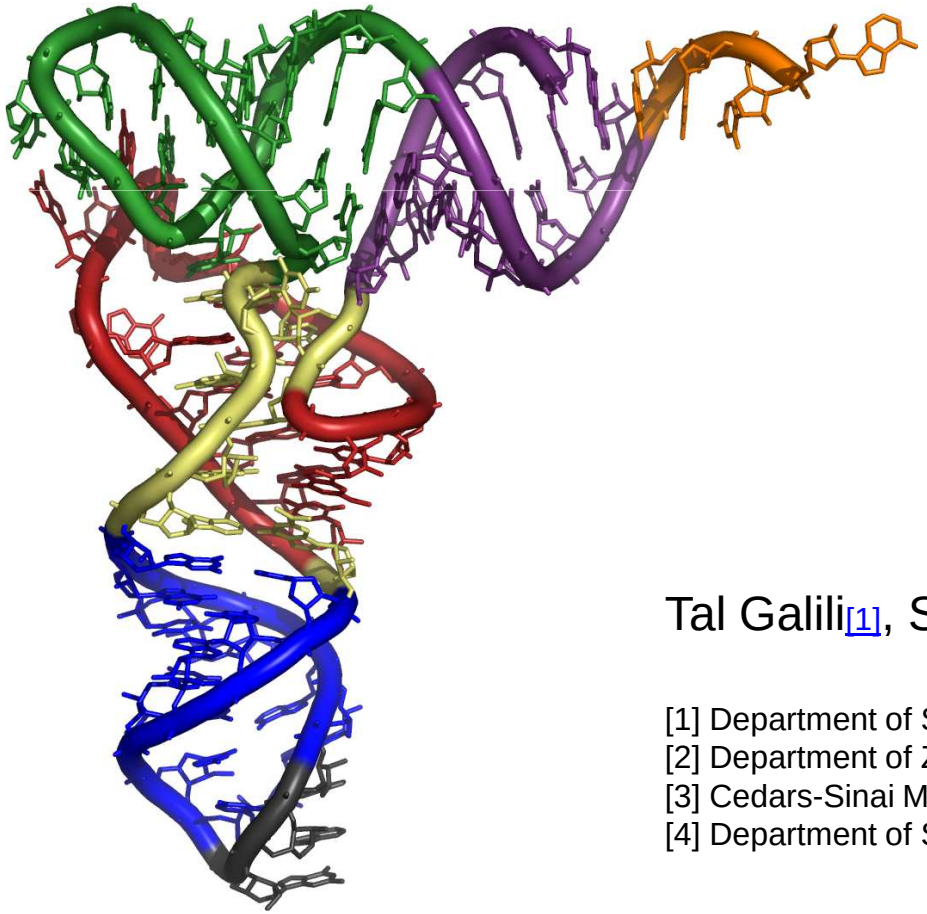
Analyzing the Operational RNA Code for Amino Acids - Using R



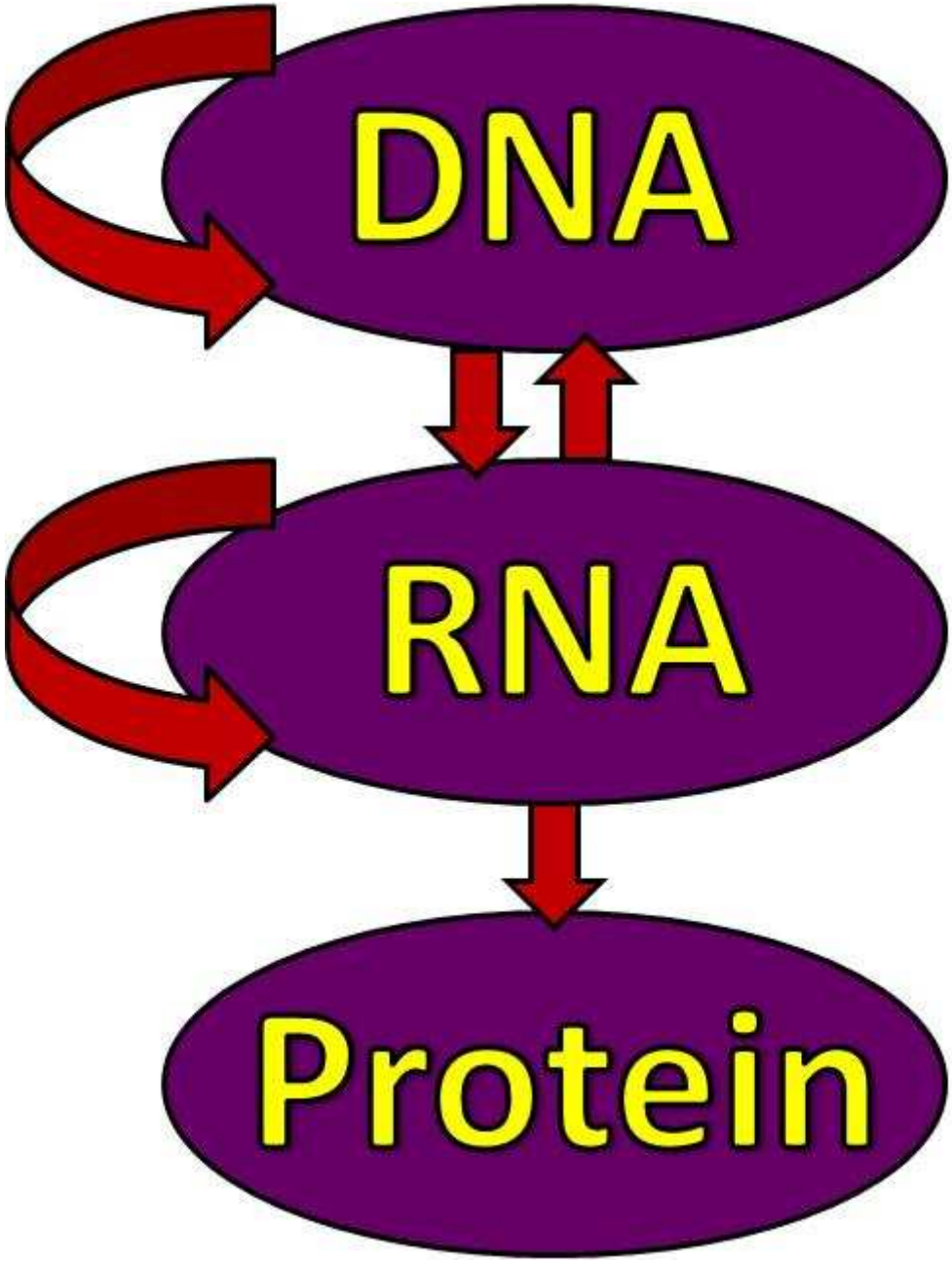
Read / Contact me:

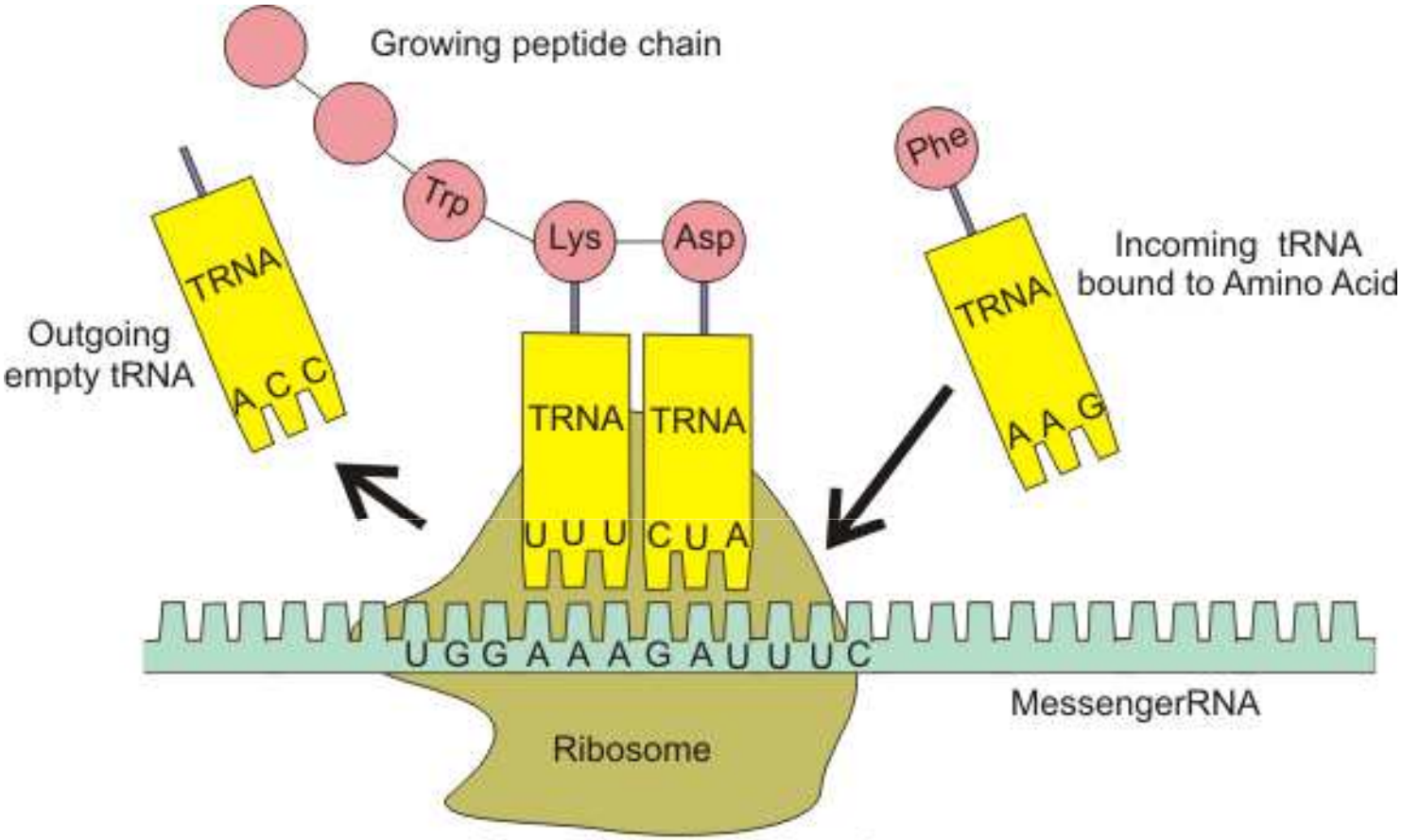
R-statistics.com

Tal Galili[1], Shaul Shaul[2], Dror Berel [3], Yoav Benjamini [4]

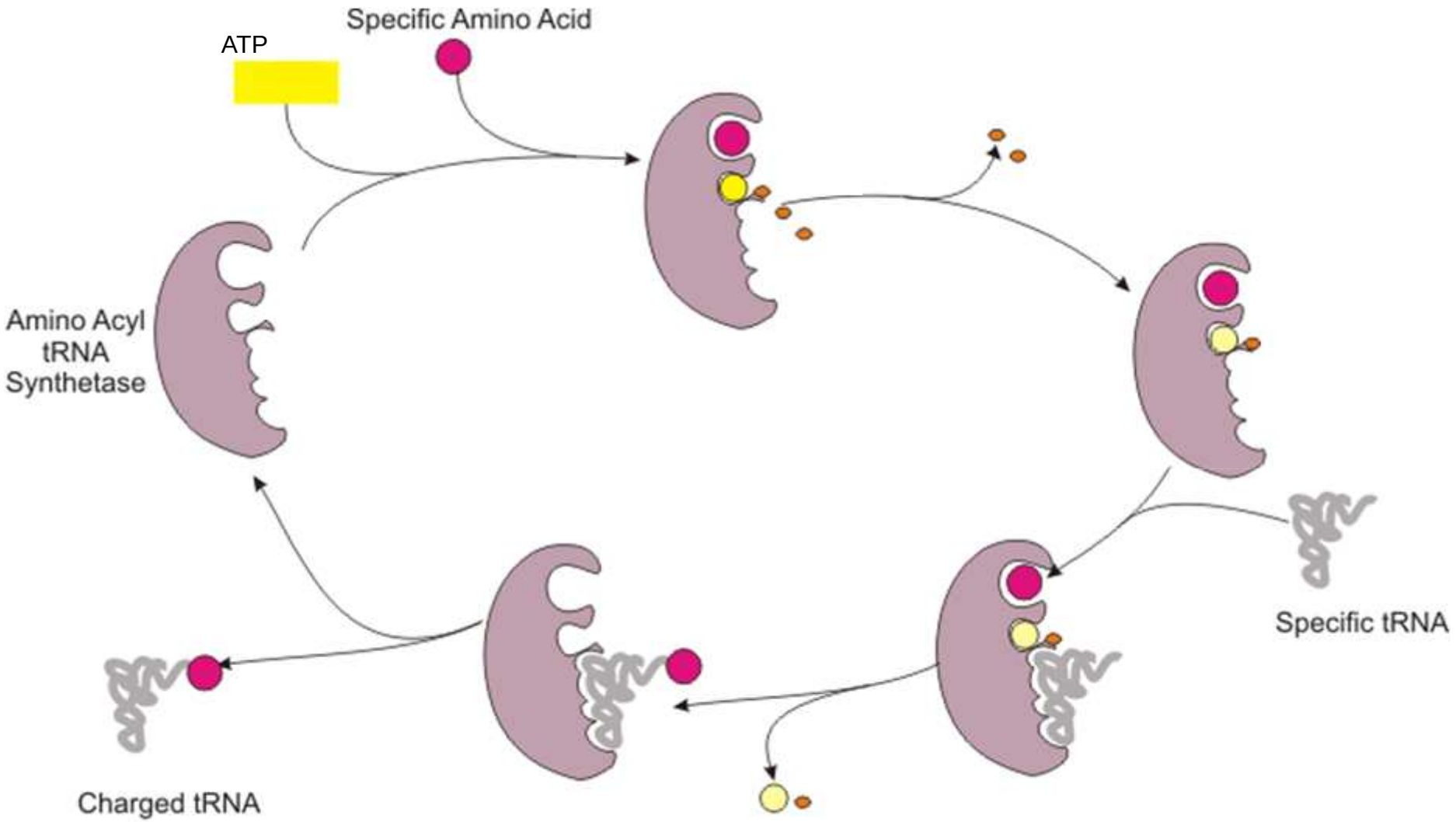
[1] Department of Statistics and Operations Research, Tel Aviv University, Israel.
[2] Department of Zoology, Tel Aviv University, Israel.
[3] Cedars-Sinai Medical Center.
[4] Department of Statistics and Operations Research, Tel Aviv University, Israel.

Story


DNA
RNA
Protein

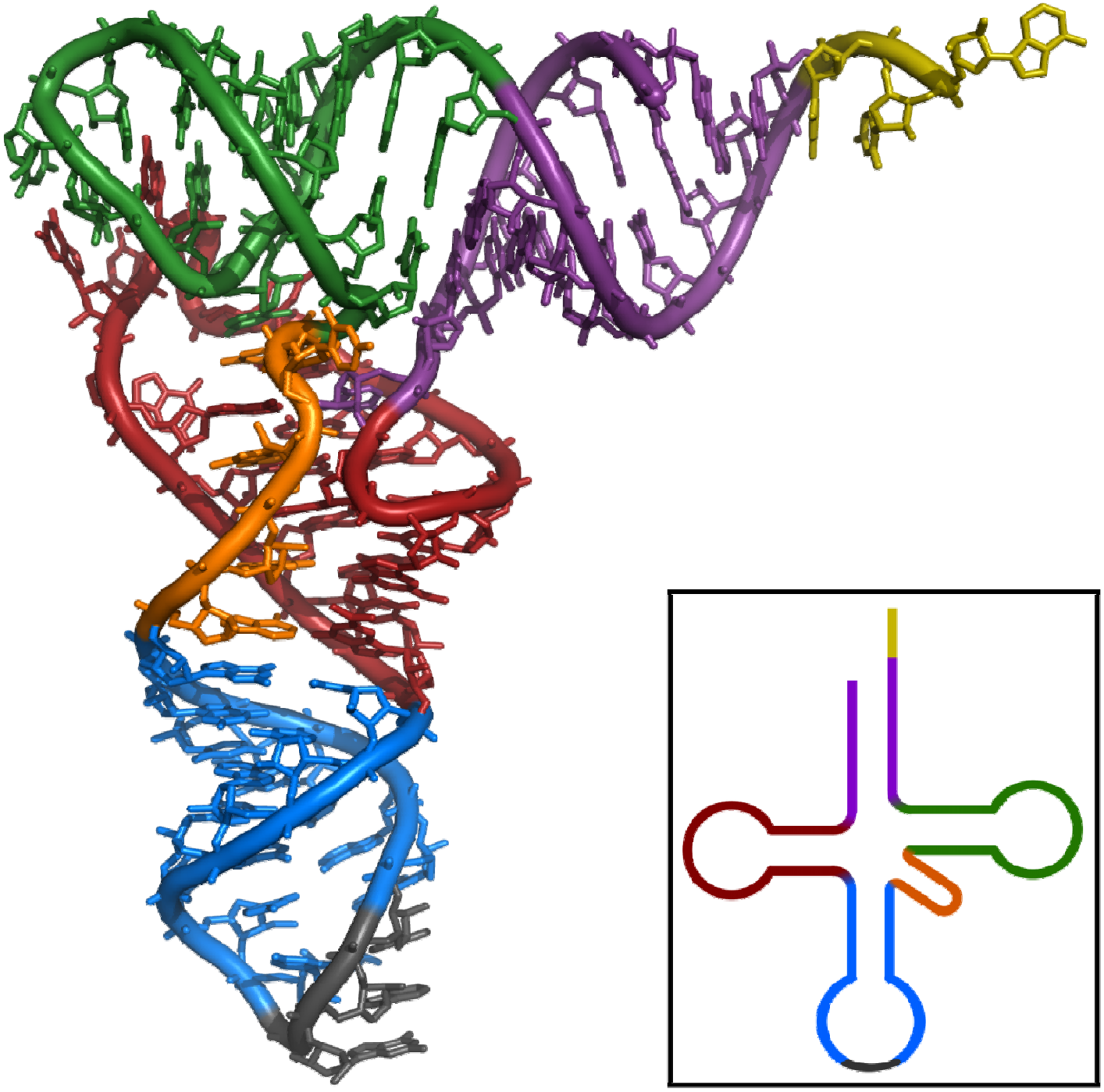


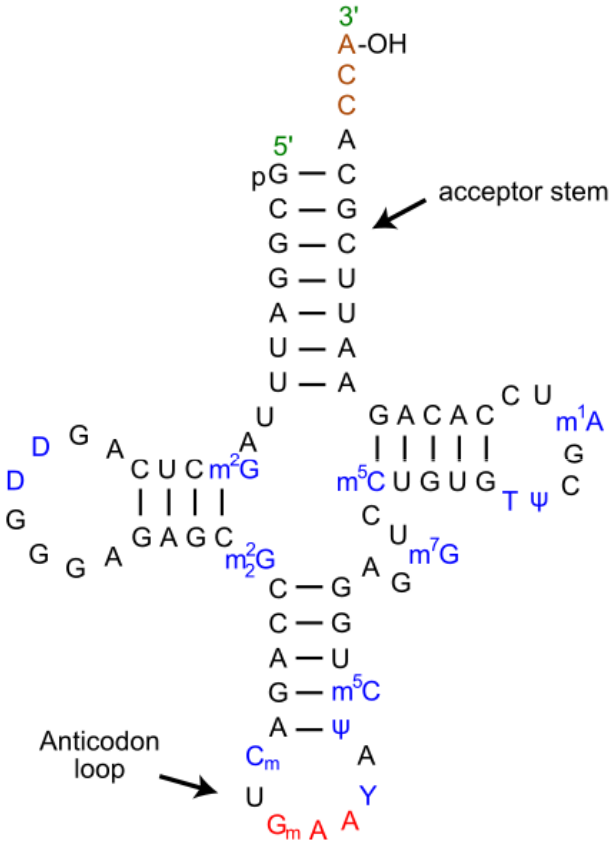
Peptide Synthesis



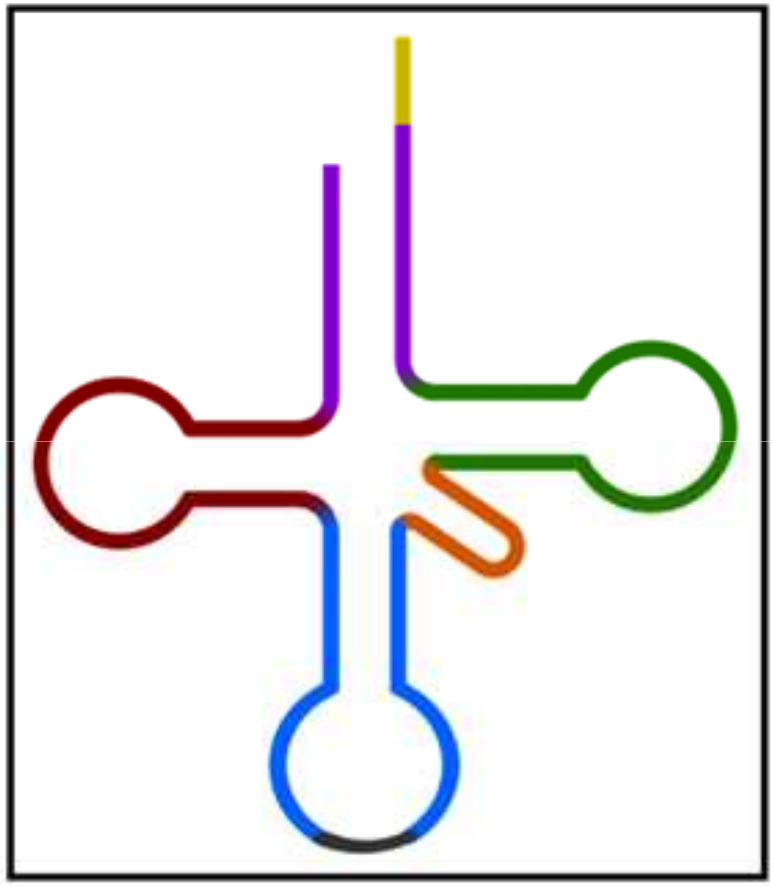
Loading a tRNA molecule

Frank Boumphrey M.D.
2009

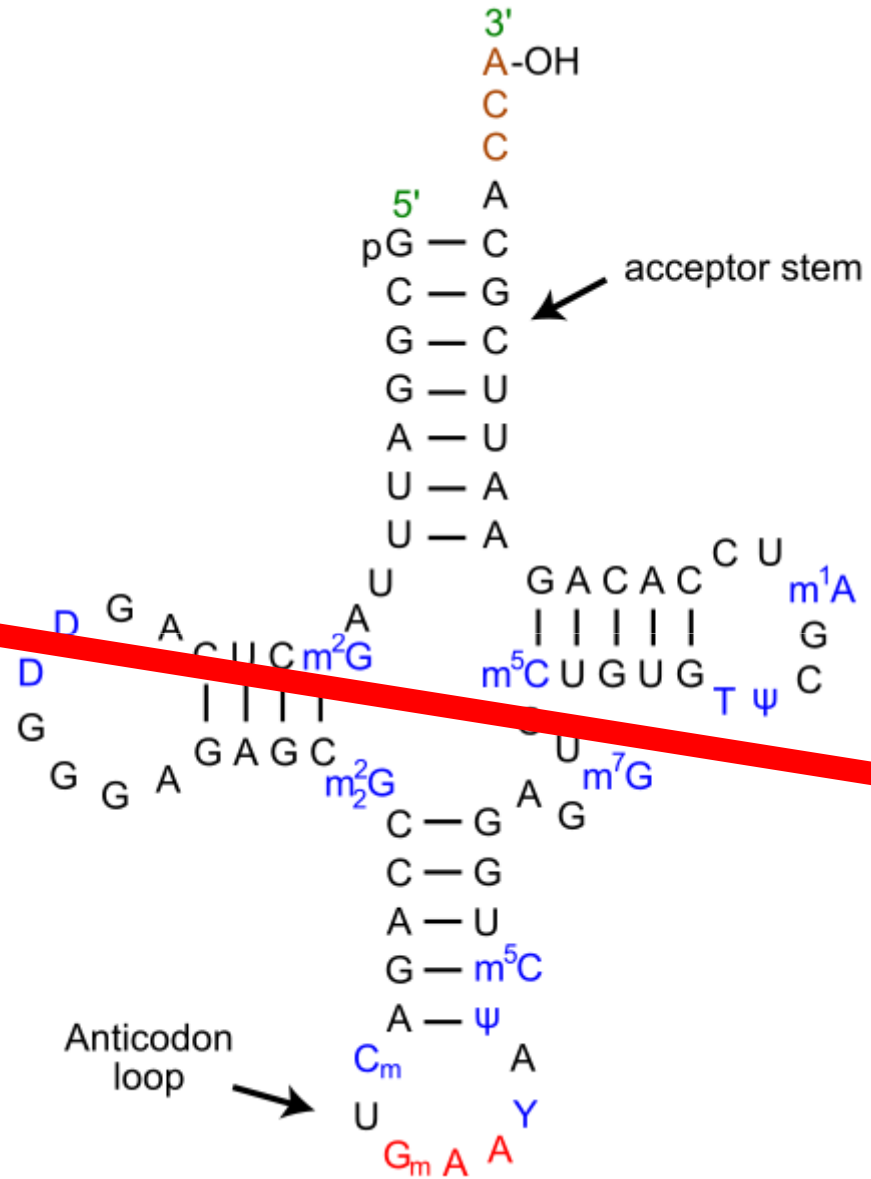



3'
A-OH
C
C
A
5'
pG — C
C — G
G — C
G — U
A — U
U — A
U — A
acceptor stem
U
G A C A C
C U
m1A
D G A
A
C U C m2G
m5C U G U G
G
D
T ψ
C
G
G A G C
C
U
m7G
G G A
m22G
A G
C — G
C — G
A — U
G — m5C
A — ψ
Anticodon loop
Cm
A
U
Y
Gm A A

=



Christopher Francklyn
and Paul Shimmel

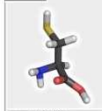


standard Amino Acids

A L-Alanine ALA

Chemical Formula	$C_3H_7NO_2$
Molecular Weight	89.09 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.616
Hydropathy Index	1.8
Isoelectric Point	6.01
pKa (α-COOH)	2.35
pKa (α-NH2)	9.87
Canonical SMILES	CC(N)C(=O)O
CAS Number	56-41-7
PubChem CID	5950

C L-Cysteine CYS



Chemical Formula	$C_3H_7NO_2S$
Molecular Weight	121.16 Da
Polarity	Polar
Acidity/ Basicity	Neutral
Hydrophobicity	0.680
Hydropathy Index	2.5
Isoelectric Point	5.05
pKa (α-COOH)	1.92
pKa (α-NH2)	10.70
Canonical SMILES	NC(C(=O)O)CS
CAS Number	52-90-4
PubChem CID	5862

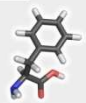
D L-Aspartic Acid ASP

Chemical Formula	$C_4H_7NO_4$
Molecular Weight	133.10 Da
Polarity	Polar
Acidity/ Basicity	Acidic
Hydrophobicity	0.028
Hydropathy Index	-3.5
Isoelectric Point	2.85
pKa (α-COOH)	1.99
pKa (α-NH2)	9.90
Canonical SMILES	NC(CC(=O)O)C(=O)O
CAS Number	56-84-8
PubChem CID	5960

E L-Glutamic Acid GLU

Chemical Formula	$C_5H_9NO_4$
Molecular Weight	147.13 Da
Polarity	Polar
Acidity/ Basicity	Acidic
Hydrophobicity	0.043
Hydropathy Index	-3.5
Isoelectric Point	3.15
pKa (α-COOH)	2.10
pKa (α-NH2)	9.47
Canonical SMILES	NC(CCC(=O)O)C(=O)O
CAS Number	56-86-0
PubChem CID	611

F L-Phenylalanine PHE

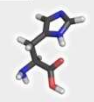


Chemical Formula	$C_9H_{11}NO_2$
Molecular Weight	165.19 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	1.00
Hydropathy Index	2.8
Isoelectric Point	5.49
pKa (α-COOH)	2.20
pKa (α-NH2)	9.31
Canonical SMILES	NC(Cc1ccccc1)C(=O)O
CAS Number	63-91-2
PubChem CID	994

G Glycine GLY

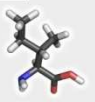
Chemical Formula	$C_2H_5NO_2$
Molecular Weight	75.07 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.501
Hydropathy Index	-0.4
Isoelectric Point	6.06
pKa (α-COOH)	2.35
pKa (α-NH2)	9.78
Canonical SMILES	NCC(=O)O
CAS Number	56-40-6
PubChem CID	750

H L-Histidine HIS



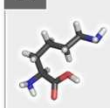
Chemical Formula	$C_6H_9N_3O_2$
Molecular Weight	155.15 Da
Polarity	Polar
Acidity/ Basicity	Basic (Weak)
Hydrophobicity	0.165
Hydropathy Index	-3.2
Isoelectric Point	7.60
pKa (α-COOH)	1.80
pKa (α-NH2)	9.33
Canonical SMILES	NC(Cc1cnc[nH]1)C(=O)O
CAS Number	71-00-1
PubChem CID	773

I L-Isoleucine ILE



Chemical Formula	$C_6H_{13}NO_2$
Molecular Weight	131.17 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.943
Hydropathy Index	4.5
Isoelectric Point	6.05
pKa (α-COOH)	2.32
pKa (α-NH2)	9.76
Canonical SMILES	CCC(C)C(N)C(=O)O
CAS Number	73-32-5
PubChem CID	791

K L-Lysine LYS

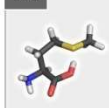


Chemical Formula	$C_6H_{14}N_2O_2$
Molecular Weight	146.19 Da
Polarity	Polar
Acidity/ Basicity	Basic
Hydrophobicity	0.283
Hydropathy Index	-3.9
Isoelectric Point	9.60
pKa (α-COOH)	2.16
pKa (α-NH2)	9.06
Canonical SMILES	NCCCCC(N)C(=O)O
CAS Number	56-87-1
PubChem CID	866

L L-Leucine LEU

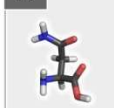
Chemical Formula	$C_6H_{13}NO_2$
Molecular Weight	131.18 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.943
Hydropathy Index	3.8
Isoelectric Point	6.01
pKa (α-COOH)	2.33
pKa (α-NH2)	9.74
Canonical SMILES	CC(C)CC(N)C(=O)O
CAS Number	61-90-5
PubChem CID	6106

M L-Methionine MET



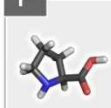
Chemical Formula	$C_5H_{11}NO_2S$
Molecular Weight	149.21 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.738
Hydropathy Index	1.9
Isoelectric Point	5.74
pKa (α-COOH)	2.13
pKa (α-NH2)	9.28
Canonical SMILES	CSCCC(N)C(=O)O
CAS Number	63-68-3
PubChem CID	876

N L-Asparagine ASN



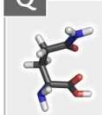
Chemical Formula	$C_4H_8N_2O_3$
Molecular Weight	132.12 Da
Polarity	Polar
Acidity/ Basicity	Neutral
Hydrophobicity	0.236
Hydropathy Index	-3.5
Isoelectric Point	5.41
pKa (α-COOH)	2.14
pKa (α-NH2)	8.72
Canonical SMILES	NC(CC(=O)N)C(=O)O
CAS Number	70-47-3
PubChem CID	236

P L-Proline PRO



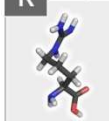
Chemical Formula	$C_5H_9NO_2$
Molecular Weight	115.13 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.711
Hydropathy Index	-1.6
Isoelectric Point	6.30
pKa (α-COOH)	1.95
pKa (α-NH2)	10.64
Canonical SMILES	OC(=O)C1CCCN1
CAS Number	147-85-3
PubChem CID	614

Q L-Glutamine GLN



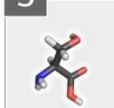
Chemical Formula	$C_5H_{10}N_2O_3$
Molecular Weight	146.14 Da
Polarity	Polar
Acidity/ Basicity	Neutral
Hydrophobicity	0.251
Hydropathy Index	-3.5
Isoelectric Point	5.65
pKa (α-COOH)	2.17
pKa (α-NH2)	9.13
Canonical SMILES	NC(CCC(=O)N)C(=O)O
CAS Number	56-85-9
PubChem CID	5950

R L-Arginine ARG



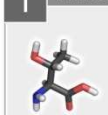
Chemical Formula	$C_6H_{14}N_4O_2$
Molecular Weight	174.20 Da
Polarity	Polar
Acidity/ Basicity	Basic (Strong)
Hydrophobicity	0.000
Hydropathy Index	-4.5
Isoelectric Point	10.76
pKa (α-COOH)	1.82
pKa (α-NH2)	8.99
Canonical SMILES	NC(CCCNC(=N)N)C(=O)O
CAS Number	74-79-3
PubChem CID	5950

S L-Serine SER



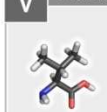
Chemical Formula	$C_3H_7NO_3$
Molecular Weight	105.09 Da
Polarity	Polar
Acidity/ Basicity	Neutral
Hydrophobicity	0.359
Hydropathy Index	-0.8
Isoelectric Point	5.68
pKa (α-COOH)	2.19
pKa (α-NH2)	9.21
Canonical SMILES	NC(CO)C(=O)O
CAS Number	56-45-1
PubChem CID	617

T L-Threonine THR



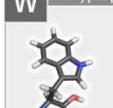
Chemical Formula	$C_4H_9NO_3$
Molecular Weight	119.12 Da
Polarity	Polar
Acidity/ Basicity	Neutral
Hydrophobicity	0.450
Hydropathy Index	-0.7
Isoelectric Point	5.60
pKa (α-COOH)	2.09
pKa (α-NH2)	9.10
Canonical SMILES	CC(O)C(N)C(=O)O
CAS Number	72-19-5
PubChem CID	6288

V L-Valine VAL



Chemical Formula	$C_5H_{11}NO_2$
Molecular Weight	117.15 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.825
Hydropathy Index	4.2
Isoelectric Point	6.00
pKa (α-COOH)	2.39
pKa (α-NH2)	9.74
Canonical SMILES	CC(C)C(N)C(=O)O
CAS Number	72-18-4
PubChem CID	1182

W L-Tryptophan TRP



Chemical Formula	$C_{11}H_{12}N_2O_2$
Molecular Weight	204.23 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.878
Hydropathy Index	-0.9
Isoelectric Point	5.89
pKa (α-COOH)	2.46
pKa (α-NH2)	9.41
Canonical SMILES	NC(Cc1c[nH]c2ccccc12)C(=O)O
CAS Number	73-22-3
PubChem CID	6305

Y L-Tyrosine TYR

Chemical Formula	$C_9H_{11}NO_3$
Molecular Weight	181.19 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.880
Hydropathy Index	-1.3
Isoelectric Point	5.64
pKa (α-COOH)	2.20
pKa (α-NH2)	9.21
Canonical SMILES	NC(Cc1ccc(O)cc1)C(=O)O
CAS Number	60-18-4
PubChem CID	1153

B L-Asparagine or L-Aspartic Acid ASX

J L-Leucine or L-Isoleucine XLE

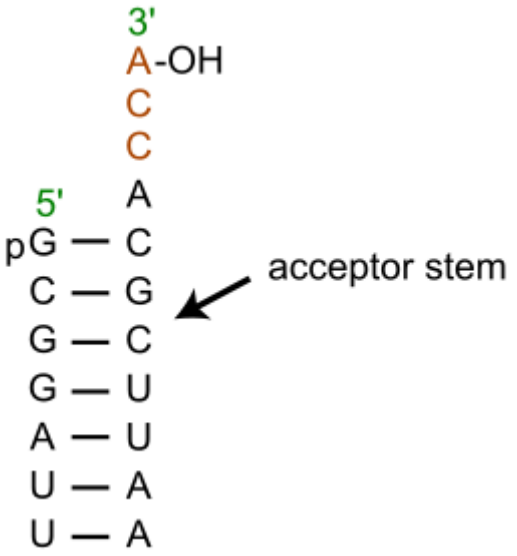
X Unspecified or Unknown XAA

Z L-Glutamine or L-Glutamic Acid GLX

Essential amino acids are denoted with yellow spots. Polarity refers to Side Chain Polarity. Acidity/Basicity refer to Side Chain Acidity/Basicity.

2D Structure Generation, 3D optimization: ChemSketch 8.17 (ACD Labs) | 3D Structure Rendering: PyMol 0.99rc6 (DeLano Scientific) | Chemical Formula and Molecular Weight Calculations: ChemSketch 8.17 (ACD Labs) | Hydrophobicity values: Analytical Biochemistry 193:11, 72-82, Elsevier, 1991. | Hydropathy Index values: J. Mol. Biol. (1982) 157, 105-132. | Canonical SMILES: www-jmg.ch.cam.ac.uk/data/molecules/smilesindex.html | Polarity, Acidity/Basicity, Isoelectric Point, pKa, CAS, PubChem CID data: en.wikipedia.org/wiki/List_of_standard_amino_acids | List of Essential Amino Acids: en.wikipedia.org/wiki/Amino_acid | Typefaces: Meta Book

20071219 Berk Zafer | MMI.unc.edu | Questions, problems, and errors should be sent to berkzafer@unc.edu
The most recent version of this document can be found at www.unc.edu/~bzafer/aminoacids



Data

Data source

http://gtrnadb.ucsc.edu/

GtRNAdb

Genomic tRNA Database

tRNAscan-SE analysis of complete genomes

[Home] [Find tRNAs] [BLAST] [Download] [Bug/Request] [FAQ] [Citation] [Legend]

Introduction

This genomic tRNA database contains tRNA gene predictions made by the program tRNAscan-SE (Lowe & Eddy, *Nucl Acids Res* **25**: 955-964, 1997) on complete or nearly co otherwise noted, all annotation is automated, and has not been inspected for agreement with published literature.

Inevitably with automated sequence analysis, we find exceptions to general identification rules, isoacceptor type predictions (*esp.* due to variable post-transcriptional anticodon modi tRNA identifications (due to pseudogenes, SINES, or other tRNA-derived elements). We attempt to document all cases we come across, and welcome feedback (lowe @soe.ucsc unrecognized discrepancies. For a more detailed description of information in tables and the tRNA search algorithm, see the [Legend].

Genomes

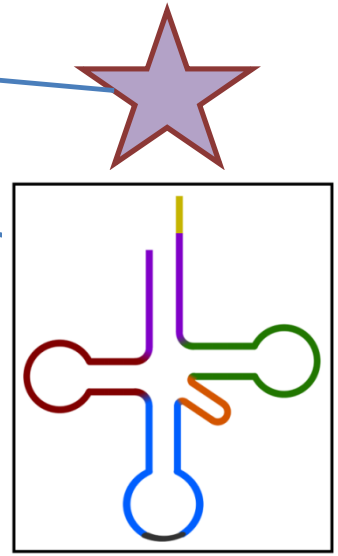
[Eukarya] [Archaea] [Bacteria]

Eukarya

[Echinozoa] [Embryophyta] [Fungi] [Haemosporida] [Insecta] [Leishmania] [Rhabditida] [Tracheophyta] [Vertebrata]

Echinozoa (1 genome)		
Strongylocentrotus purpuratus (Sea urchin) (Version 2.1) (1065 tRNAs)		
Embryophyta (3 genomes)		

```
>Aeropyrum_pernix_chr.trna3-AlaCGC (1-77)  Ala (CGC) 77 bp  Sc: 93.45
GGGCCGGTAGCTCAGCCTGGAAGAGCGCCGCCCTCGCACGGCGGAGGCCCCGGGTTCAAA
TCCCGGCCGGTCCACCA
>Aeropyrum_pernix_chr.trna39-AlaGGC (1-77)  Ala (GGC) 77 bp  Sc: 88.60
GGGCCGGTAGCTCAGCCTGGAAGAGCGCCCGGTTGGCATCCGGGAGGCCCCGGGTTCAAA
TCCCGGCCGGTCCACCA
>Aeropyrum_pernix_chr.trna27-AlaTGC (1-77)  Ala (TGC) 77 bp  Sc: 93.25
GGGCCGGTAGCTCAGCCTGGAAGAGCGCCCGGCTTGCACCCGGGAGGCCCCGGGTTCAAA
TCCCGGCCGGTCCACCA
>Aeropyrum_pernix_chr.trna18-ArgCCG (1-77)  Arg (CCG) 77 bp  Sc: 28.48
ATAAATGCAACACAGGTATTTATAGAGCGGGTCTCCGGAGCCGGAGGTCCCGGGTTCAAA
TCCCGGCGGGCCCGCCA
>Aeropyrum_pernix_chr.trna2-ArgCCG (1-72)  Arg (CCG) 72 bp  Sc: 24.36
TGCAACGTGAGTATTTATAGCGTCGTTCTCCGGAGCCGGAGGTCCCGGGTTCAAATCCCG
GCGGGCCCGCCA
>Aeropyrum_pernix_chr.trna1-ArgCCG (1-78)  Arg (CCG) 78 bp  Sc: 92.70
GGGCCCGTAGCTCAGCCAGGATAGAGCGCCGGCCTCCGGAGCCGGAGGTCCCGGGTTCAA
ATCCCGGCGGGCCCGCCA
>Aeropyrum_pernix_chr.trna26-ArgCCT (1-78)  Arg (CCT) 78 bp  Sc: 91.62
GGGCCCGTAGCTCAGCCAGGATAGAGCGCCGGCCTCCTAAGCCGGTGGTCGGGGGTTCAA
ATCCCCCCGGGCCCGCCA
>Aeropyrum_pernix_chr.trna9-ArgGCG (1-79)  Arg (GCG) 79 bp  Sc: 92.86
GGGCCCGTAGCTCAGCCAGGACCAGAGCGCCGGCCTGCGGAGCCGGAGGTCCCGGGTTCA
```

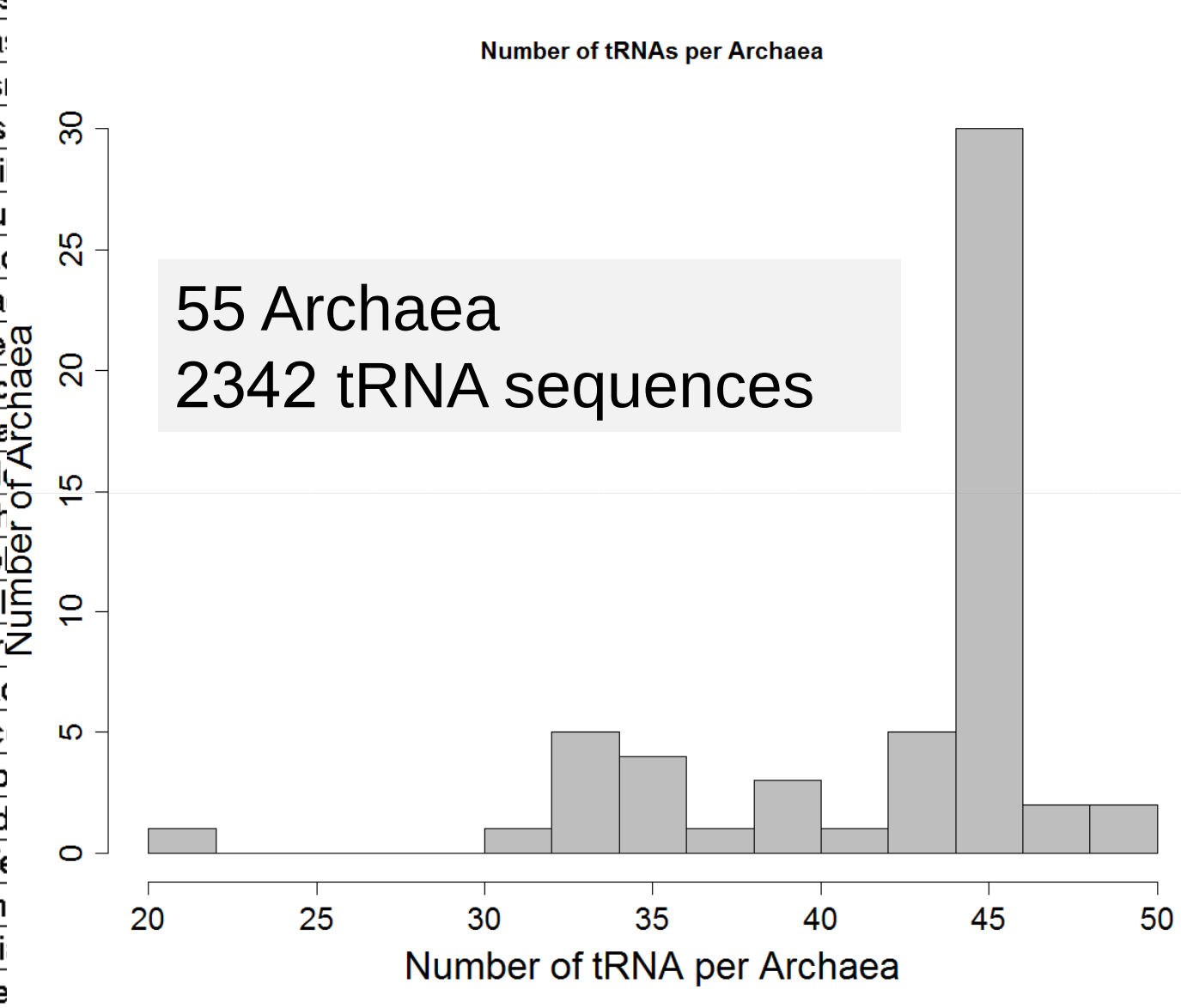


Species	N1	N2	N3	N4	N5	N6	N7	N73	N72	N71	N70	N69	N68	N67	N66	AA
Mycoplasma capricolum	G	G	C	A	A	C	A	U	C	C	G	U	U	G	U	C
Mycoplasma genitalium	G	G	U	G	C	C	U	U	C	C	A	C	G	G	A	C
Mycoplasma pneumoniae	G	G	U	G	C	C	U	U	C	C	A	C	G	G	A	C
Ureaplasma urealyticum	G	G	U	G	C	C	A	U	C	C	A	C	G	G	U	C
Bacillus subtilis	G	G	C	G	G	C	A	U	C	C	G	C	U	G	U	C
Bacillus halodurans	G	G	C	G	G	C	A	U	C	C	G	C	C	G	U	C
Bifidobacterium longum	G	G	U	G	G	G	U	U	C	C	A	C	C	C	A	C
Mycobacterium leprae	G	G	U	G	G	A	G	U	C	C	A	C	C	U	C	C
Mycobacterium tuberculosis	G	G	U	G	G	A	G	U	C	C	A	C	C	U	C	C
Synechocystis sp.	G	G	C	G	G	C	A	U	C	C	G	C	C	G	U	C
Helicobacter pylori J99	G	G	C	G	A	C	A	U	C	C	G	C	U	G	U	C
Campylcobacter jejuni	G	G	C	G	A	C	A	U	C	C	G	C	U	G	U	C
Escherichia coli	G	G	C	G	C	G	U	U	C	C	G	C	G	C	A	C
Haemophilus influenzae	G	G	C	G	U	G	U	U	C	C	G	C	A	C	A	C
Xylella fastidiosa	G	G	C	C	U	G	G	U	C	C	G	G	A	C	C	C
Vibro cholerae	G	G	C	G	C	G	U	U	C	C	G	C	G	C	A	C
Caulobacter crescentus	G	G	C	C	C	G	G	U	C	C	G	G	G	C	C	C
Deinococcus radiodurans	G	G	C	G	C	U	G	U	C	C	G	C	G	A	C	C
Thermotoga maritima	G	G	C	G	G	C	G	U	C	C	G	C	C	G	C	C
Chlamidia trachomatis	G	G	U	G	G	C	A	U	C	C	A	C	C	G	U	C
Aqifex aeolicus	G	G	C	G	A	C	G	U	C	C	G	C	U	G	C	C
Treponema pallidum	G	G	C	G	C	C	G	U	C	C	G	C	G	G	C	C
Borrelia burgdorferi	G	G	C	G	C	U	G	U	C	C	G	C	G	A	C	C
Acinetobacter sp	G	G	C	G	G	G	G	U	C	C	G	C	C	U	C	C
Agrobacterium tumefaciens	G	G	C	C	U	C	G	U	C	C	G	G	A	G	C	C

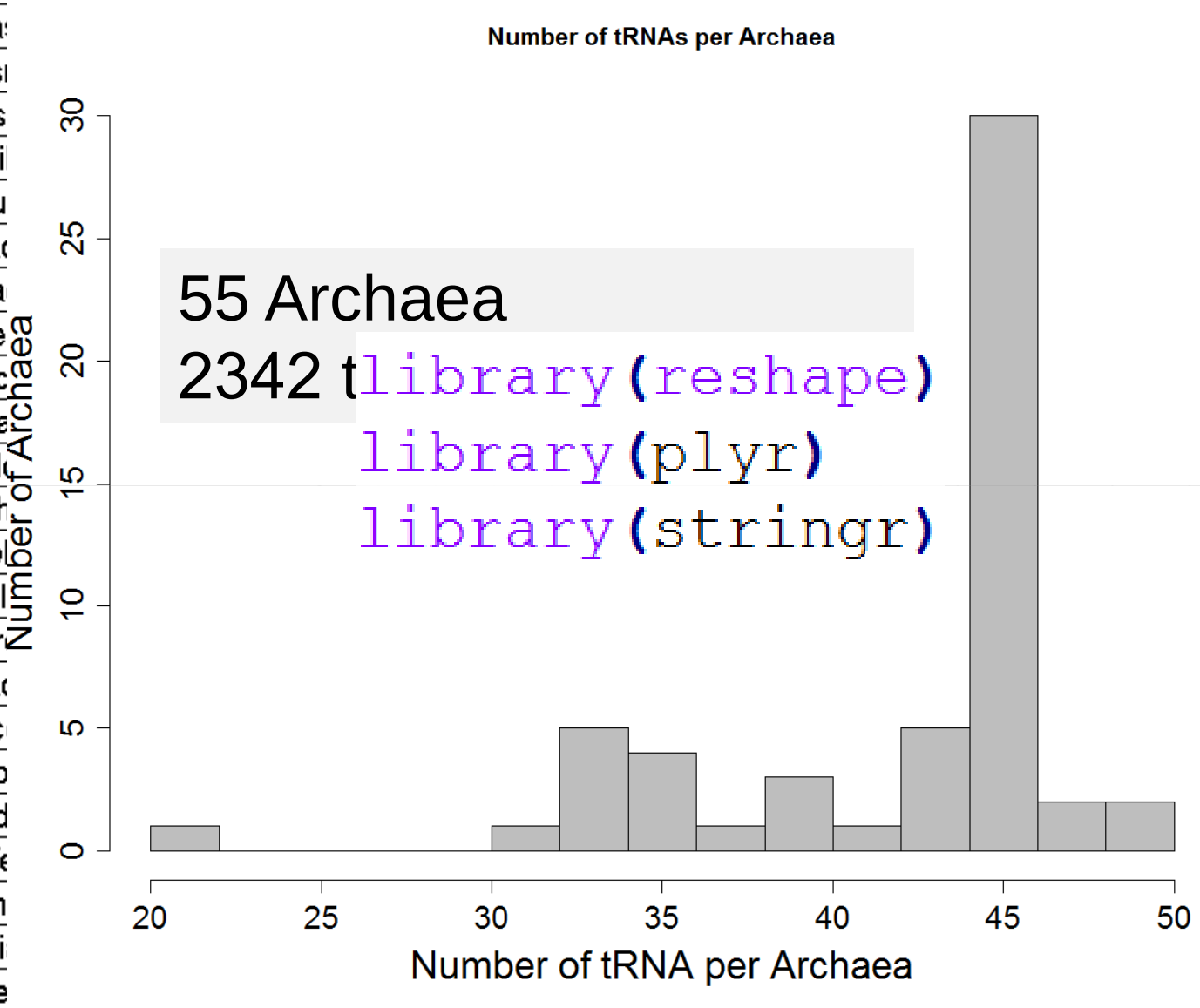
```
library(Biostrings)
readFASTA("http://gtrnadb....")
```

Species	N1	N2	N3	N4	N5	N6	N7	N73	N72	N71	N70	N69	N68	N67	N66	AA
Mycoplasma capricolum	G	G	C	A	A	C	A	U	C	C	G	U	U	G	U	C
Mycoplasma genitalium	G	G	U	G	C	C	U	U	C	C	A	C	G	G	A	C
Mycoplasma pneumoniae	G	G	U	G	C	C	U	U	C	C	A	C	G	G	A	C
Ureaplasma urealyticum	G	G	U	G	C	C	A	U	C	C	A	C	G	G	U	C
Bacillus subtilis	G	G	C	G	G	C	A	U	C	C	G	C	U	G	U	C
Bacillus halodurans	G	G	C	G	G	C	A	U	C	C	G	C	C	G	U	C
Bifidobacterium longum	G	G	U	G	G	G	U	U	C	C	A	C	C	C	A	C
Mycobacterium leprae	G	G	U	G	G	A	G	U	C	C	A	C	C	U	C	C
Mycobacterium tuberculosis	G	G	U	G	G	A	G	U	C	C	A	C	C	U	C	C
Synechocystis sp.	G	G	C	G	G	C	A	U	C	C	G	C	C	G	U	C
Helicobacter pylori J99	G	G	C	G	A	C	A	U	C	C	G	C	U	G	U	C
Campylcobacter jejuni	G	G	C	G	A	C	A	U	C	C	G	C	U	G	U	C
Escherichia coli	G	G	C	G	C	G	U	U	C	C	G	C	G	C	A	C
Haemophilus influenzae	G	G	C	G	U	G	U	U	C	C	G	C	A	C	A	C
Xylella fastidiosa	G	G	C	C	U	G	G	U	C	C	G	G	A	C	C	C
Vibro cholerae	G	G	C	G	C	G	U	U	C	C	G	C	G	C	A	C
Caulobacter crescentus	G	G	C	C	C	G	G	U	C	C	G	G	G	C	C	C
Deinococcus radiodurans	G	G	C	G	C	U	G	U	C	C	G	C	G	A	C	C
Thermotoga maritima	G	G	C	G	G	C	G	U	C	C	G	C	C	G	C	C
Chlamidia trachomatis	G	G	U	G	G	C	A	U	C	C	A	C	C	G	U	C
Aqifex aeolicus	G	G	C	G	A	C	G	U	C	C	G	C	U	G	C	C
Treponema pallidum	G	G	C	G	C	C	G	U	C	C	G	C	G	G	C	C
Borrelia burgdorferi	G	G	C	G	C	U	G	U	C	C	G	C	G	A	C	C
Acinetobacter sp	G	G	C	G	G	G	G	U	C	C	G	C	C	U	C	C
Agrobacterium tumefaciens	G	G	C	C	U	C	G	U	C	C	G	G	A	G	C	C

Species	N1	N2	N3	N4	N5	N6	N7	N73	N72	N71	N70	N69	N68	N67	N66	AA
Mycoplasma capricolum	G	G	C	A	A	C	A	U	C	C	G	U	U	G	U	C
Mycoplas															A	C
Mycoplasi															A	C
Ureaplas															U	C
Baci															U	C
Bacillu															U	C
Bifidobac															A	C
Mycoba															C	C
Mycobacte															C	C
Synec															U	C
Helicoba															U	C
Campyl															U	C
Esch															A	C
Haemop															A	C
Xylell															C	C
Vibr															A	C
Caulobac															C	C
Deinococ															C	C
Thermo															C	C
Chlamid															U	C
Aqife															C	C
Trepon															C	C
Borreli															C	C
Acine															C	C
Agrobacterium tumefaciens	G	G	C	C	U	C	G	U	C	C	G	G	A	G	C	C



Species	N1	N2	N3	N4	N5	N6	N7	N73	N72	N71	N70	N69	N68	N67	N66	AA
Mycoplasma capricolum	G	G	C	A	A	C	A	U	C	C	G	U	U	G	U	C
Mycoplas															A	C
Mycoplasi															A	C
Ureaplas															U	C
Baci															U	C
Bacillu															U	C
Bifidobac															A	C
Mycoba															C	C
Mycobacte															C	C
Synec															U	C
Helicoba															U	C
Campyl															U	C
Esch															A	C
Haemop															A	C
Xylell															C	C
Vibr															A	C
Caulobac															C	C
Deinococ															C	C
Thermo															C	C
Chlamid															U	C
Aqife															C	C
Trepon															C	C
Borreli															C	C
Acine															C	C
Agrobacterium tumefaciens	G	G	C	C	U	C	G	U	C	C	G	G	A	G	C	C



55 Archaea

2342 t

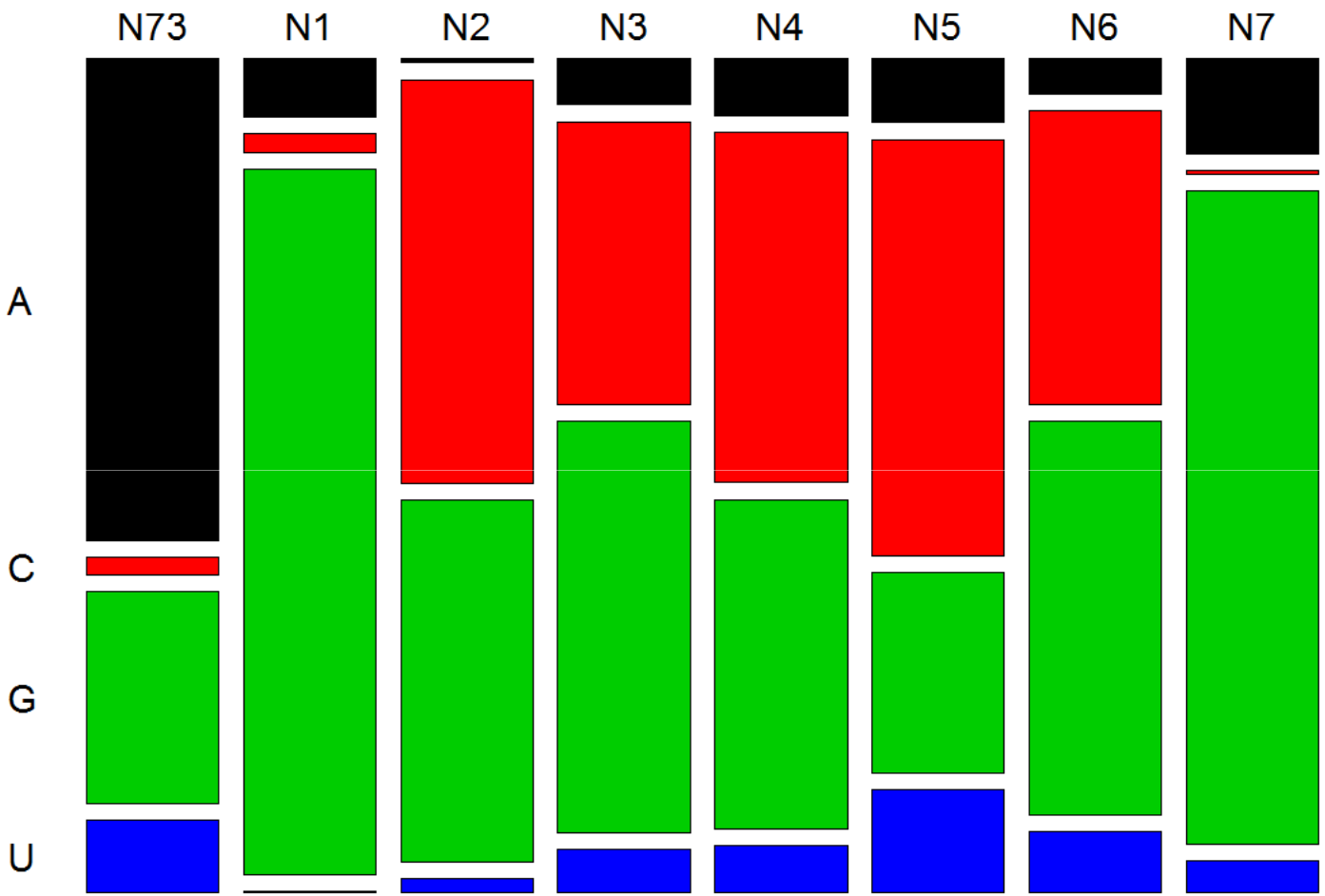
```
library(reshape)
library(plyr)
library(stringr)
```

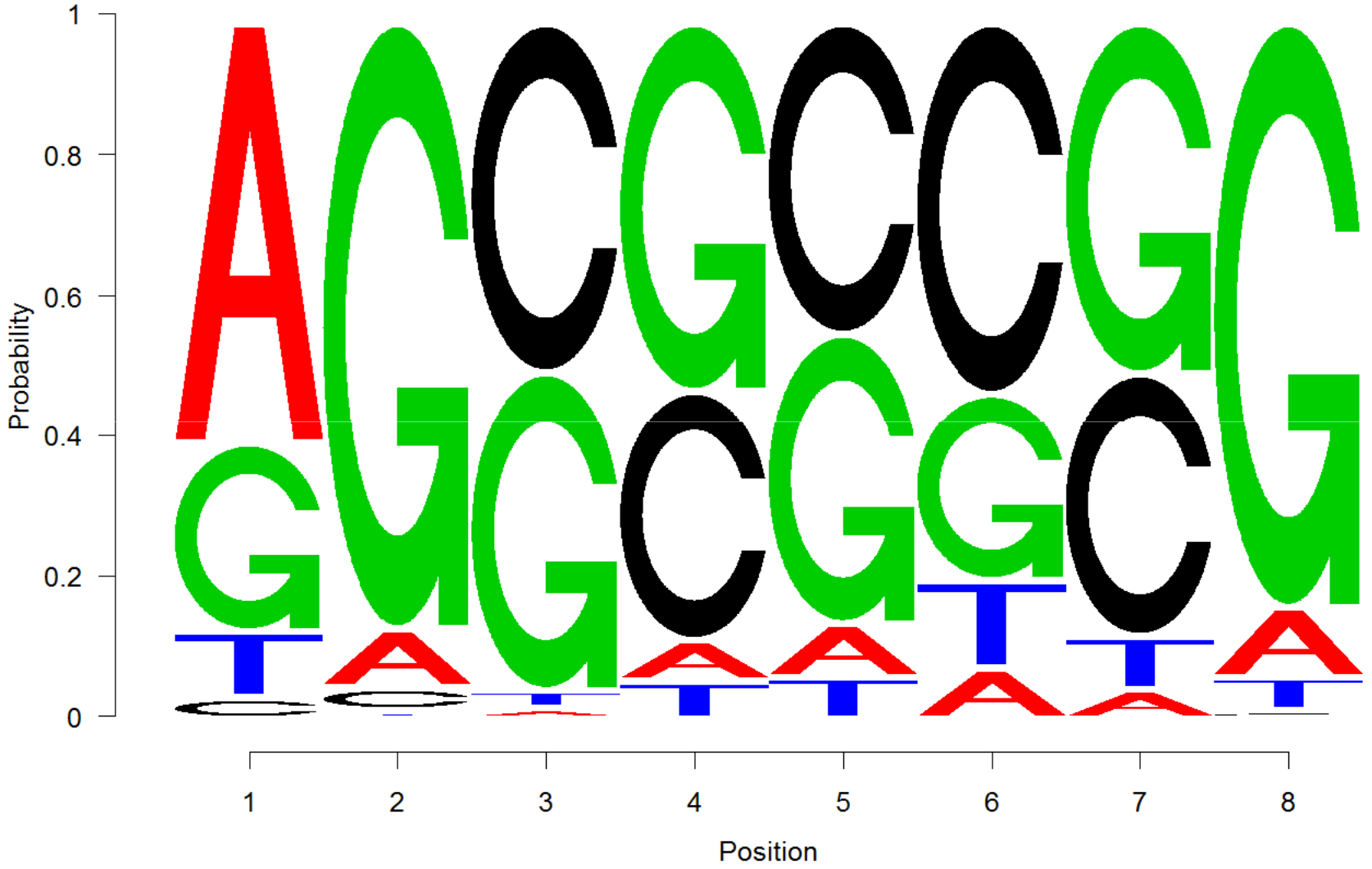
Visualize

2342
Data rows

Species	N1	N2	N3	N4	N5	N6	N7	N73
Mycoplasma capricolum	G	G	C	A	A	C	A	U
Mycoplasma genitalium	G	G	U	G	C	C	U	U
Mycoplasma pneumoniae	G	G	U	G	C	C	U	U
Ureaplasma urealyticum	G	G	U	G	C	C	A	U
Bacillus subtilis	G	G	C	G	G	C	A	U
Bacillus halodurans	G	G	C	G	G	C	A	U
Bifidobacterium longum	G	G	U	G	G	G	U	U
Mycobacterium leprae	G	G	U	G	G	A	G	U
Mycobacterium tuberculosis	G	G	U	G	G	A	G	U
Synechocystis sp.	G	G	C	G	G	C	A	U
Helicobacter pylori J99	G	G	C	G	A	C	A	U
Campylcobacter jejuni	G	G	C	G	A	C	A	U
Escherichia coli	G	G	C	G	C	G	U	U

Mosaic plot: nucleotide distribution in tRNA acceptor stem

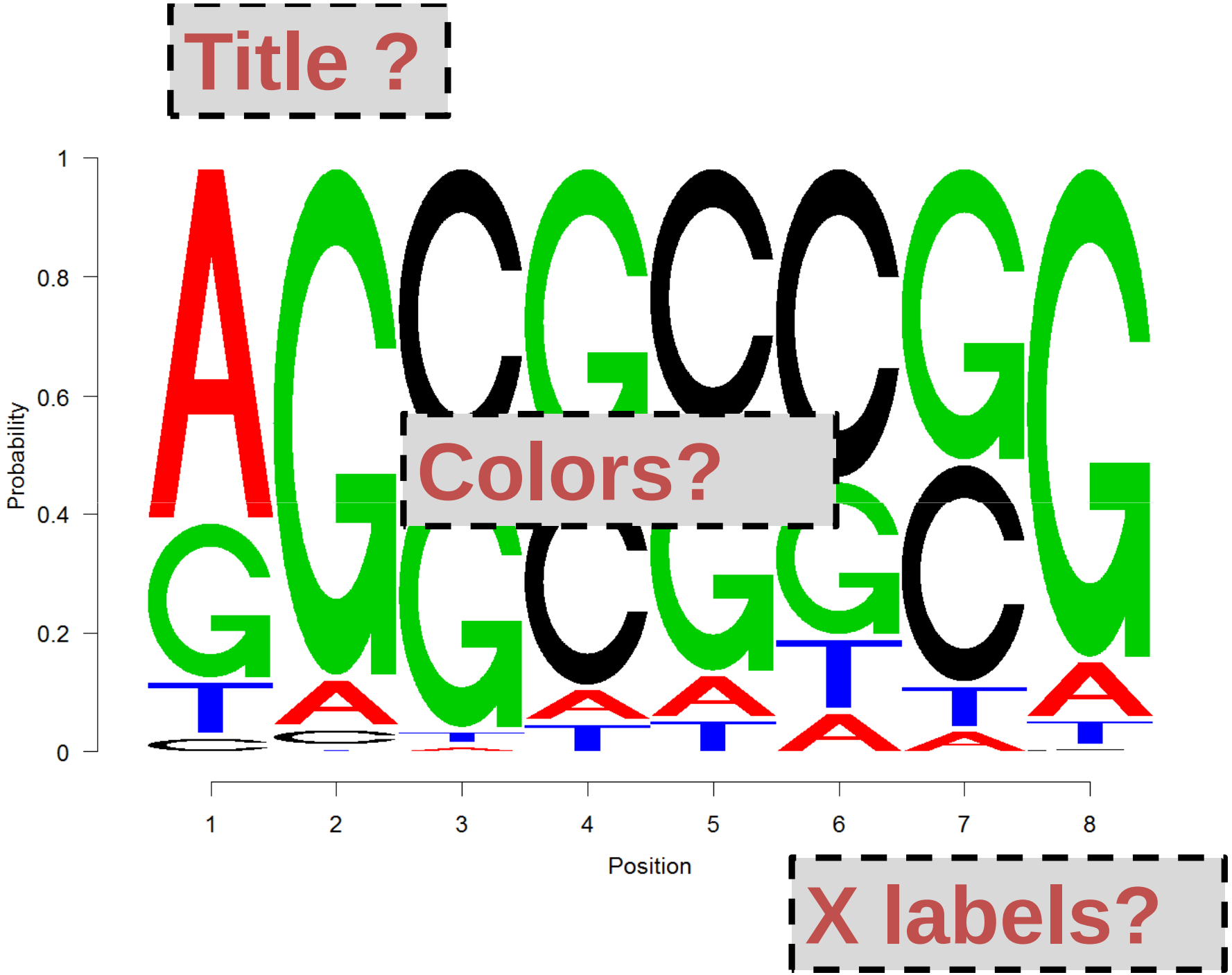




Sequence logo plot

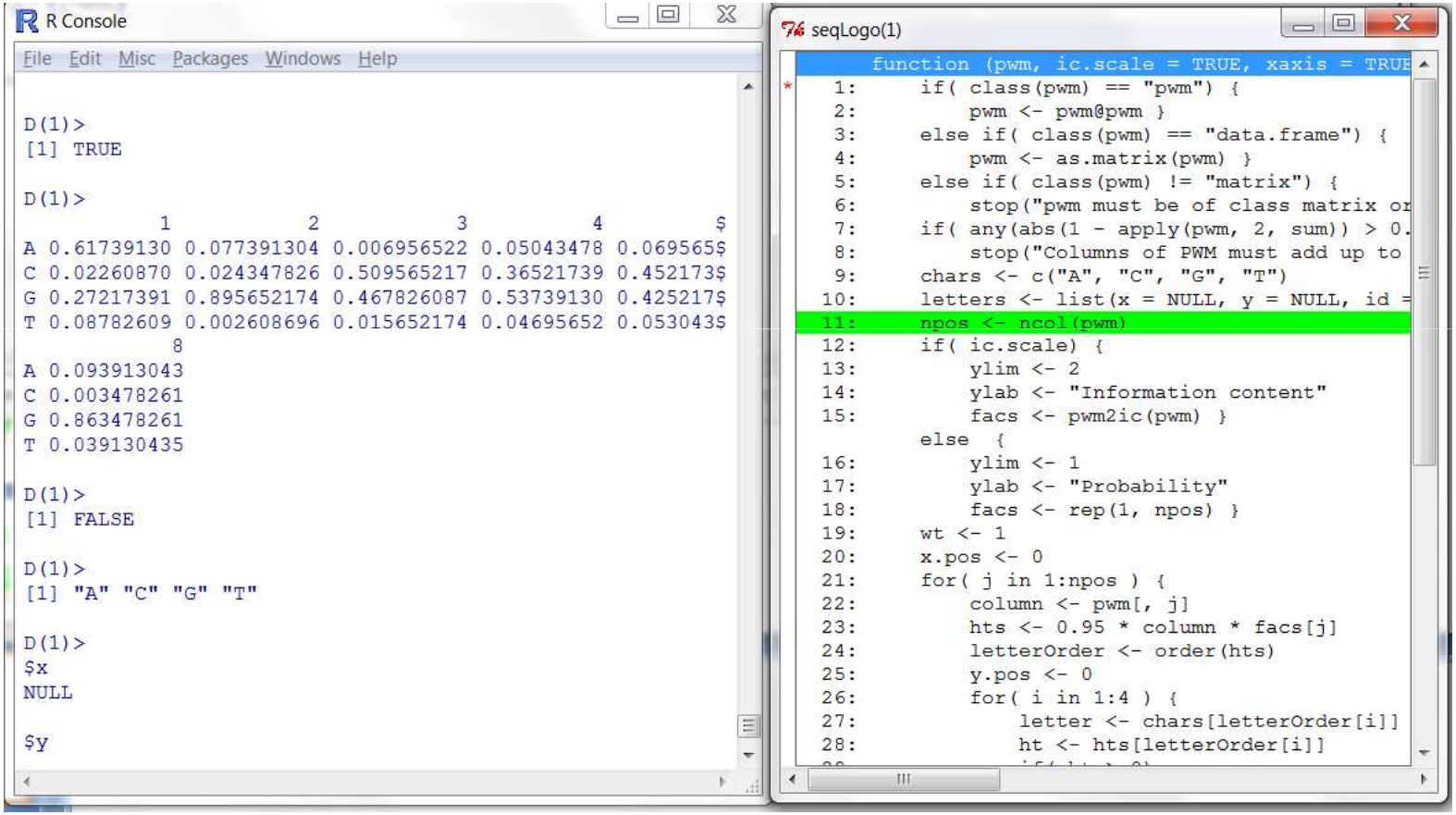
(Schneider and Stephens - 1990)

```
library("seqLogo")
seqLogo(...)
```

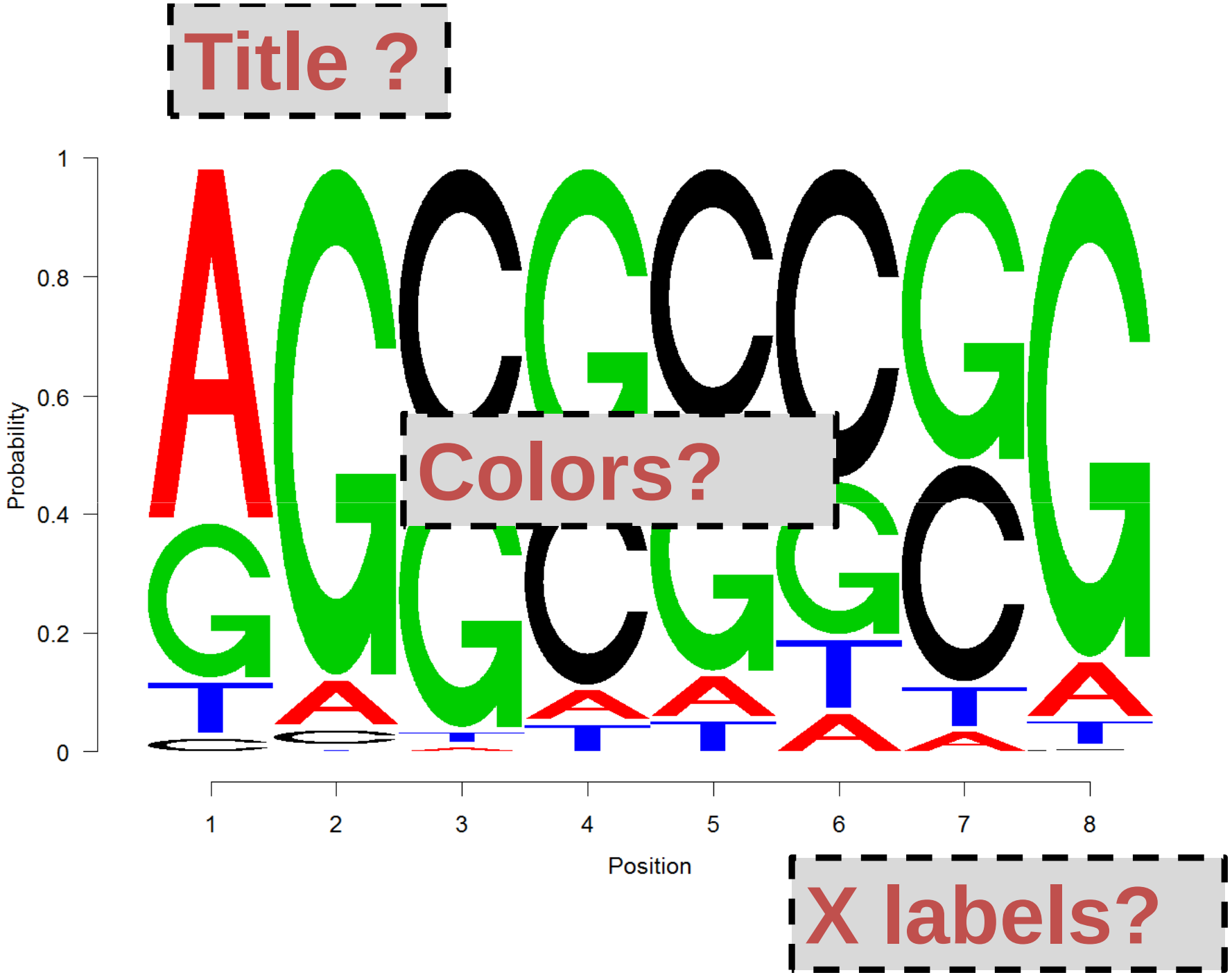

Title ?
Colors?
X labels?
Probability
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
7
8
Position

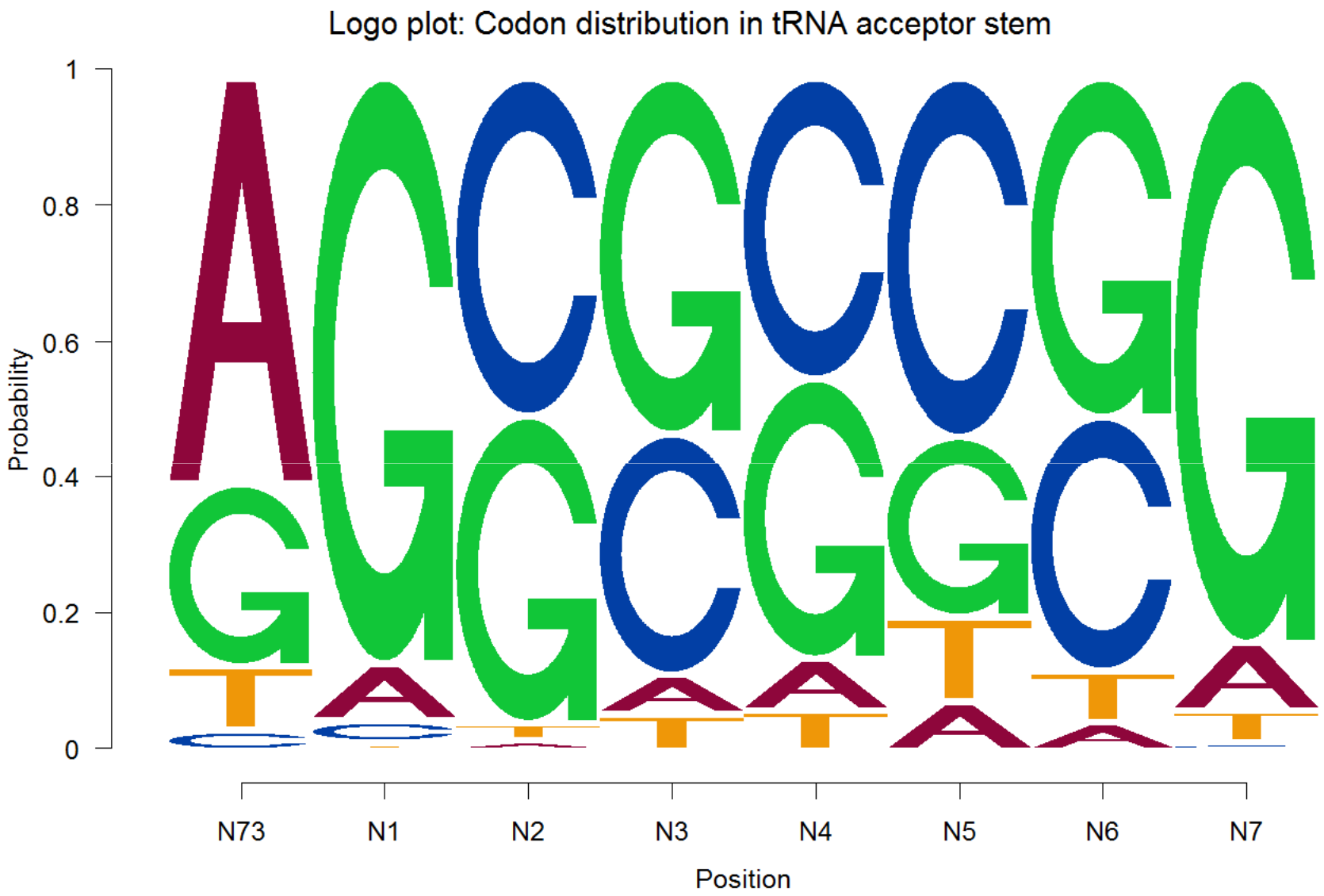
```
library(debug)
mtrace(seqLogo)
seqLogo(pwm)

mtrace.off()
```



-> “grid“ graphics….


Title ?
Colors?
X labels?
Probability
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
7
8
Position

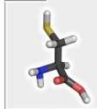

Logo plot: Codon distribution in tRNA acceptor stem
Probability
1
0.8
0.6
0.4
0.2
0
N73
N1
N2
N3
N4
N5
N6
N7
Position

standard Amino Acids

A L-Alanine ALA

Chemical Formula	$C_3H_7NO_2$
Molecular Weight	89.09 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.616
Hydropathy Index	1.8
Isoelectric Point	6.01
pKa (α-COOH)	2.35
pKa (α-NH2)	9.87
Canonical SMILES	CAS Number 56-41-7
CC(N)C(=O)O	PubChem CID 5950

C L-Cysteine CYS



Chemical Formula	$C_3H_7NO_2S$
Molecular Weight	121.16 Da
Polarity	Polar
Acidity/ Basicity	Neutral
Hydrophobicity	0.680
Hydropathy Index	2.5
Isoelectric Point	5.05
pKa (α-COOH)	1.92
pKa (α-NH2)	10.70
Canonical SMILES	CAS Number 52-90-4
NC(CS)C(=O)O	PubChem CID 5862

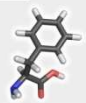
D L-Aspartic Acid ASP

Chemical Formula	$C_4H_7NO_4$
Molecular Weight	133.10 Da
Polarity	Polar
Acidity/ Basicity	Acidic
Hydrophobicity	0.028
Hydropathy Index	-3.5
Isoelectric Point	2.85
pKa (α-COOH)	1.99
pKa (α-NH2)	9.90
Canonical SMILES	CAS Number 56-84-8
NC(CC(=O)O)C(=O)O	PubChem CID 5960

E L-Glutamic Acid GLU

Chemical Formula	$C_5H_9NO_4$
Molecular Weight	147.13 Da
Polarity	Polar
Acidity/ Basicity	Acidic
Hydrophobicity	0.043
Hydropathy Index	-3.5
Isoelectric Point	3.15
pKa (α-COOH)	2.10
pKa (α-NH2)	9.47
Canonical SMILES	CAS Number 56-86-0
NC(CCC(=O)O)C(=O)O	PubChem CID 611

F L-Phenylalanine PHE

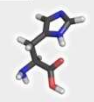


Chemical Formula	$C_9H_{11}NO_2$
Molecular Weight	165.19 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	1.00
Hydropathy Index	2.8
Isoelectric Point	5.49
pKa (α-COOH)	2.20
pKa (α-NH2)	9.31
Canonical SMILES	CAS Number 63-91-2
NC(Cc1ccccc1)C(=O)O	PubChem CID 994

G Glycine GLY

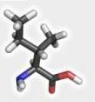
Chemical Formula	$C_2H_5NO_2$
Molecular Weight	75.07 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.501
Hydropathy Index	-0.4
Isoelectric Point	6.06
pKa (α-COOH)	2.35
pKa (α-NH2)	9.78
Canonical SMILES	CAS Number 56-40-6
NCC(=O)O	PubChem CID 750

H L-Histidine HIS



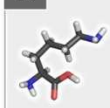
Chemical Formula	$C_6H_9N_3O_2$
Molecular Weight	155.15 Da
Polarity	Polar
Acidity/ Basicity	Basic (Weak)
Hydrophobicity	0.165
Hydropathy Index	-3.2
Isoelectric Point	7.60
pKa (α-COOH)	1.80
pKa (α-NH2)	9.33
Canonical SMILES	CAS Number 71-00-1
NC(Cc1cnc[nH]1)C(=O)O	PubChem CID 773

I L-Isoleucine ILE



Chemical Formula	$C_6H_{13}NO_2$
Molecular Weight	131.17 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.943
Hydropathy Index	4.5
Isoelectric Point	6.05
pKa (α-COOH)	2.32
pKa (α-NH2)	9.76
Canonical SMILES	CAS Number 73-32-5
CCC(C)C(N)C(=O)O	PubChem CID 791

K L-Lysine LYS

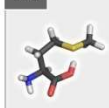


Chemical Formula	$C_6H_{14}N_2O_2$
Molecular Weight	146.19 Da
Polarity	Polar
Acidity/ Basicity	Basic
Hydrophobicity	0.283
Hydropathy Index	-3.9
Isoelectric Point	9.60
pKa (α-COOH)	2.16
pKa (α-NH2)	9.06
Canonical SMILES	CAS Number 56-87-1
NCCCCC(N)C(=O)O	PubChem CID 866

L L-Leucine LEU

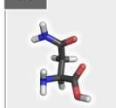
Chemical Formula	$C_6H_{13}NO_2$
Molecular Weight	131.18 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.943
Hydropathy Index	3.8
Isoelectric Point	6.01
pKa (α-COOH)	2.33
pKa (α-NH2)	9.74
Canonical SMILES	CAS Number 61-90-5
CC(C)CC(N)C(=O)O	PubChem CID 6106

M L-Methionine MET



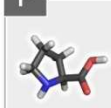
Chemical Formula	$C_5H_{11}NO_2S$
Molecular Weight	149.21 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.738
Hydropathy Index	1.9
Isoelectric Point	5.74
pKa (α-COOH)	2.13
pKa (α-NH2)	9.28
Canonical SMILES	CAS Number 63-68-3
CSCCC(N)C(=O)O	PubChem CID 876

N L-Asparagine ASN



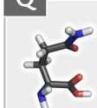
Chemical Formula	$C_4H_8N_2O_3$
Molecular Weight	132.12 Da
Polarity	Polar
Acidity/ Basicity	Neutral
Hydrophobicity	0.236
Hydropathy Index	-3.5
Isoelectric Point	5.41
pKa (α-COOH)	2.14
pKa (α-NH2)	8.72
Canonical SMILES	CAS Number 70-47-3
NC(CC(=O)N)C(=O)O	PubChem CID 236

P L-Proline PRO



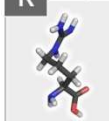
Chemical Formula	$C_5H_9NO_2$
Molecular Weight	115.13 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.711
Hydropathy Index	-1.6
Isoelectric Point	6.30
pKa (α-COOH)	1.95
pKa (α-NH2)	10.64
Canonical SMILES	CAS Number 147-85-3
OC(=O)C1CCCN1	PubChem CID 614

Q L-Glutamine GLN



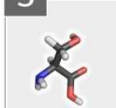
Chemical Formula	$C_5H_{10}N_2O_3$
Molecular Weight	146.14 Da
Polarity	Polar
Acidity/ Basicity	Neutral
Hydrophobicity	0.251
Hydropathy Index	-3.5
Isoelectric Point	5.65
pKa (α-COOH)	2.17
pKa (α-NH2)	9.13
Canonical SMILES	CAS Number 56-85-9
NC(CCC(=O)N)C(=O)O	PubChem CID 5950

R L-Arginine ARG



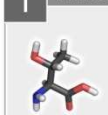
Chemical Formula	$C_6H_{14}N_4O_2$
Molecular Weight	174.20 Da
Polarity	Polar
Acidity/ Basicity	Basic (Strong)
Hydrophobicity	0.000
Hydropathy Index	-4.5
Isoelectric Point	10.76
pKa (α-COOH)	1.82
pKa (α-NH2)	8.99
Canonical SMILES	CAS Number 74-79-3
NC(CCCNC(=N)N)C(=O)O	PubChem CID 5950

S L-Serine SER



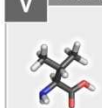
Chemical Formula	$C_3H_7NO_3$
Molecular Weight	105.09 Da
Polarity	Polar
Acidity/ Basicity	Neutral
Hydrophobicity	0.359
Hydropathy Index	-0.8
Isoelectric Point	5.68
pKa (α-COOH)	2.19
pKa (α-NH2)	9.21
Canonical SMILES	CAS Number 56-45-1
NC(CO)C(=O)O	PubChem CID 617

T L-Threonine THR



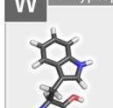
Chemical Formula	$C_4H_9NO_3$
Molecular Weight	119.12 Da
Polarity	Polar
Acidity/ Basicity	Neutral
Hydrophobicity	0.450
Hydropathy Index	-0.7
Isoelectric Point	5.60
pKa (α-COOH)	2.09
pKa (α-NH2)	9.10
Canonical SMILES	CAS Number 72-19-5
CC(O)C(N)C(=O)O	PubChem CID 6288

V L-Valine VAL



Chemical Formula	$C_5H_{11}NO_2$
Molecular Weight	117.15 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.825
Hydropathy Index	4.2
Isoelectric Point	6.00
pKa (α-COOH)	2.39
pKa (α-NH2)	9.74
Canonical SMILES	CAS Number 72-18-4
CC(C)C(N)C(=O)O	PubChem CID 1182

W L-Tryptophan TRP



Chemical Formula	$C_{11}H_{12}N_2O_2$
Molecular Weight	204.23 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.878
Hydropathy Index	-0.9
Isoelectric Point	5.89
pKa (α-COOH)	2.46
pKa (α-NH2)	9.41
Canonical SMILES	CAS Number 73-22-3
NC(Cc1c[nH]c2ccccc12)C(=O)O	PubChem CID 6305

Y L-Tyrosine TYR

Chemical Formula	$C_9H_{11}NO_3$
Molecular Weight	181.19 Da
Polarity	Nonpolar
Acidity/ Basicity	Neutral
Hydrophobicity	0.880
Hydropathy Index	-1.3
Isoelectric Point	5.64
pKa (α-COOH)	2.20
pKa (α-NH2)	9.21
Canonical SMILES	CAS Number 60-18-4
NC(Cc1ccc(O)cc1)C(=O)O	PubChem CID 1153

B L-Asparagine or L-Aspartic Acid ASX

J L-Leucine or L-Isoleucine XLE

X Unspecified or Unknown XAA

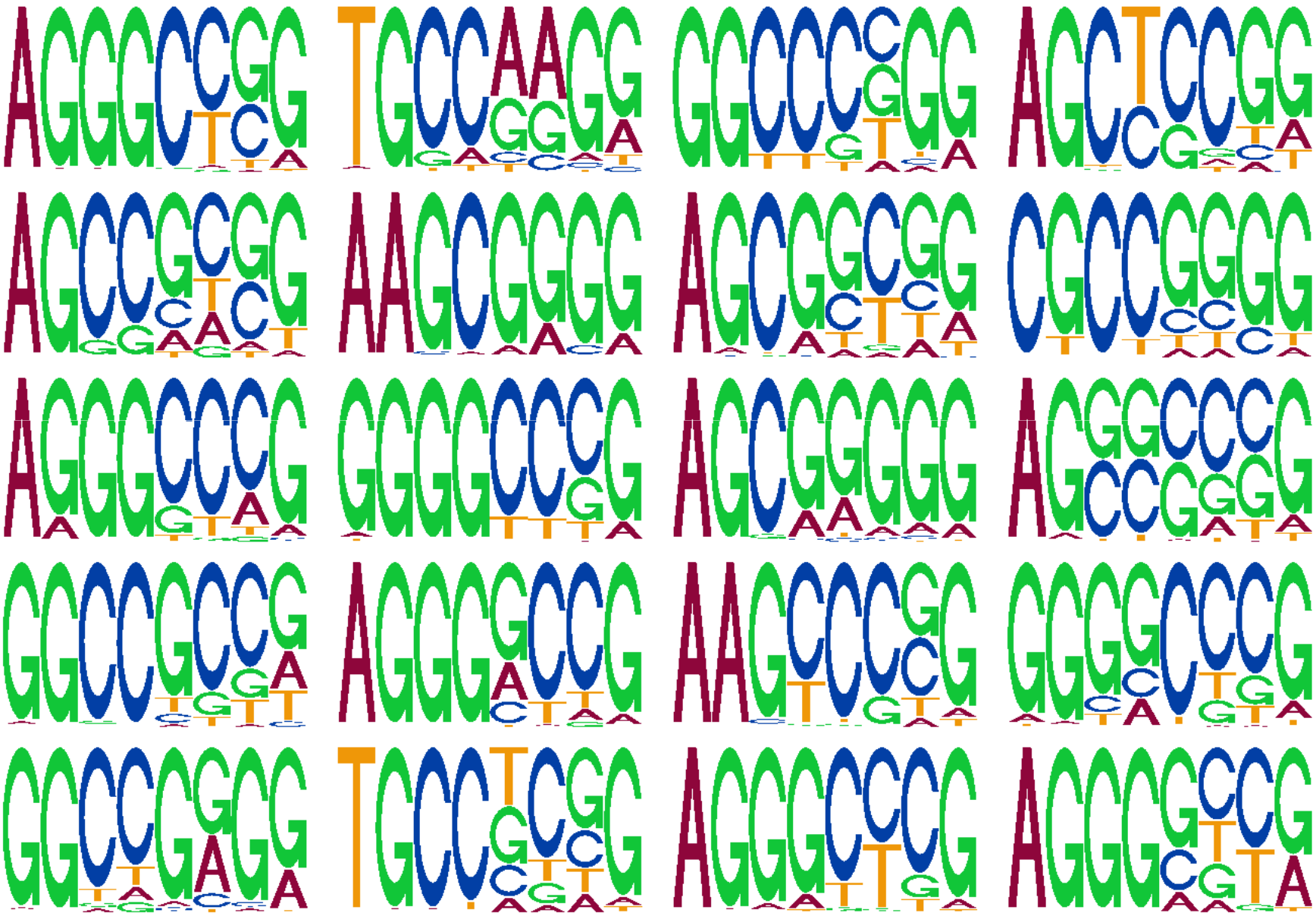
Z L-Glutamine or L-Glutamic Acid GLX

Essential amino acids are denoted with yellow spots. Polarity refers to Side Chain Polarity. Acidity/Basicity refer to Side Chain Acidity/Basicity.

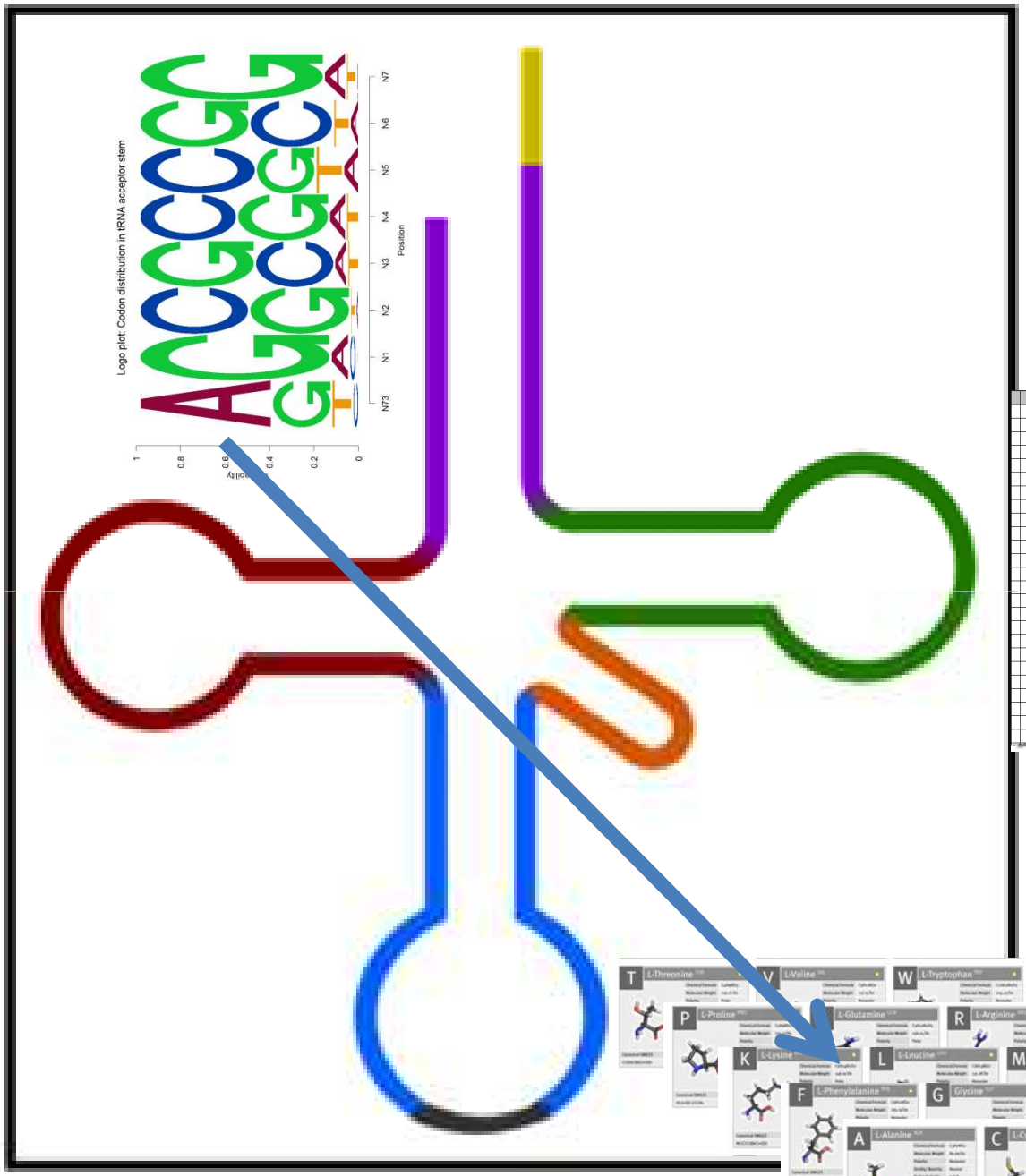
2D Structure Generation, 3D optimization: ChemSketch 8.17 (ACD Labs) | 3D Structure Rendering: PyMol 0.99rc6 (DeLano Scientific) | Chemical Formula and Molecular Weight Calculations: ChemSketch 8.17 (ACD Labs) | Hydrophobicity values: Analytical Biochemistry 193:11, 72-82, Elsevier, 1991. | Hydropathy Index values: J. Mol. Biol. (1982) 157, 105-132. | Canonical SMILES: www-jmg.zb.cam.ac.uk/data/molecules/smilesindex.html | Polarity, Acidity/Basicity, Isoelectric Point, pKa, CAS, PubChem CID data: en.wikipedia.org/wiki/List_of_standard_amino_acids | List of Essential Amino Acids: en.wikipedia.org/wiki/Amino_acid | Typeface: Meta Book

2007/2129 Berk Zafer | MMI.unc.edu | Questions, problems, and errors should be sent to berkzafer@unc.edu
The most recent version of this document can be found at www.unc.edu/~bzafer/aminoacids

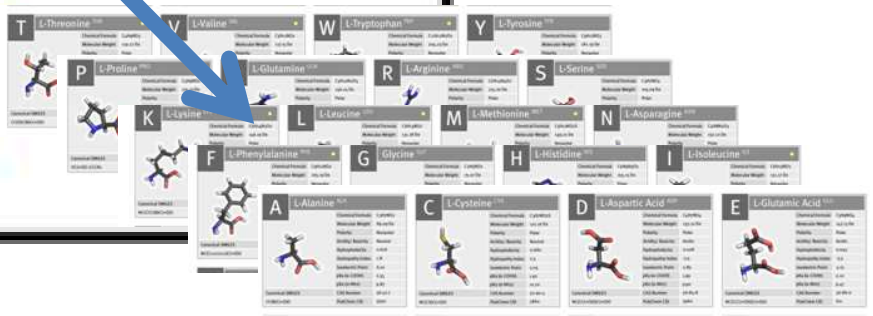
Codon distribution, per location, per Amino Acid



Analyze



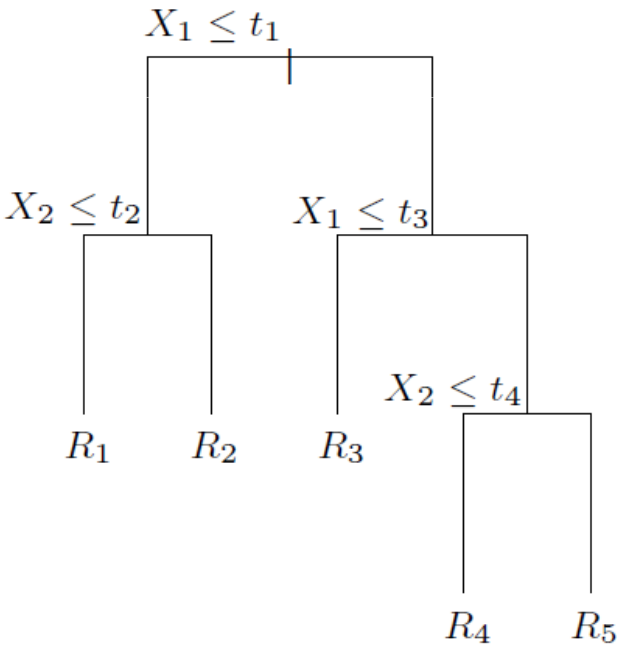
Species	N1	N2	N3	N4	N5	N6	N7	N73	N72	N71	N70	N69	N68	N67	N66	AA
Mycoplasma capricolum	G	G	C	A	A	C	A	U	C	C	G	U	U	G	U	C
Mycoplasma genitalium	G	G	U	G	C	C	U	U	C	C	A	C	G	G	A	C
Mycoplasma pneumoniae	G	G	U	G	C	C	U	U	C	C	A	C	G	G	A	C
Ureaplasma urealyticum	G	G	U	G	C	C	A	U	C	C	A	C	G	G	U	C
Bacillus subtilis	G	G	C	G	G	C	A	U	C	C	G	C	U	G	U	C
Bacillus halodurans	G	G	C	G	G	C	A	U	C	C	G	C	C	G	U	C
Bifidobacterium longum	G	G	U	G	G	G	U	U	C	C	A	C	C	C	A	C
Mycobacterium leprae	G	G	U	G	G	A	G	U	C	C	A	C	C	U	C	C
Mycobacterium tuberculosis	G	G	U	G	G	A	G	U	C	C	A	C	C	U	C	C
Synechocystis sp.	G	G	C	G	G	C	A	U	C	C	G	C	C	G	U	C
Helicobacter pylori J99	G	G	C	G	A	C	A	U	C	C	G	C	U	G	U	C
Campylcobacter jejuni	G	G	C	G	A	C	A	U	C	C	G	C	U	G	U	C
Escherichia coli	G	G	C	G	C	G	U	U	C	C	G	C	G	C	A	C
Haemophilus influenzae	G	G	C	G	U	G	U	U	C	C	G	C	A	C	A	C
Xylella fastidiosa	G	G	C	C	U	G	G	U	C	C	G	G	A	C	C	C
Vibro cholerae	G	G	C	G	C	G	U	U	C	C	G	C	G	C	A	C
Caulobacter crescentus	G	G	C	C	C	G	G	U	C	C	G	G	G	C	C	C
Deinococcus radiodurans	G	G	C	G	C	U	G	U	C	C	G	C	G	A	C	C
Thermotoga maritima	G	G	C	G	G	C	G	U	C	C	G	C	C	G	C	C
Chlamidia trachomatis	G	G	U	G	G	C	A	U	C	C	A	C	C	G	U	C
Aqifex aeolicus	G	G	C	G	A	C	G	U	C	C	G	C	U	G	C	C
Treponema pallidum	G	G	C	G	C	C	G	U	C	C	G	C	G	G	C	C
Borrelia burgdorferi	G	G	C	G	C	U	G	U	C	C	G	C	G	A	C	C
Acinetobacter sp	G	G	C	G	G	G	G	U	C	C	G	C	C	U	C	C
Agrobacterium tumefaciens	G	G	C	C	U	C	G	U	C	C	G	G	A	G	C	C



CART
(Classification and regression trees)

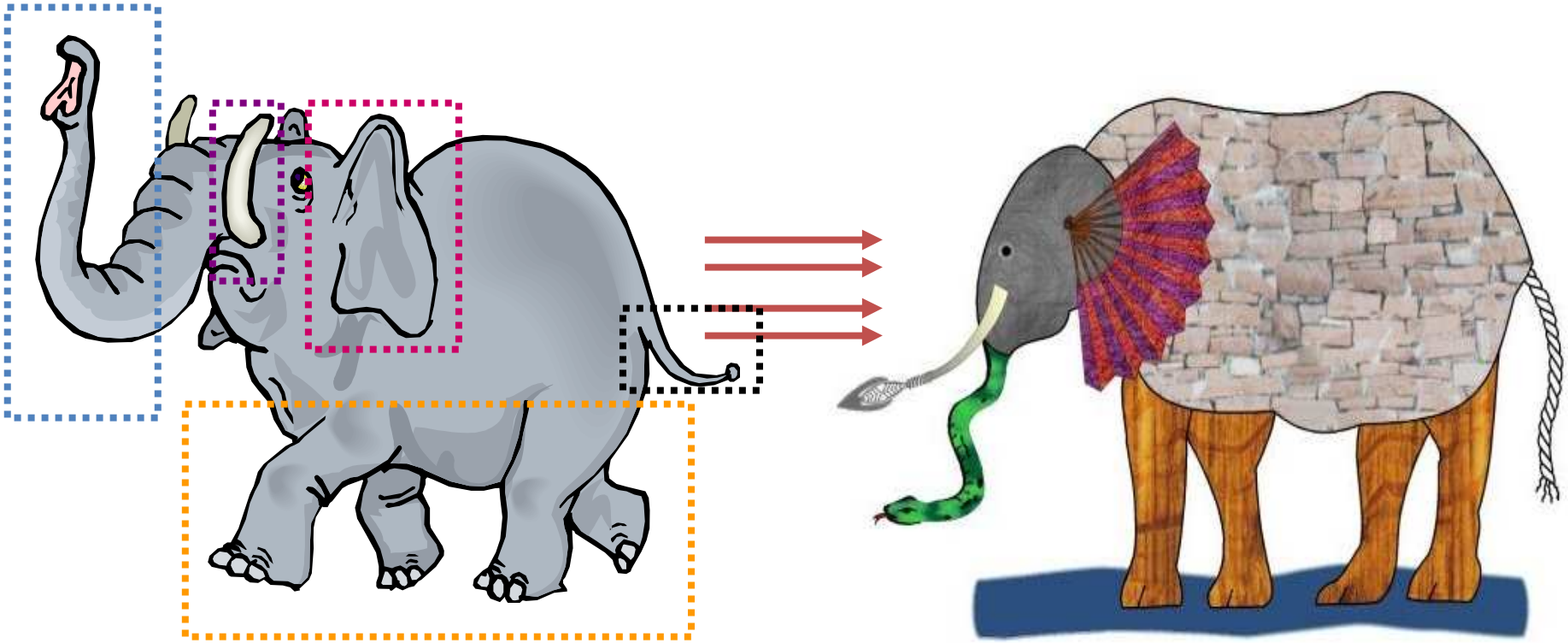
- **Goal** predict target variable

- **Method**: recursively partition explanatory variables



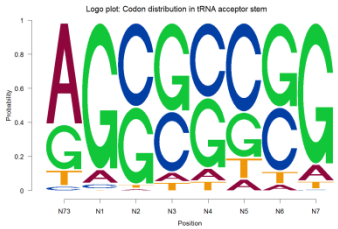
CART

(**Classification** and regression **trees**)



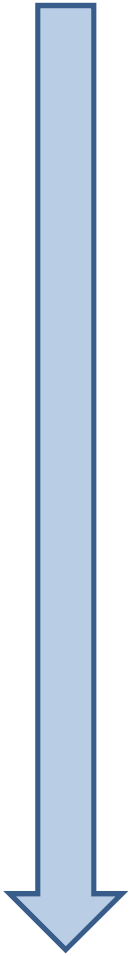
It was six men of Indostan ; To learning much inclined, Who went to the Elephant; (Though all of them were blind), That each by observation; Might satisfy his mind …. etc

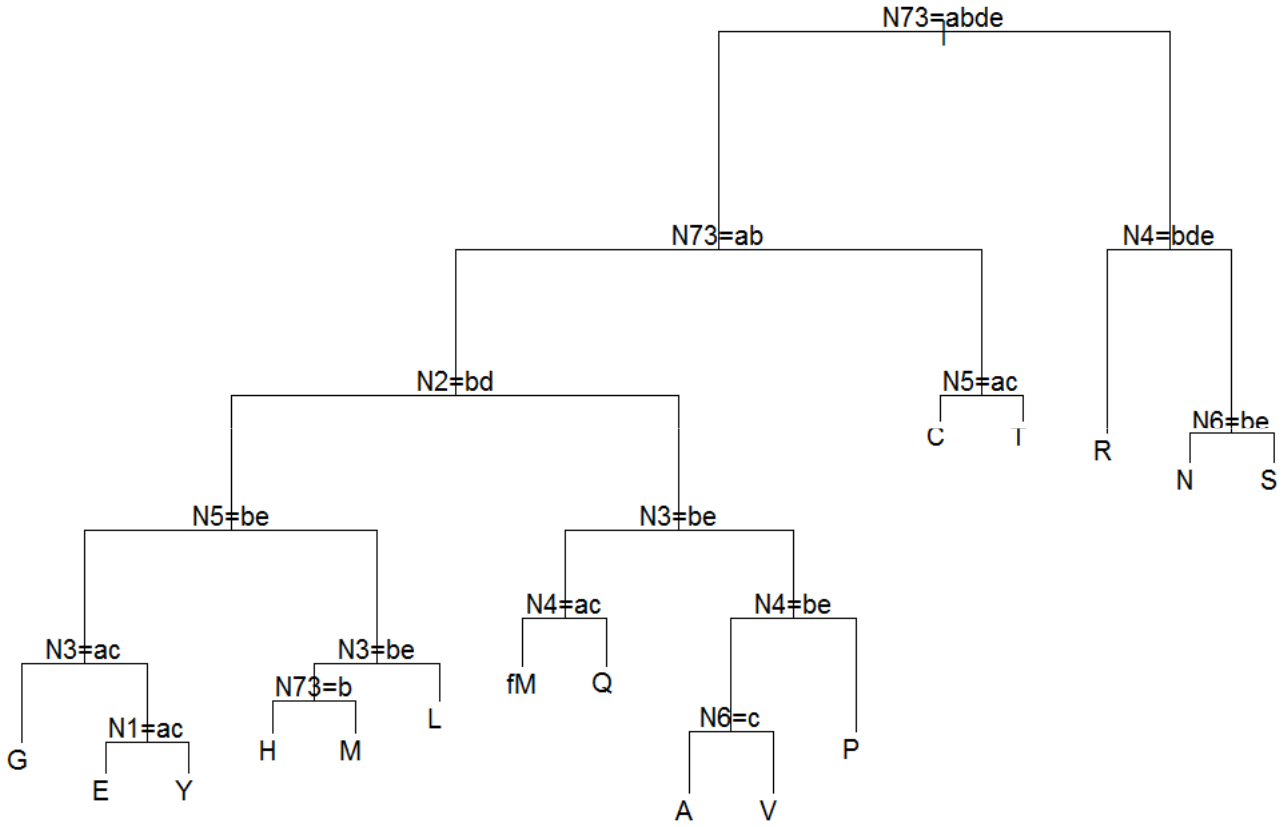
-- "**The Blind Men and the Elephant**" by John Godfrey Saxe (1816-1887)

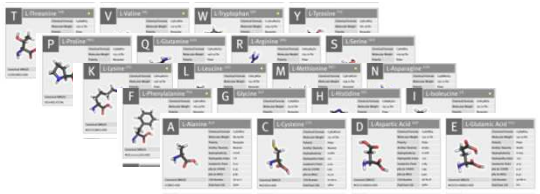


CART

with default values



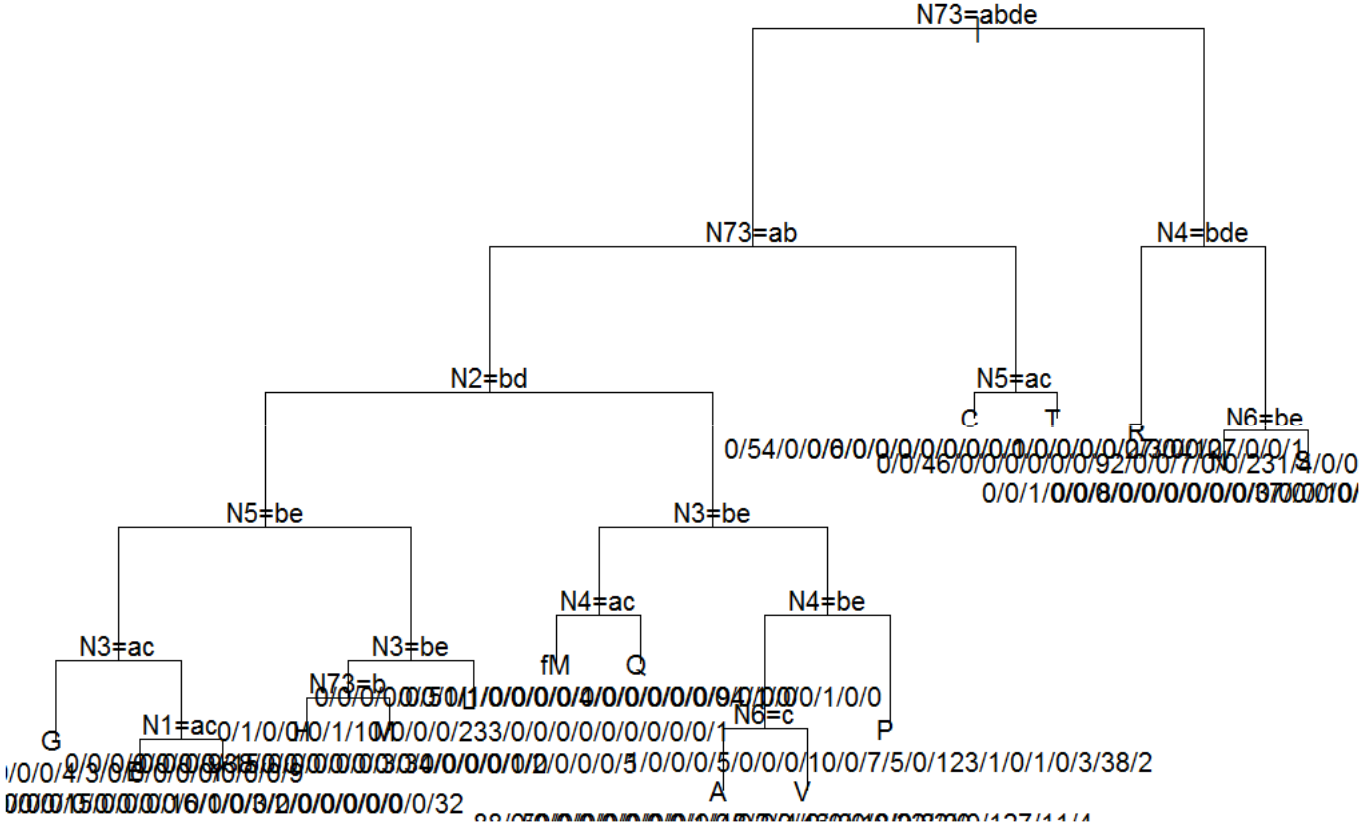




Library(rpart)

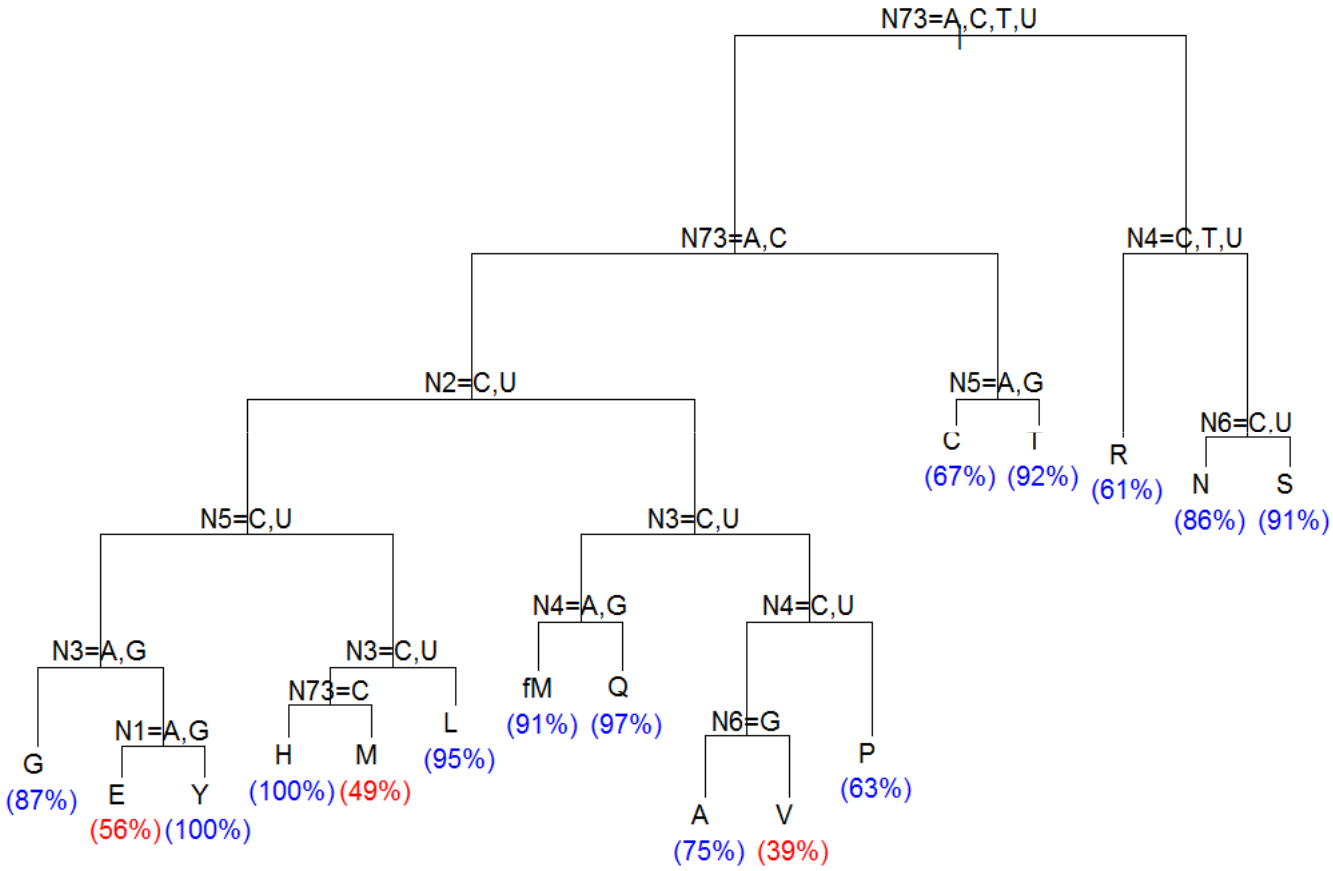
CART

Predictive success ? (use.n = T)



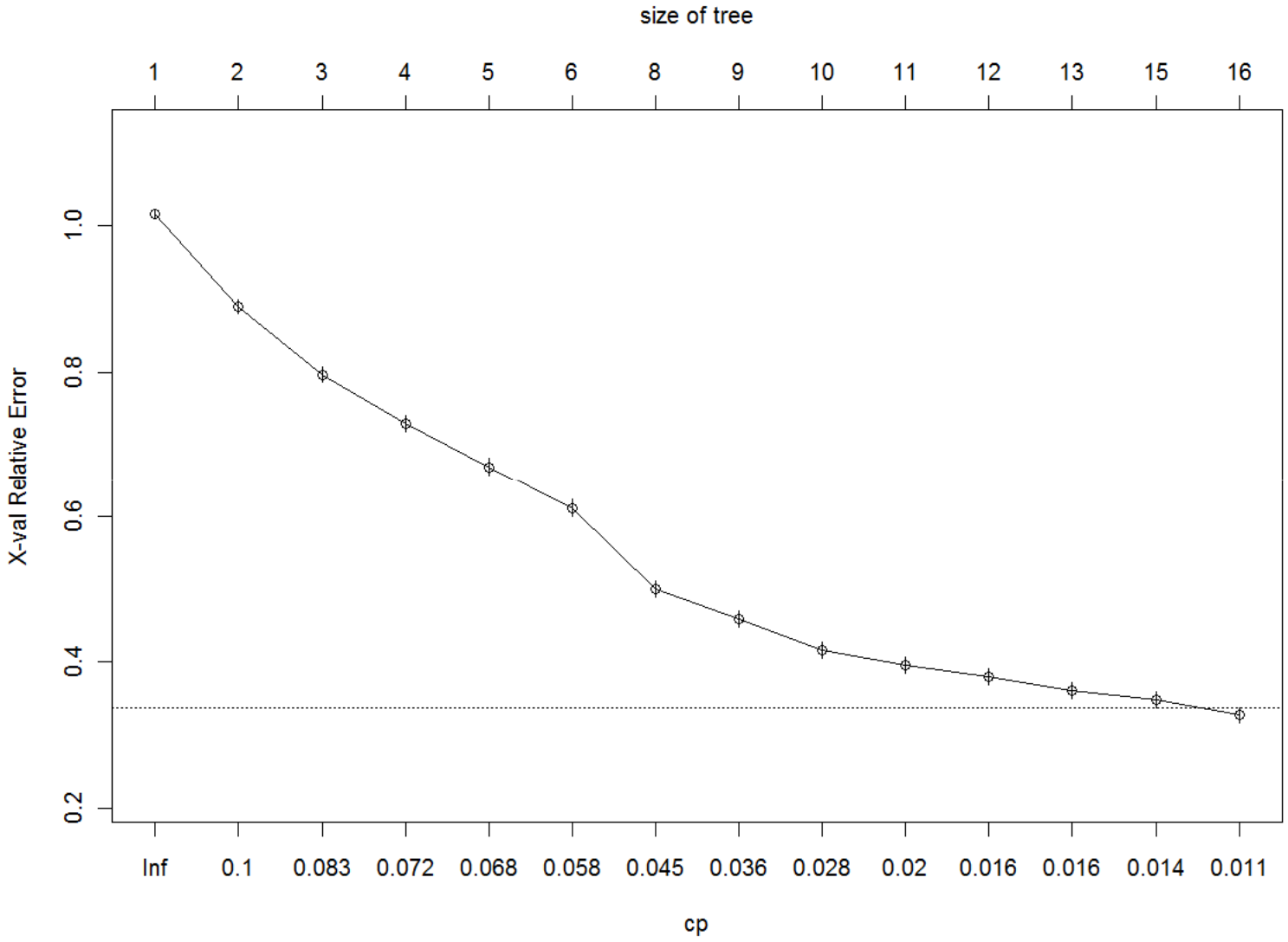
CART

Percent of success in prediction

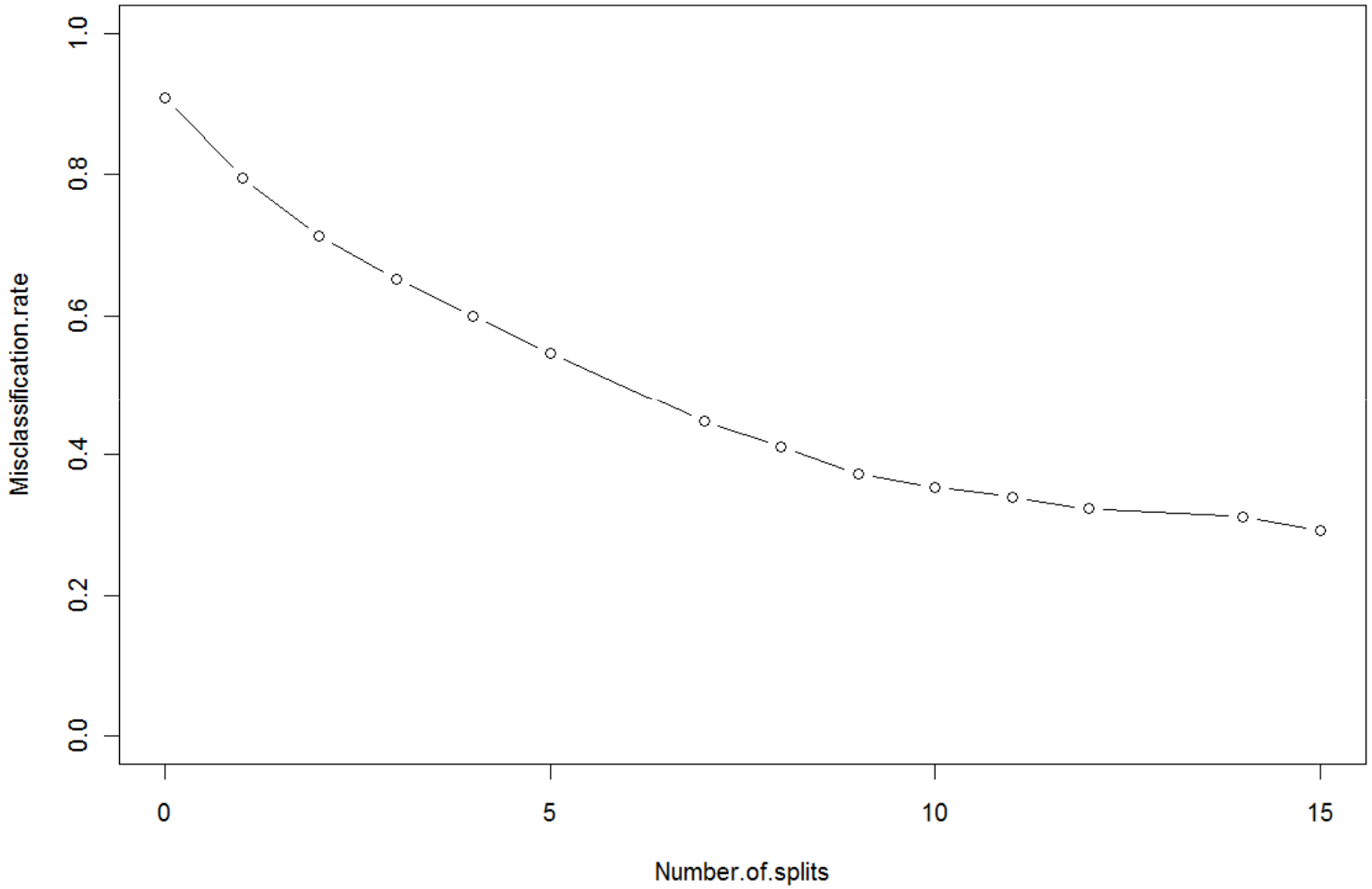


Red = bellow 60% fit

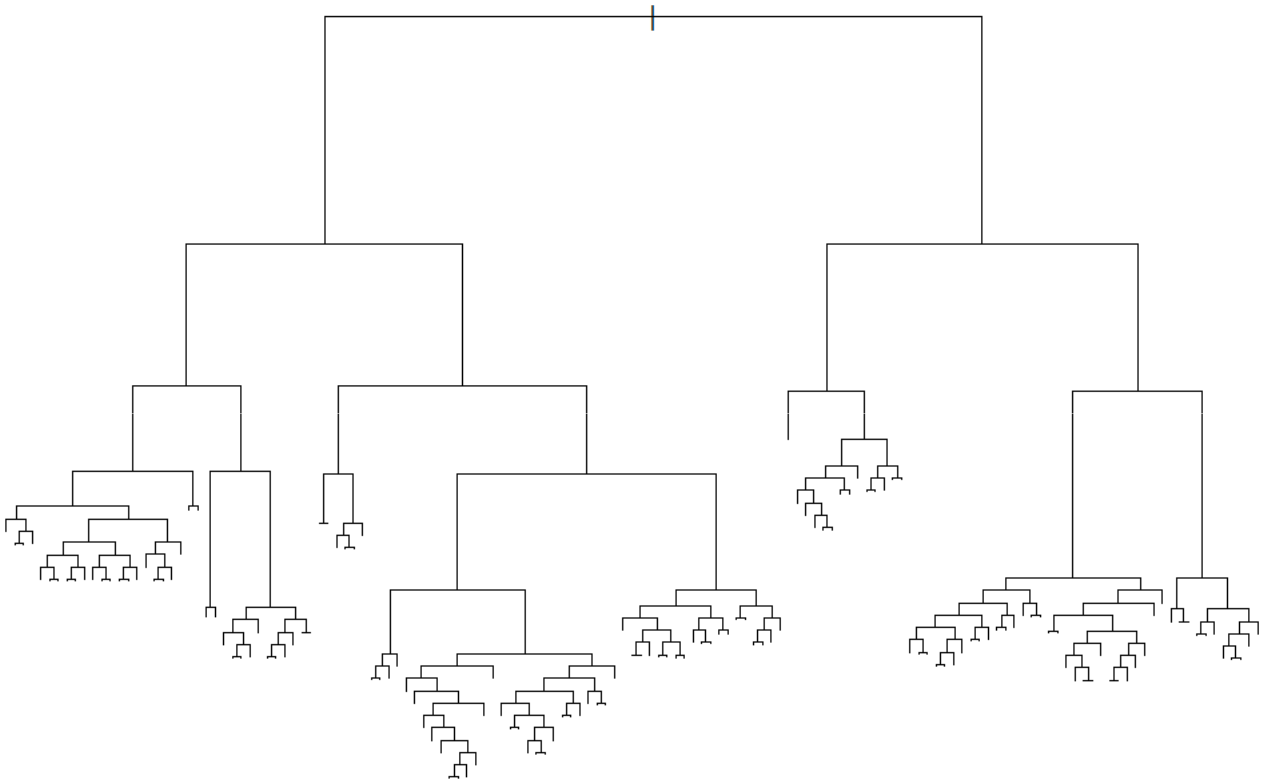
Cross validation – Relative error



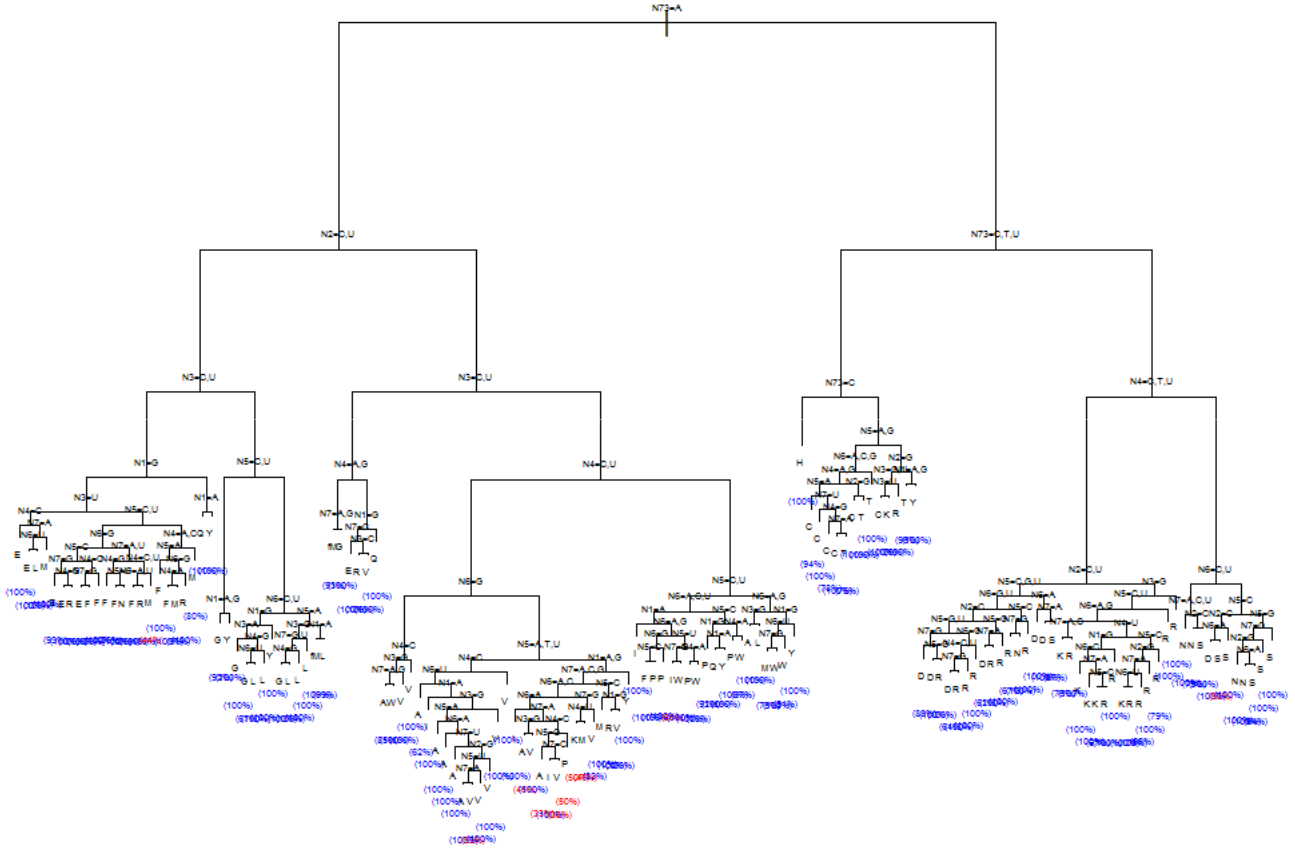
Cross validation – Misclassification rate



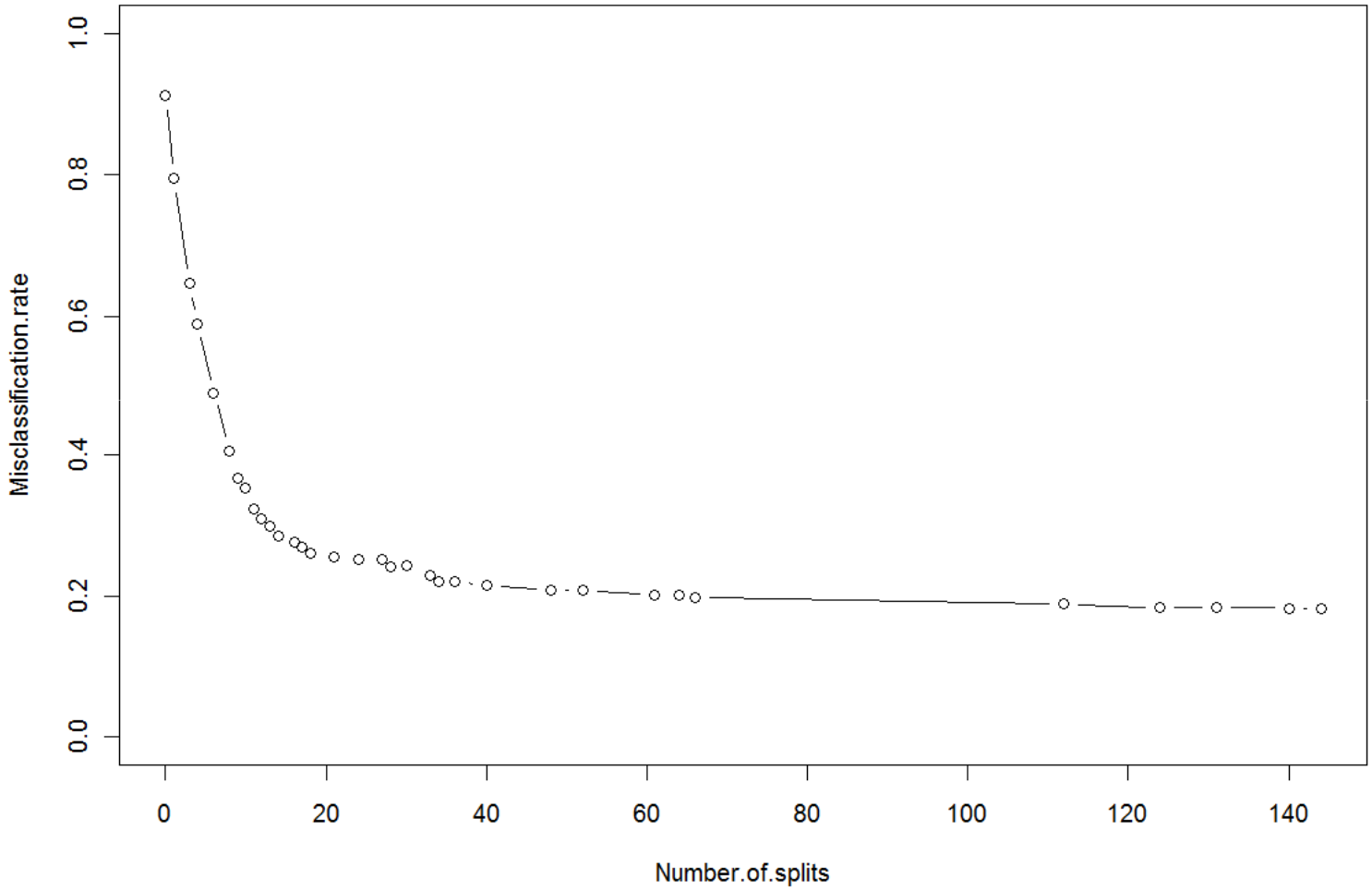
Bigger tree?



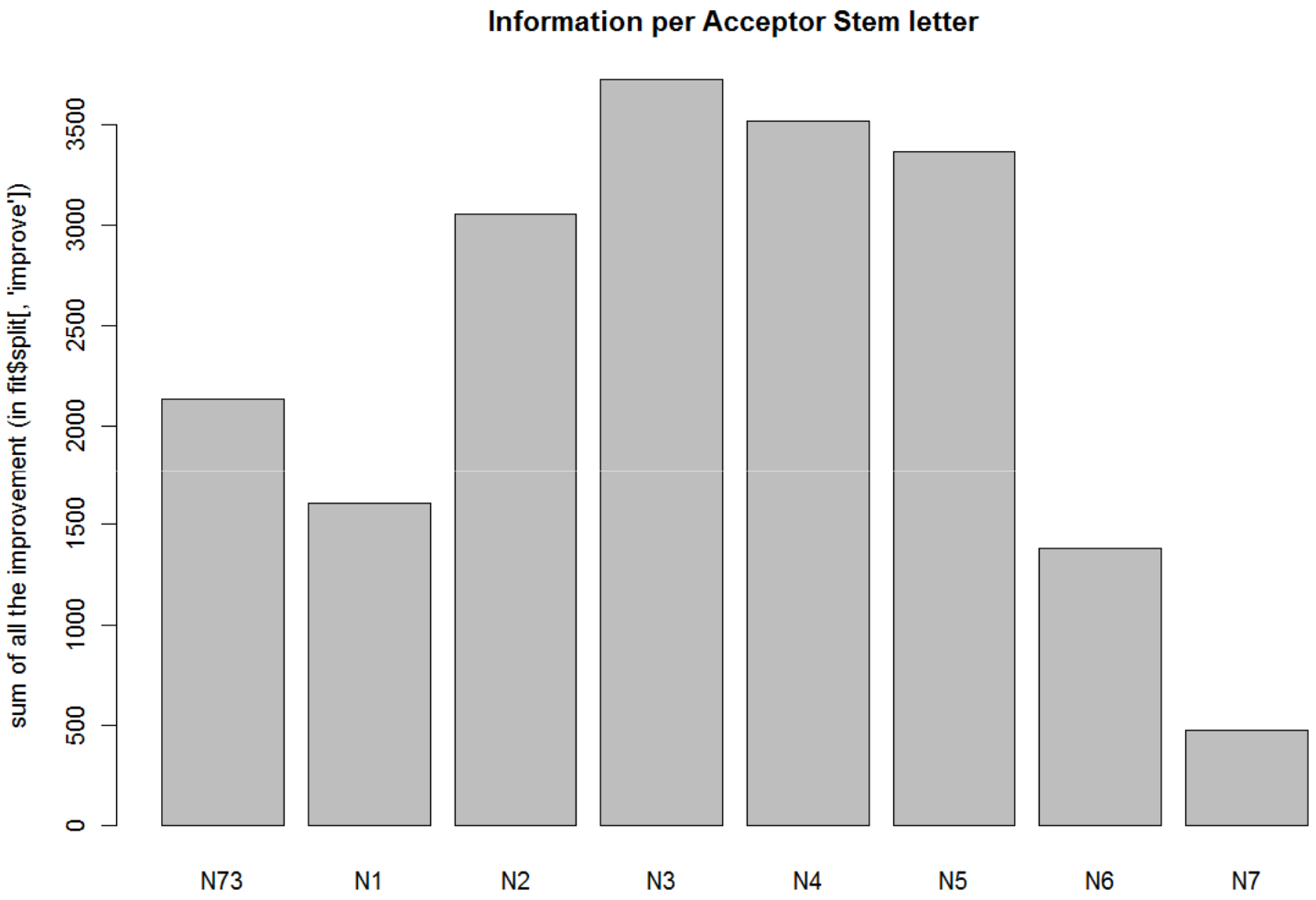
Bigger tree?



Cross validation – Misclassification rate

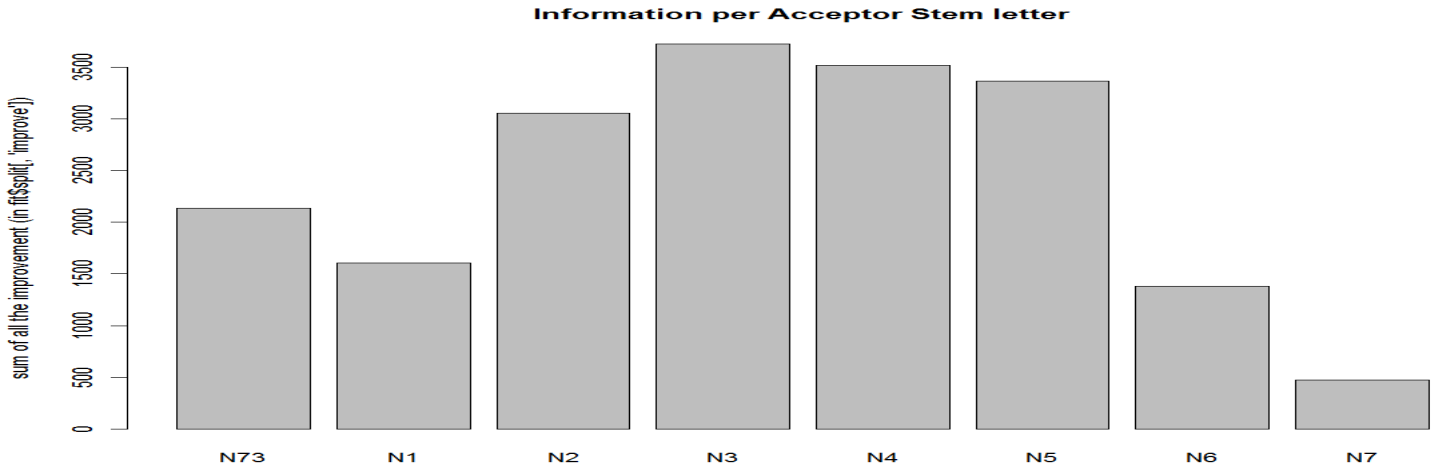


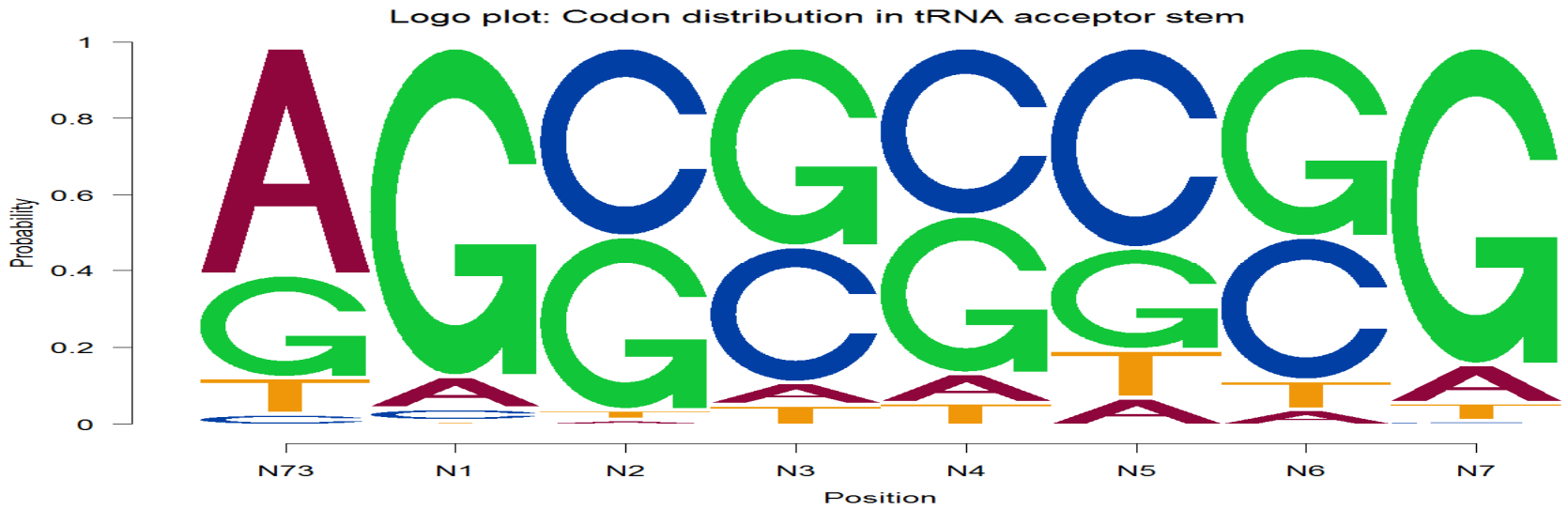
Relative importance for each letter in the model



(Sum of information gain in the use of each letter in the model)

Relative importance for each letter in the model



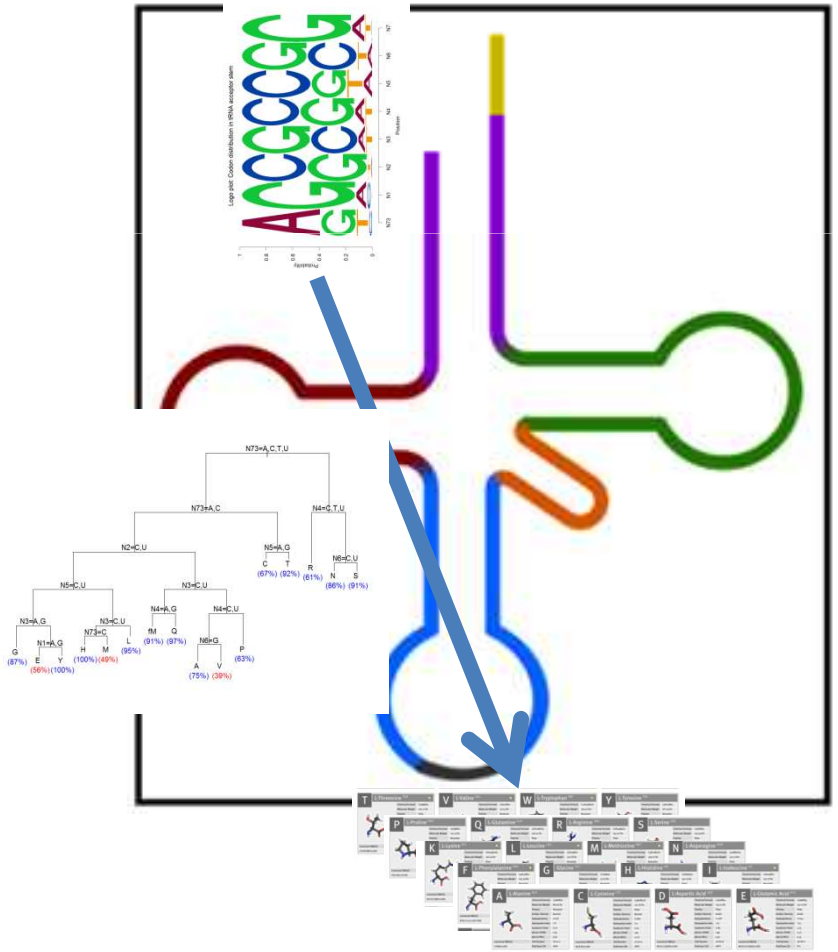


Results

- 20% misclassification in prediction
- Same acceptor stem RNA can code for for different Amino Acids
  - In **different** Archaeas
  - In the **same** Archaea

Conclusions

- There is code in the acceptor stem
- It is not enough



Thank you !

Read / Contact me:

R-statistics.com

Tal Galili[1], Shaul Shaul[2], Yoav Benjamini [3], Dror Berel [4]

[1] Department of Statistics and Operations Research, Tel Aviv University, Israel.
[2] Department of Zoology, Tel Aviv University, Israel.
[3] Department of Statistics and Operations Research, Tel Aviv University, Israel.
[4] Cedars-Sinai Medical Center.