

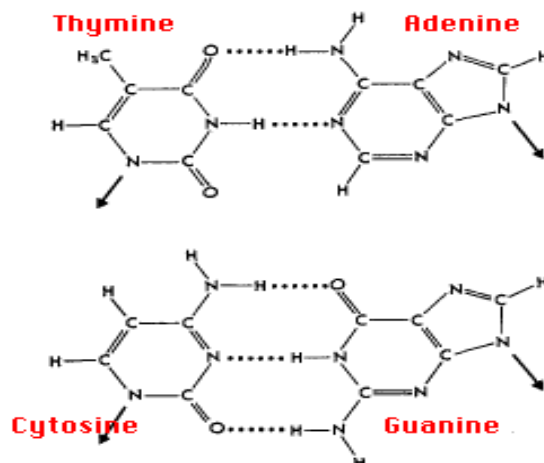
BIOLOGICAL SEQUENCES
BRIEFLY ANNOTATED DEFINITIONS AND CONCEPTS
PART I

MARLOS VIANA

1. BIOLOGICAL SEQUENCES

- 1.1. A *biological sequence* is a string of symbols from a finite alphabet (\mathcal{A}) of residues, e.g., strings of
- nucleotides: adenine (A), guanine (G), thymine (T), cytosine (C) in DNA (deoxyribonucleic acid), $\mathcal{A} = \{A, G, T, C\}$;
 - nucleotides: adenine (A), guanine (G), cytosine (C), uracil (U) in RNA (ribonucleic acid), $\mathcal{A} = \{A, G, T, U\}$;
 - Purine (u=A or u=G), Pyrimidine (y=C or y=T) residues, $\mathcal{A} = \{u, y\}$;
 - amino acids: Alanine (A), Arginine (R), Asparagine (N), Aspartic (D), Cysteine (C), Glutamic (E), Glutamine (Q), Glycine (G), Histidine (H), Isoleucine (I), Leucine (L), Lysine (K), Methionine (M), Phenylalanine (F), Proline (P), Serine (S), Threonine (T), Tryptophan (W), Tyrosine (Y), Valine (V), in protein sequences, $\mathcal{A} = \{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$;
 - Magnitude range of global sequences, in base pairs: 10^3 (single-stranded virus), 10^9 (mammals);
- 1.2. Figure 1.1 illustrates¹ the base-pairing rules between purines and pyrimidines.

FIGURE 1.1. Watson-Crick rules for nucleotides pairing



Date: November 5, 2002.

Lectures notes prepared for the GRCR's Fall 2002 Biostatistics Rotation.

¹e.g., <http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/N/Nucleotides.html>.

the relative frequency of all 3-sequences,

$$\begin{bmatrix}
 aaa & aac & aag & aat & caa & cac & cag & cat \\
 aca & acc & acg & act & cca & ccc & ccg & cct \\
 aga & agc & agg & agt & cga & cgc & cgg & cgt \\
 ata & atc & atg & att & cta & ctc & ctg & ctt \\
 gaa & gac & gag & gat & taa & tac & tag & tat \\
 gca & gcc & gcg & gct & tca & tcc & tcg & tct \\
 gga & ggc & ggg & ggt & tga & tgc & tgg & tgt \\
 gta & gtc & gtg & gtt & tta & ttc & ttg & ttt
 \end{bmatrix}
 : \frac{1}{66}
 \begin{bmatrix}
 0 & 2 & 1 & 0 & 0 & 1 & 1 & 0 \\
 0 & 2 & 0 & 2 & 2 & 1 & 0 & 1 \\
 2 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 1 & 2 & 4 & 4 & 1 \\
 1 & 1 & 2 & 1 & 2 & 0 & 2 & 0 \\
 0 & 1 & 0 & 3 & 0 & 0 & 0 & 5 \\
 2 & 1 & 2 & 2 & 1 & 1 & 3 & 0 \\
 0 & 1 & 0 & 1 & 2 & 0 & 1 & 1
 \end{bmatrix},$$

and the relative frequency of all 4-sequences,

$$\begin{bmatrix}
 aaaa & aaac & aaag & aaat & acaa & acac & acag & acat & caaa & caac & caag & caat & ccaa & ccac & ccag & ccat \\
 aaca & aacc & aacg & aact & acca & accc & accg & acct & caca & cacc & cacg & cact & ccca & cccc & cccg & ccct \\
 aaga & aagc & aagg & aagt & acga & acgc & acgg & acgt & caga & cagc & cagg & cagt & ccga & ccgc & ccgg & ccgt \\
 aata & aatc & aatg & aatt & acta & actc & actg & actt & cata & catc & catg & catt & ccta & cctc & cctg & cctt \\
 agaa & agac & agag & agat & ataa & atac & atag & atat & cgaa & cgac & cgag & cgat & ctaa & ctac & ctag & ctat \\
 agca & agcc & agcg & agct & atca & atcc & atcg & atct & cgca & cgcc & cgcg & cgct & ctca & ctcc & cteg & ctct \\
 agga & aggc & aggg & aggt & atga & atgc & atgg & atgt & cgga & cgge & cggg & cggt & ctga & ctgc & ctgg & ctgt \\
 agta & agtc & agtg & agtt & atta & attc & attg & attt & cgta & cgtc & cgtg & cgtt & ctta & cttc & cttg & cttt \\
 gaaa & gaac & gaag & gaat & gcaa & gcac & gcag & gcat & taaa & taac & taag & taat & tcaa & tcac & tcag & tcat \\
 gaca & gacc & gacg & gact & gcca & gccc & gccg & gcct & taca & tacc & tacg & tact & tcca & tccc & tccg & tcct \\
 gaga & gagc & gagg & gagt & gcga & gcgc & gcgg & gcgt & taga & tage & tagg & tagt & tcga & tegc & tcgg & tegt \\
 gata & gate & gatg & gatt & gcta & gctc & gctg & gctt & tata & tate & tatg & tatt & tcta & tctc & tctg & tctt \\
 ggaa & ggac & ggag & ggat & gtaa & gtac & gtag & gtat & tgaa & tgac & tgag & tgat & ttaa & ttac & ttag & ttat \\
 ggca & ggcc & ggcg & ggct & gtca & gtcc & gtcg & gtct & tgca & tgcc & tgcg & tgct & ttca & ttcc & ttcg & ttct \\
 gggaa & gggc & gggg & gggg & gtga & gtgc & gtgg & gtgt & tggaa & tggc & tggg & tggg & ttga & ttgc & ttgg & ttgt \\
 ggta & ggtc & ggtg & ggtt & gtta & gttc & gttg & gttt & tgta & tgtc & tgtg & tggt & ttta & tttc & tttg & tttt
 \end{bmatrix} :$$

1.8. The human immunodeficiency virus type 1, isolate BRU. Here is a fragment of the entire nucleotide sequence. To locate the sequence in the NCBI³ data base, use the accession number K02013. Figures 1.2 to 1.5 illustrate the distributed frequencies of indicated nucleotides along the global sequences of two isolates (BRU and OYI) of the same virus. The BRU isolate is 9229 bp (base-pair) long and the OYI is 9102bp (accession number M26727).

LOCUS HIVBRUCG 2586 bp ss-RNA linear VRL 02-AUG-1993
 DEFINITION Human immunodeficiency virus type 1, isolate BRU, complete genome (LAV-1).
 ACCESSION K02013 REGION: 5803..8388

```

1 atgagagtga aggagaaata tcagcacttg tggagatggg ggtggaaatg gggcaccatg
61 ctcccttggga tattgatgat ctgtagtgtc acagaaaaat tgtgggtcac agtctattat
121 ggggtacctg tgtggaagga agcaaccacc actctattht gtgcatcaga tgctaaagca
181 tatgatacag aggtacataa tgtttgggcc acacatgcct gtgtaccac agacccaac
241 ccacaagaag tagtattggt aaatgtgaca gaaaatttta acatgtggaa aaatgacatg
301 gtagaacaga tgcattgagga tataatcagt ttatgggatc aaagcctaaa gccatgtgta
361 aaattaaccc cactctgtgt tagtttaaag tgcactgatt tggggaatgc tactaatacc
421 aatagtagta ataccaatag tagtagcggg gaaatgatga tggagaaagg agagataaaa
481 aactgctctt tcaatatcag cacaagcata agaggttaagg tgcagaaaga atatgcattt
541 ttttataaac ttgatataat accaatagat aatgatacta ccagctatac gttgacaagt
601 tgtaacacct cagtcattac acaggcctgt ccaaaggatc ctttgagcc aattcccata
661 cattattgtg ccccggctgg ttttgcgatt ctaaaatgta ataataagac gttcaatgga
721 acaggacat gtacaaatgt cagcacagta caatgtacac atggaattag gccagtagta
781 tcaactcaac tgctgttgaa tggcagtcta gcagaagaag aggtagtaat tagatctgcc
841 aatttcacag acaatgctaa aaccataata gtacagctga accaatctgt agaaattaat
901 tgtacaagac ccaacaaca tacaagaaa agtatccgta tccagagggg accagggaga
961 gcatttgtaa caatagaaa aataggaat atgagacaag cacattgtaa cattagtaga
1021 gcaaaaatgga atgccacttt aaaacagata gctagcaaat taagagaaca atttggaaat
1081 aataaaacaa taatctttaa gcaatcctca ggaggggacc cagaaattgt aacgcacagt
1141 ttttaattgtg gaggggaatt ttttactgt aattcaacac aactgtttaa tagtacttgg
1201 ttttaatagta cttggagtac tgaagggtca aataaactg aaggaagtga cacaatcaca
1261 ctccccatgca gaataaaaca atttataaac atgtggcagg aagtaggaaa agcaatgtat
1321 gcccctccca tcagcggaca aattagatgt tcatcaaata ttacagggct gctattaaca
1381 agagatggtg tgaataaaca caatgggtcc gagatcttca gacctggagg aggagatag
1441 agggacaatt ggagaagtga attatataaa tataaagtag taaaaattga accattagga
1501 gtagcaccca ccaaggcaaa gagaagagtg gtgcagagag aaaaagagc agtgggaata
1561 ggagctttgt tccttgggtt cttgggagca gcaggaagca ctatgggagc acggtcaatg
1621 acgctgacgg tacagccag acaattattg tctggtatag tgcagcagca gaacaattg
1681 ctgagggcta ttgagcgcga acagcatctg ttgcaactca cagtctgggg catcaagcag
1741 ctccaggcaa gaatcctggc tgtggaaaga tacctaagg atcaacagct cctggggatt
1801 tggggttgct ctggaaaact cttttgcacc actgctgtgc cttggaatgc tagttggagt
1861 aataaatctc tggaaacagat ttggaataac atgacctgga tggagtggga cagagaaat
1921 aacaattaca caagcttaat acattccta attgaagaat cgcaaaacca gcaagaaaag
1981 aatgaacaag aattattgga attagataaa tgggcaagtt tgtggaattg gtttaacata
2041 acaaatggc tgtggtatat aaaaatattc ataatgatag taggaggctt gtaggttta
2101 agaatagttt ttgctgtact ttctatagtg aatagagtta ggaggggata ttaccatta
2161 tcttttcaga cccacctccc aaccccgagg ggaccgaca ggcccgaagg aatagaagaa
2221 gaaggtggag agagagacag agacagatcc attcgattag tgaacggatc cttagcactt
2281 atctgggacg atctgaggag cctgtgcctc ttcagctacc accgcttgag agacttactc
2341 ttgattgtaa cgaggattgt ggaacttctg ggacgcaggg ggtgggaagc cctcaaatat
2401 tgggtggaatc tcctacagta ttggagttag gaactaaaga atagtgtgt tagcttgctc
2461 aatgccacag ccatagcagt agctgagggg acagataggg ttatagaagt agtacaagga
2521 gctttagtag ctattcgcca catacctaga agaataagac agggcttggga aaggattttg
2581 ctataa

```

³National Center for Biotechnology Information, <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide>

FIGURE 1.2. Adenine distribution in the BRU isolate K02013 (top) and OYI isolate M26727 (bottom).

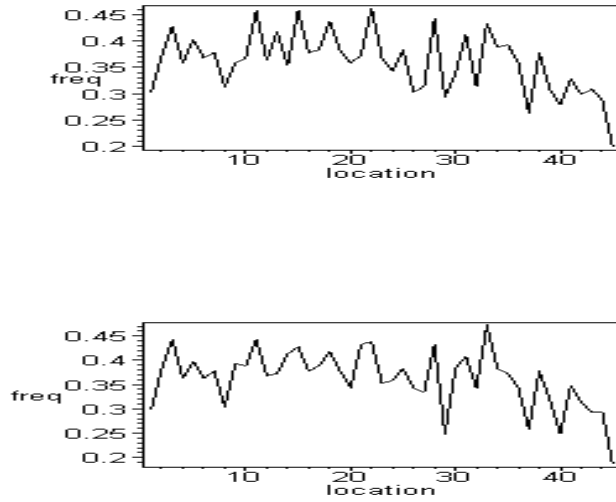


FIGURE 1.3. Cytosine distribution in the BRU isolate K02013 (top) and OYI isolate M26727 (bottom).

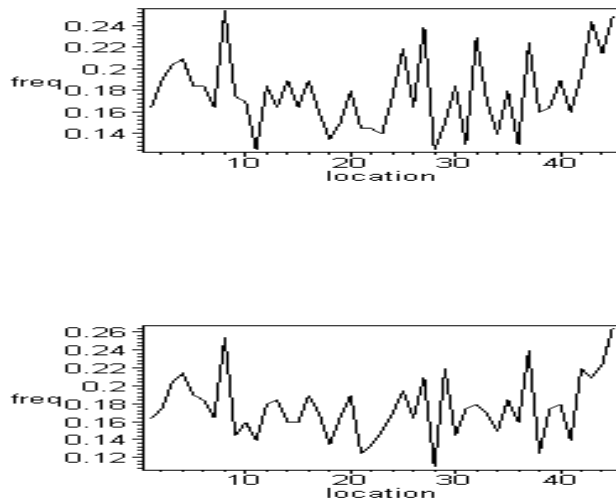


FIGURE 1.4. Guanine distribution in the BRU isolate K02013 (top) and OYI isolate M26727 (bottom).

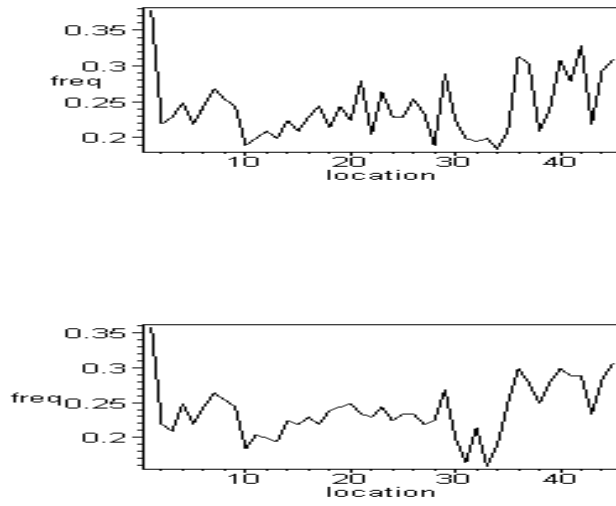
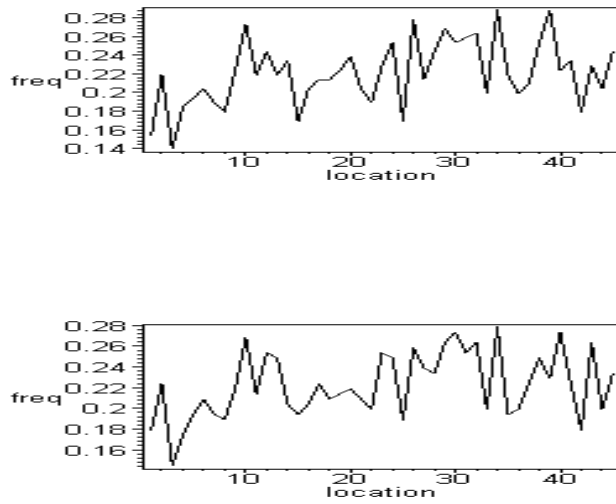


FIGURE 1.5. Thymine distribution in the BRU isolate K02013 (top) and OYI isolate M26727 (bottom).



1.9. *Independent letters* probability models. Given a probability model, π , in \mathcal{A} , then, defining

$$P(s) = \pi(s(1)) \times \pi(s(2)) \times \dots \times \pi(s(\ell)),$$

gives a probability model in \mathcal{A}^ℓ , the space of all ℓ bases long sequences in length of $|\mathcal{A}|$;

1.10. The *generating function* for the independent letters probability model in \mathcal{A}^ℓ is

$$(x_1 + \dots + x_c)^\ell,$$

from which $P(s)$ is shown to be well-defined;

1.11. *Example:* Consider the space of all 3-letter sequences in length of 2 (that is $|\mathcal{A}| = 3$, $\ell = 2$). There are $3^2 = 9$ sequences in \mathcal{A}^ℓ . The generating function, with, say $\mathcal{A} = \{a, b, c\}$, is

$$(a + b + c)^2 = a^2 + 2ab + 2ac + b^2 + 2bc + c^2,$$

which has 9 distinct (up to a permutation of the factors) coefficients, one for each sequence member of \mathcal{A}^ℓ . Now replace the letters $\{a, b, c\}$ by their corresponding probabilities, and we see that in fact $P(s)$, as defined above in Comment 1.9, is a probability law in \mathcal{A}^ℓ ;

1.12. Matrices 1.1 and 1.2 show the observed probability distribution of four-sequences in lengths of one and two, respectively, from the BRU isolate K02013.

$$(1.1) \quad \mathcal{L}_1 = [\ 0.356329 \quad 0.179520 \quad 0.242000 \quad 0.222149 \],$$

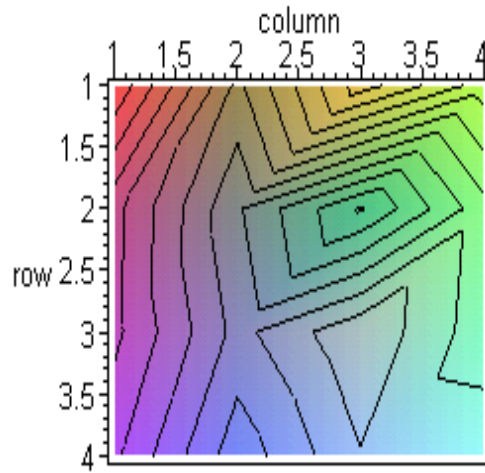
$$(1.2) \quad \mathcal{L}_2 = \begin{bmatrix} & \text{A} & \text{C} & \text{G} & \text{T} \\ \text{A} & 0.118680 & 0.057496 & 0.105228 & 0.074853 \\ \text{C} & 0.083098 & 0.041332 & 0.008787 & 0.046322 \\ \text{G} & 0.082555 & 0.046647 & 0.068453 & 0.044369 \\ \text{T} & 0.072032 & 0.034063 & 0.059448 & 0.056628 \end{bmatrix}.$$

The probability distribution \mathcal{L}_2 is also represented in Figure 1.6.

1.13. Matrix 1.3 shows the observed probability distribution of four-sequences in lengths of three, from the BRU isolate K02013, with entries given by

$$\begin{bmatrix} aaa & aac & aag & aat & | & caa & cac & cag & cat \\ aca & acc & acg & act & | & cca & ccc & ccg & cct \\ aga & agc & agg & agt & | & cga & cgc & cgg & cgt \\ ata & atc & atg & att & | & cta & ctc & ctg & ctt \\ \hline gaa & gac & gag & gat & | & taa & tac & tag & tat \\ gca & gcc & gcg & gct & | & tca & tcc & tcg & tct \\ gga & ggc & ggg & ggt & | & tga & tgc & tgg & tgt \\ gta & gtc & gtg & gtt & | & tta & ttc & ttg & ttt \end{bmatrix}.$$

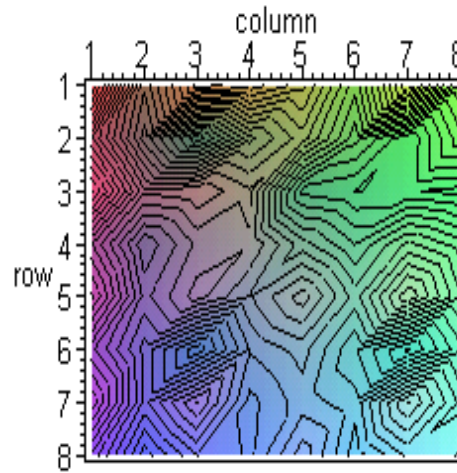
FIGURE 1.6. Observed contours of constant probability of four-sequences in lengths of two, from the BRU isolate K02013.



$$(1.3) \quad \mathcal{L}_3 = \begin{bmatrix} 0.042422 & 0.016383 & 0.030704 & 0.029185 & | & 0.024737 & 0.012694 & 0.029945 & 0.015623 \\ 0.028426 & 0.013128 & 0.0024954 & 0.013453 & | & 0.019095 & 0.0087881 & 0.0020614 & 0.011392 \\ 0.038624 & 0.022458 & 0.023760 & 0.020397 & | & 0.0033634 & 0.0019529 & 0.0022784 & 0.0011934 \\ 0.024194 & 0.012368 & 0.018444 & 0.019855 & | & 0.012911 & 0.0079202 & 0.014538 & 0.010958 \\ \hline 0.030487 & 0.015298 & 0.020940 & 0.015840 & | & 0.021048 & 0.013128 & 0.023652 & 0.014213 \\ 0.020723 & 0.011175 & 0.0026039 & 0.012152 & | & 0.014864 & 0.0082456 & 0.0016274 & 0.0093306 \\ 0.027558 & 0.011500 & 0.020072 & 0.0093306 & | & 0.013019 & 0.010741 & 0.022242 & 0.013453 \\ 0.017793 & 0.0048823 & 0.012043 & 0.0096561 & | & 0.017142 & 0.0088966 & 0.014430 & 0.016166 \end{bmatrix}$$

The distribution is also illustrated in Figure 1.7.

FIGURE 1.7. Observed contours of constant probability of four-sequences in lengths of three, from the BRU isolate K02013.



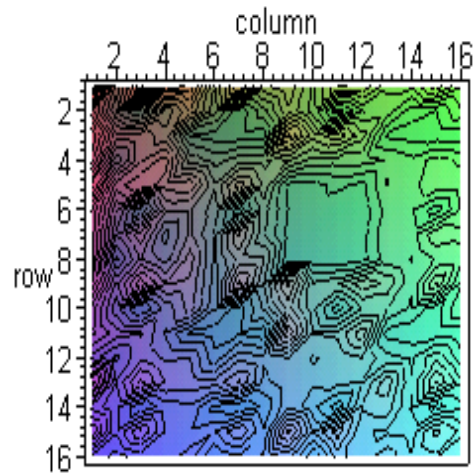
1.14. Matrix 1.4 shows the observed probability distribution of four-sequences in lengths of four, from the BRU isolate K02013, with entries given by

<i>aaaa</i>	<i>aaac</i>	<i>aaag</i>	<i>aaat</i>	<i>acaa</i>	<i>acac</i>	<i>acag</i>	<i>acat</i>	<i>caaa</i>	<i>caac</i>	<i>caag</i>	<i>caat</i>	<i>ccaa</i>	<i>ccac</i>	<i>ccag</i>	<i>ccat</i>
<i>aaca</i>	<i>aacc</i>	<i>aacg</i>	<i>aact</i>	<i>acca</i>	<i>accc</i>	<i>accg</i>	<i>acct</i>	<i>caca</i>	<i>cacc</i>	<i>cacg</i>	<i>cact</i>	<i>ccca</i>	<i>cccc</i>	<i>cccg</i>	<i>ccct</i>
<i>aaga</i>	<i>aagc</i>	<i>aagg</i>	<i>aagt</i>	<i>acga</i>	<i>acgc</i>	<i>acgg</i>	<i>acgt</i>	<i>caga</i>	<i>cagc</i>	<i>cagg</i>	<i>cagt</i>	<i>ccga</i>	<i>ccgc</i>	<i>ccgg</i>	<i>ccgt</i>
<i>aata</i>	<i>aatc</i>	<i>aatg</i>	<i>aatt</i>	<i>acta</i>	<i>actc</i>	<i>actg</i>	<i>actt</i>	<i>cata</i>	<i>catc</i>	<i>catg</i>	<i>catt</i>	<i>ccta</i>	<i>cctc</i>	<i>cctg</i>	<i>cctt</i>
<i>agaa</i>	<i>agac</i>	<i>agag</i>	<i>agat</i>	<i>ataa</i>	<i>atac</i>	<i>atag</i>	<i>atat</i>	<i>cgaa</i>	<i>cgac</i>	<i>cgag</i>	<i>cgat</i>	<i>ctaa</i>	<i>ctac</i>	<i>ctag</i>	<i>ctat</i>
<i>agca</i>	<i>agcc</i>	<i>agcg</i>	<i>agct</i>	<i>atca</i>	<i>atcc</i>	<i>atcg</i>	<i>atct</i>	<i>cgca</i>	<i>cgcc</i>	<i>cgcg</i>	<i>cgct</i>	<i>ctca</i>	<i>ctcc</i>	<i>ctcg</i>	<i>ctct</i>
<i>agga</i>	<i>aggc</i>	<i>aggg</i>	<i>aggt</i>	<i>atga</i>	<i>atgc</i>	<i>atgg</i>	<i>atgt</i>	<i>cgga</i>	<i>cggc</i>	<i>cggg</i>	<i>cggt</i>	<i>ctga</i>	<i>ctgc</i>	<i>ctgg</i>	<i>ctgt</i>
<i>agta</i>	<i>agtc</i>	<i>agtg</i>	<i>agtt</i>	<i>atta</i>	<i>attc</i>	<i>attg</i>	<i>attt</i>	<i>cgta</i>	<i>cgtc</i>	<i>cgtg</i>	<i>cgtt</i>	<i>ctta</i>	<i>cttc</i>	<i>cttg</i>	<i>cttt</i>
<i>gaaa</i>	<i>gaac</i>	<i>gaag</i>	<i>gaat</i>	<i>gcaa</i>	<i>gcac</i>	<i>gcag</i>	<i>gcat</i>	<i>taaa</i>	<i>taac</i>	<i>taag</i>	<i>taat</i>	<i>tcaa</i>	<i>tcac</i>	<i>tcag</i>	<i>tcat</i>
<i>gaca</i>	<i>gacc</i>	<i>gacg</i>	<i>gact</i>	<i>gcca</i>	<i>gccc</i>	<i>gccg</i>	<i>gctt</i>	<i>taca</i>	<i>tacc</i>	<i>tacg</i>	<i>tact</i>	<i>tcca</i>	<i>tccc</i>	<i>tccg</i>	<i>tcct</i>
<i>gaga</i>	<i>gagc</i>	<i>gagg</i>	<i>gagt</i>	<i>gcga</i>	<i>gcgc</i>	<i>gcgg</i>	<i>gcgt</i>	<i>taga</i>	<i>tage</i>	<i>tagg</i>	<i>tagt</i>	<i>tcga</i>	<i>tcgc</i>	<i>tcgg</i>	<i>tcgt</i>
<i>gata</i>	<i>gatc</i>	<i>gatg</i>	<i>gatt</i>	<i>gcta</i>	<i>gctc</i>	<i>gctg</i>	<i>gctt</i>	<i>tata</i>	<i>tatc</i>	<i>tatg</i>	<i>tatt</i>	<i>tcta</i>	<i>tctc</i>	<i>tctg</i>	<i>tctt</i>
<i>ggaa</i>	<i>ggac</i>	<i>ggag</i>	<i>ggat</i>	<i>gtaa</i>	<i>gtac</i>	<i>gtag</i>	<i>gtat</i>	<i>tgaa</i>	<i>tgac</i>	<i>tgag</i>	<i>tgat</i>	<i>ttaa</i>	<i>ttac</i>	<i>ttag</i>	<i>ttat</i>
<i>ggca</i>	<i>ggcc</i>	<i>ggcg</i>	<i>ggct</i>	<i>gtca</i>	<i>gtcc</i>	<i>gtcg</i>	<i>gtct</i>	<i>tgca</i>	<i>tgcc</i>	<i>tgcg</i>	<i>tgct</i>	<i>ttca</i>	<i>ttcc</i>	<i>ttcg</i>	<i>ttct</i>
<i>ggga</i>	<i>gggc</i>	<i>gggg</i>	<i>gggt</i>	<i>gtga</i>	<i>gtgc</i>	<i>gtgg</i>	<i>gtgt</i>	<i>tgga</i>	<i>tggc</i>	<i>tggg</i>	<i>tggt</i>	<i>ttga</i>	<i>ttgc</i>	<i>ttgg</i>	<i>ttgt</i>
<i>ggta</i>	<i>ggtc</i>	<i>ggtg</i>	<i>ggtt</i>	<i>gtta</i>	<i>gttc</i>	<i>gttg</i>	<i>gttt</i>	<i>tgta</i>	<i>tgtc</i>	<i>tgtg</i>	<i>tgtt</i>	<i>ttta</i>	<i>tttc</i>	<i>tttg</i>	<i>tttt</i>

$$(1.4) \quad \mathcal{L}_4 = \frac{1}{9226} \begin{bmatrix} 152 & 42 & 102 & 97 & 91 & 36 & 83 & 52 & 63 & 36 & 65 & 64 & 39 & 38 & 60 & 38 \\ 78 & 27 & 8 & 38 & 64 & 31 & 2 & 24 & 54 & 33 & 3 & 27 & 33 & 13 & 9 & 26 \\ 111 & 60 & 62 & 50 & 9 & 5 & 5 & 4 & 96 & 59 & 71 & 50 & 8 & 5 & 4 & 2 \\ 97 & 30 & 63 & 79 & 34 & 24 & 33 & 33 & 47 & 33 & 32 & 32 & 27 & 21 & 32 & 26 \\ 136 & 69 & 83 & 68 & 72 & 38 & 70 & 44 & 7 & 6 & 12 & 6 & 28 & 24 & 35 & 32 \\ 102 & 48 & 9 & 48 & 47 & 32 & 4 & 32 & 9 & 5 & 1 & 3 & 22 & 20 & 4 & 28 \\ 97 & 36 & 62 & 24 & 48 & 26 & 66 & 29 & 6 & 4 & 6 & 5 & 27 & 23 & 44 & 40 \\ 94 & 16 & 47 & 32 & 60 & 26 & 50 & 47 & 1 & 3 & 2 & 5 & 23 & 18 & 29 & 31 \\ 101 & 44 & 80 & 56 & 54 & 23 & 77 & 37 & 77 & 29 & 36 & 52 & 44 & 20 & 56 & 17 \\ 71 & 31 & 8 & 31 & 50 & 20 & 4 & 29 & 59 & 30 & 4 & 28 & 28 & 17 & 4 & 27 \\ 68 & 51 & 47 & 27 & 8 & 5 & 8 & 3 & 81 & 37 & 39 & 62 & 6 & 3 & 4 & 2 \\ 43 & 25 & 44 & 35 & 37 & 12 & 34 & 29 & 37 & 27 & 30 & 37 & 21 & 17 & 35 & 13 \\ 104 & 37 & 67 & 46 & 42 & 40 & 57 & 25 & 34 & 29 & 31 & 27 & 52 & 19 & 57 & 30 \\ 44 & 29 & 10 & 23 & 26 & 6 & 1 & 11 & 36 & 21 & 4 & 38 & 43 & 18 & 6 & 15 \\ 73 & 29 & 56 & 27 & 20 & 27 & 46 & 19 & 78 & 37 & 61 & 29 & 26 & 23 & 49 & 35 \\ 27 & 15 & 22 & 22 & 32 & 10 & 18 & 29 & 42 & 10 & 41 & 30 & 43 & 28 & 36 & 44 \end{bmatrix}$$

The distribution is also illustrated in Figure 1.8.

FIGURE 1.8. Observed contours of constant probability of four-sequences in lengths of four, from the BRU isolate K02013.



1.15. *Partition-dependent* probability models, P_λ , may be generated according to

$$(x_{11} + \dots + x_{1c})^{\lambda_1} \times \dots \times (x_{\ell 1} + \dots + x_{\ell c})^{\lambda_\ell},$$

where $\lambda = (\lambda_1, \dots, \lambda_\ell)$ is any integer partition of ℓ , $c = |\mathcal{A}|$, and

$$(x_{i1}, x_{i2}, \dots, x_{ic}), \quad i = 1, \dots, \ell$$

are probability laws in \mathcal{A} ;

1.16. *Example* of partition-dependent models for two-sequences in length of three ($|\mathcal{A}| = 2, \ell = 3$):

- $\lambda = (3, 0, 0) :$ $P_\lambda(s) = \pi_1(s(1)) \times \pi_1(s(2)) \times \pi_1(s(3))$,
with generating function $(x_{11} + x_{12})^3$;
- $\lambda = (2, 1, 0) :$ $P_\lambda(s) = \pi_1(s(1)) \times \pi_1(s(2)) \times \pi_2(s(3))$,
with generating function $(x_{11} + x_{12})^2(x_{21} + x_{22})$;
- $\lambda = (1, 1, 1) :$ $P_\lambda(s) = \pi_1(s(1)) \times \pi_2(s(2)) \times \pi_3(s(3))$,
with generating function $(x_{11} + x_{12})(x_{21} + x_{22})(x_{31} + x_{32})$;

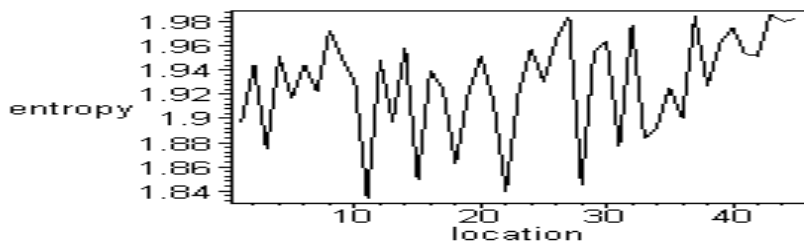
where $\pi_i = (x_{i1}, x_{i2})$, $i = 1, 2, 3$ are distinct probability laws in \mathcal{A} .

1.17. The *entropy* of the probability law⁴ P is

$$H(P) = - \sum_s P(s) \log_2 P(s) \leq \log_2 |\mathcal{A}|,$$

which attains its maximum value when the law of P is uniform. In this sense, the entropy of the law is a measure of its *uniformity*⁵. Figure 1.9 shows the $\{a, c, g, t\}$ entropy along the sequence for the BRU isolate of the human immunodeficiency virus type 1, HIV-1, (K02013). The *information content* at a given region may be defined as $\log_2 |\mathcal{A}| - H(P)$ bits, or, in this case ($|\mathcal{A}| = 4$), as $2 - H(P)$ bits. The higher the information content the more *conserved* the local sequence;

FIGURE 1.9. Four-base entropy along the BRU isolate K02013.



⁴The base of the logarithm is irrelevant- when the base is 2 the entropy is usually expressed in units called *bits*.

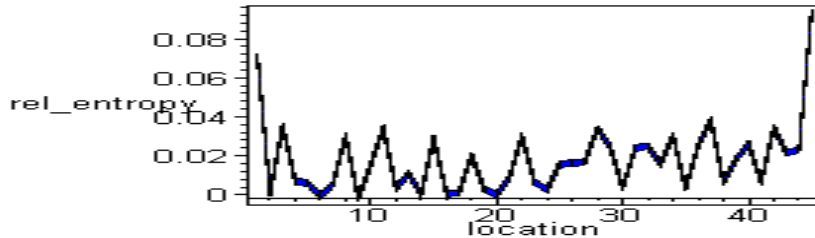
⁵Historically, the notion of entropy was central to the 2nd law of thermodynamics, in which the entropy of a system of gas molecules left to itself almost always increases (towards uniformity).

1.18. The *relative entropy* of laws P and Q is given by

$$H(P | Q) = \sum_s P(s) \log_2 \frac{P(s)}{Q(s)} \geq 0,$$

which reduces, when the law of Q is uniform, to the information content of P. Moreover, $H(P | Q) = 0$ if and only if $P = Q$;

FIGURE 1.10. Four-base self-relative entropy along the BRU isolate K02013.



1.19. The *mutual information* is the relative entropy

$$H(P_{sf} | P_s P_f) \leq H(P_s) = H(P_s)$$

between the joint probability model (P_{sf}) and the product of their marginal laws ($P_s P_f$). The mutual information is maximum when the two laws covary the most⁶.

2. SEQUENCE CHARACTERIZATION BY SYMMETRIES.

Characterizing a sequence by symmetries is like searching for *mosaic*-like patterns in the sequence. A set of objects or labels share a similarity relation by symmetry when these objects remain indifferent, invariant or constant under a set of transformations. The transformations defining symmetries are simple *permutations*, or one-to-one mappings onto a set of ℓ objects or labels. Permutations, together with the operation of composition of functions, have the algebraic properties of finite *group*⁷. The set of all permutations of ℓ objects is usually indicated by S_ℓ .

2.1. *Example* The set S_3 of all permutations of 3 symbols includes the identity (1) transformation

$$1 = \begin{bmatrix} 1 \rightarrow 1 \\ 2 \rightarrow 2 \\ 3 \rightarrow 3 \end{bmatrix},$$

three transpositions,

$$(12) = \begin{bmatrix} 1 \rightarrow 2 \\ 2 \rightarrow 1 \\ 3 \rightarrow 3 \end{bmatrix}, \quad (13) = \begin{bmatrix} 1 \rightarrow 3 \\ 2 \rightarrow 2 \\ 3 \rightarrow 1 \end{bmatrix}, \quad (23) = \begin{bmatrix} 1 \rightarrow 1 \\ 2 \rightarrow 3 \\ 3 \rightarrow 2 \end{bmatrix},$$

⁶This occurs when the pattern of the joint probabilities is similar to a permutation matrix (with a probability distribution replacing the unit entries), in which case $P_{sf} = P_s = P_{s\tau}$, for a permutation τ

⁷The group properties say that the composing (i.e., one followed by the other) of two permutations is again a permutation, that there is a identity permutation, that to each permutation there corresponds a unique inverse permutation, and that composing is associative.

and two cyclic permutations,

$$(123) = \begin{bmatrix} 1 \rightarrow 2 \\ 2 \rightarrow 3 \\ 3 \rightarrow 1 \end{bmatrix}, \quad (132) = \begin{bmatrix} 1 \rightarrow 3 \\ 2 \rightarrow 1 \\ 3 \rightarrow 2 \end{bmatrix}.$$

In summary,

$$S_3 = \{1, (12), (13), (23), (123), (132)\}.$$

The set C_3 of all cyclic permutations of 3 symbols:

$$C_3 = \{1, (123), (132)\}.$$

2.2. *Composing sequences and symmetries.* Given a sequence s in length of ℓ and a symmetry τ in S_ℓ then the composite $s\tau$ is also a sequence in length of ℓ . Say $\ell = 4$ (and $\mathcal{A} = \{A, C, G, T\}$), and

$$\tau = \begin{bmatrix} 1 \rightarrow 2 \\ 2 \rightarrow 3 \\ 3 \rightarrow 4 \\ 4 \rightarrow 1 \end{bmatrix}, \quad s = \begin{bmatrix} 1 \rightarrow A \\ 2 \rightarrow A \\ 3 \rightarrow G \\ 4 \rightarrow C \end{bmatrix}. \quad \text{Then, } s\tau = \begin{bmatrix} 1 \rightarrow A \\ 2 \rightarrow G \\ 3 \rightarrow C \\ 4 \rightarrow A \end{bmatrix};$$

2.3. *Cyclic symmetries.* Given a local sequence s in length of ℓ , define the *cyclic orbit*

$$[s] = \{s\tau; \tau \in C_\ell\},$$

so that, for example,

$$[CGG] = \{CGG, GCG, GGC\}, \quad [uuyuyu] = \{uuyuyu, yuuyuu, uyuyuu\};$$

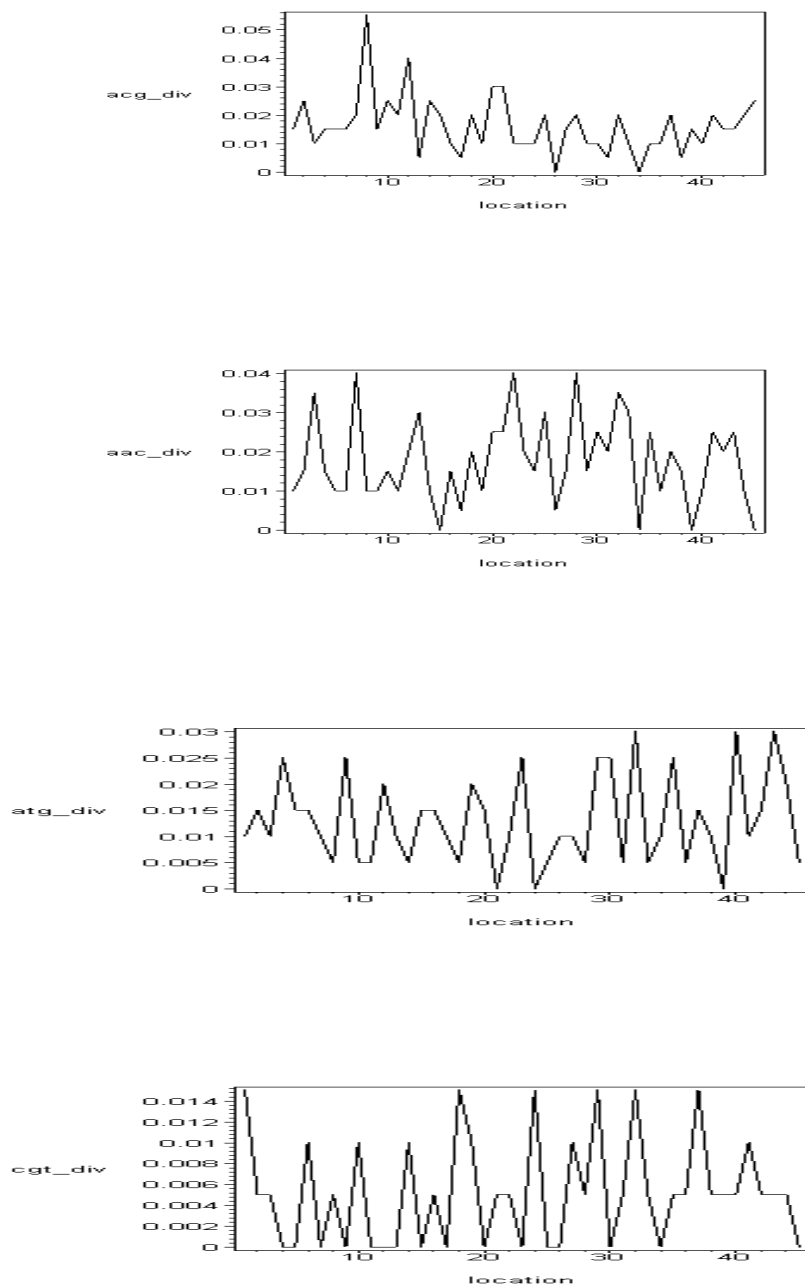
2.4. In Doi (1991), the *frequency diversity* of $[s]$ is defined as the ratio

$$\frac{\max_{f \in [s]} \widehat{f}}{\min_{f \in [s]} \widehat{f}},$$

where \widehat{f} is the observed relative frequency of sequence f .

2.5. *Example.* Figure 2.1 shows the observed cyclic diversity for the cyclic orbits $[agc]$, $[aac]$, $[atg]$ and $[cgt]$ along the BRU isolate K02013.

FIGURE 2.1. Diversity, expressed as the range $\max_{f \in [s]} \hat{f} - \min_{f \in [s]} \hat{f}$, of selected cyclic sets (indicated in the y-axis: ACG, AAC, ATG, CGT, respectively) along the BRU isolate K02013.



2.6. *Baseline variation.* The independent letters model of Comment 1.9 implies that $P(s) = P(s\tau)$ for all permutation $\tau \in S_\ell$, which implies that P is constant in each one of the *orbits*

$$[s] = \{s\tau; \tau \in S_\ell\}.$$

For two-sequences, say $\mathcal{A} = \{u, y\}$, these orbits are defined by collecting together the sequences with the same number of, say, purines (u). There are, in this case, $\ell + 1$ orbits, as the number of purines ranges from 0 to $\ell + 1$. The uniformity within each orbit may serve as a baseline or reference variation. Figure 2.4 illustrates the argument with purine (u)-pyrimidine(y) sequences in length of two. The orbits of interest are the single purine u_1 and the single pyrimidine u_2 ones, that is,

$$u_1 = \{uyy, yuy, yyu\}, \quad u_2 = \{yuu, uyu, uuy\},$$

in addition to the trivial ones $u_0 = \{yyy\}$ and $u_3 = \{uuu\}$. Figures 2.4 and 2.5 illustrate the relative frequency ratios and diversity within each one of the two orbits. The ratios and diversities imply the inadequacy of the independent letters model.

FIGURE 2.2. Relative frequency ratios f_{yyu}/f_{uyy} (red, steady curve) and f_{yyu}/f_{yuy} (green, variable curve) in the single-purine orbit u_1 (top) and corresponding orbit diversity (bottom), along the BRU isolate K02013.

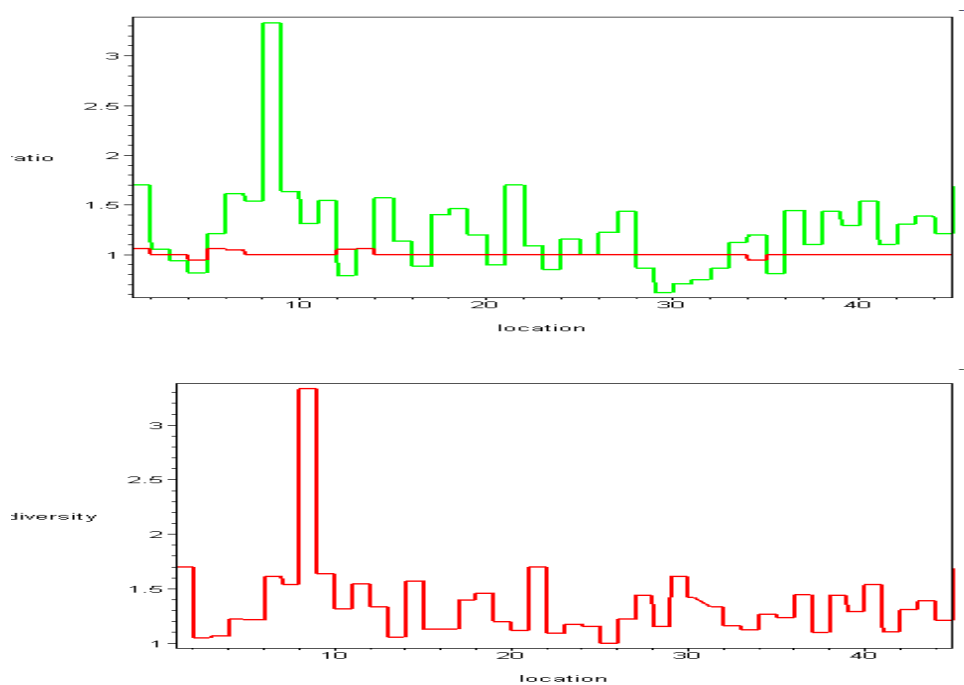


FIGURE 2.3. Relative frequency ratios f_{uuy}/f_{yuu} (red, steady curve) and f_{uuy}/f_{uyu} (green, variable curve) in the single-pyrimidine orbit u_2 (top) and corresponding orbit diversity (bottom), along the BRU isolate K02013.

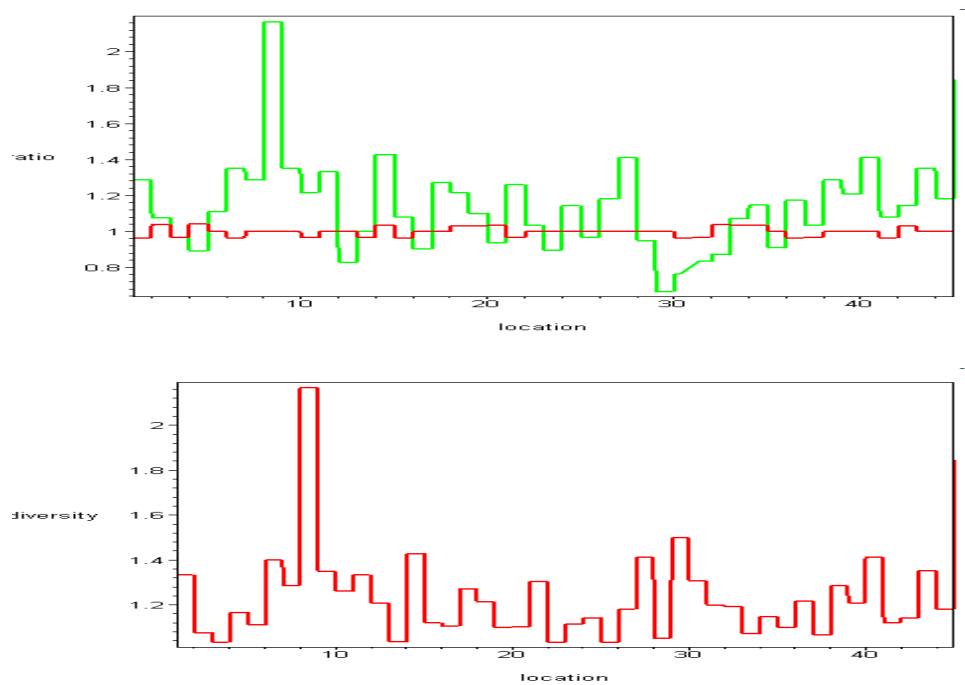


FIGURE 2.4. Relative frequency ratios f_{yyu}/f_{uyy} (green, steady curve) and f_{yyu}/f_{yuy} (red, variable curve) in the single-purine orbit u_1 (top) and corresponding orbit diversity (bottom), along the isolate M26727.

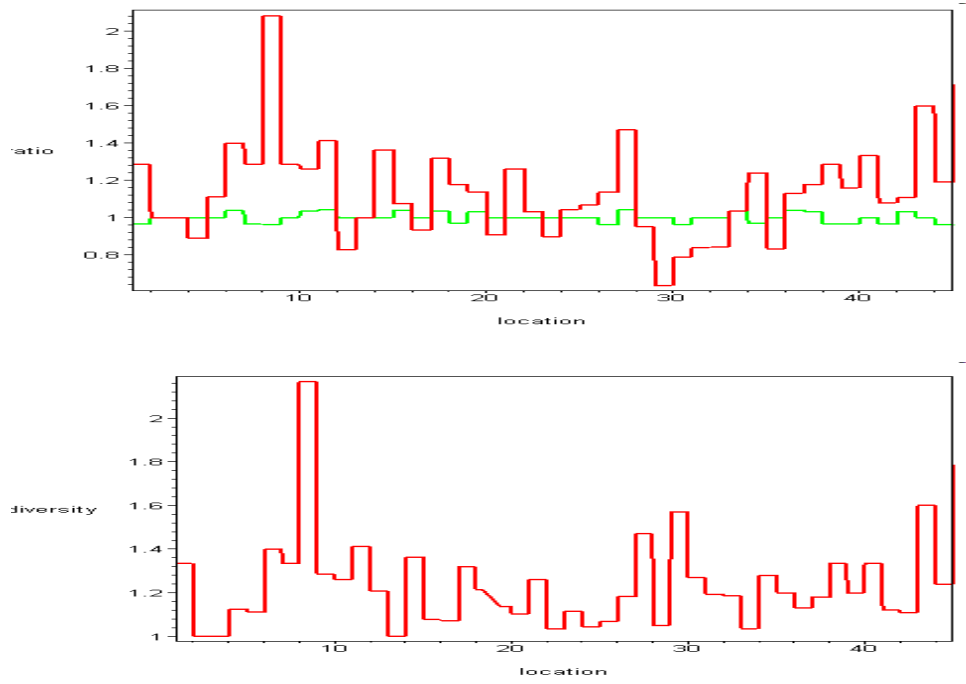
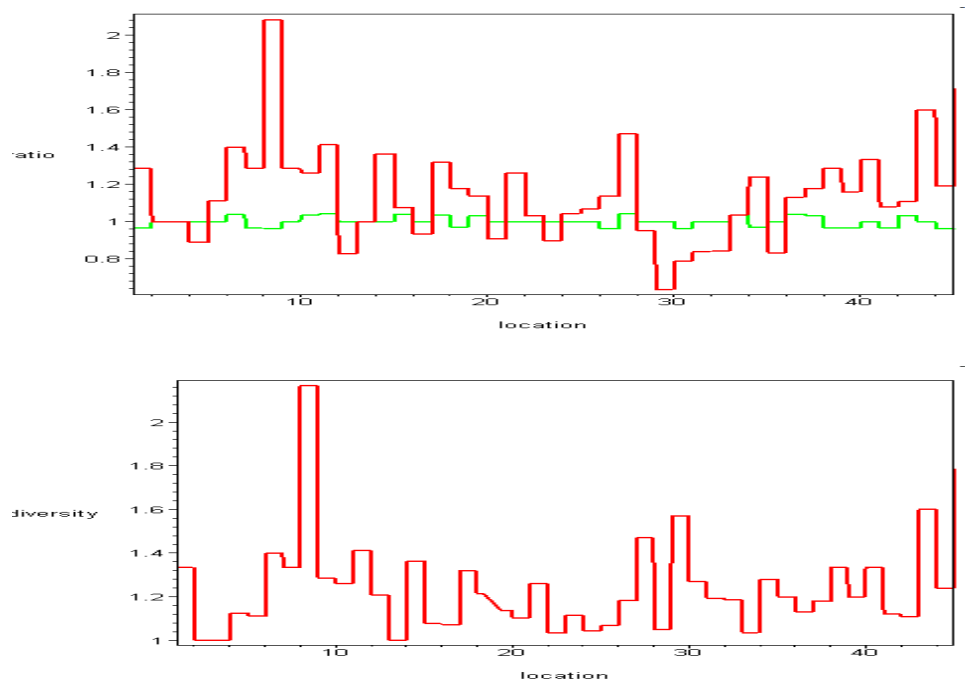


FIGURE 2.5. Relative frequency ratios f_{uuu}/f_{yuu} (green, steady curve) and f_{uuu}/f_{uyu} (red, variable curve) in the single-pyrimidine orbit u_2 (top) and corresponding orbit diversity (bottom), along the isolate M26727.



Example 2.1. *Permutation orbits* for 2-sequences in length of 4. The map space $V = C^L$ has 16 points, each representing one sequence, namely

$$V = \left[\begin{array}{c|cccccccccccccccc} s & 1 & 16 & 15 & 14 & 12 & 8 & 13 & 11 & 7 & 10 & 6 & 4 & 9 & 5 & 3 & 2 \\ \hline s(1) & y & u & y & u & u & u & y & y & y & u & u & u & y & y & y & u \\ s(2) & y & u & u & y & u & u & y & u & u & y & y & u & y & y & u & y \\ s(3) & y & u & u & u & y & u & u & y & u & y & u & y & y & u & y & y \\ s(4) & y & u & u & u & u & y & u & u & y & u & y & y & u & y & y & y \end{array} \right].$$

Consider the left regular action of S_4 on V with $\ell = 4, c = 2$. Recall that S_4 has 6 transpositions, 3 cycles of order 2, 8 cycles of order 3 and 6 cycles of order 4. These permutations are indicated in the first column of matrix (2.1). The resulting actions are shown in the adjacent 16 columns.

$$(2.1) \quad \left[\begin{array}{c|cccccccccccccccc} S_4 \backslash s & 1 & 16 & 15 & 14 & 12 & 8 & 13 & 11 & 7 & 10 & 6 & 4 & 9 & 5 & 3 & 2 \\ \hline 1 & 1 & 16 & 15 & 14 & 12 & 8 & 13 & 11 & 7 & 10 & 6 & 4 & 9 & 5 & 3 & 2 \\ (34) & 1 & 16 & 15 & 14 & 8 & 12 & 13 & 7 & 11 & 6 & 10 & 4 & 5 & 9 & 3 & 2 \\ (23) & 1 & 16 & 15 & 12 & 14 & 8 & 11 & 13 & 7 & 10 & 4 & 6 & 9 & 3 & 5 & 2 \\ (24) & 1 & 16 & 15 & 8 & 12 & 14 & 7 & 11 & 13 & 4 & 6 & 10 & 3 & 5 & 9 & 2 \\ (12) & 1 & 16 & 14 & 15 & 12 & 8 & 13 & 10 & 6 & 11 & 7 & 4 & 9 & 5 & 2 & 3 \\ (13) & 1 & 16 & 12 & 14 & 15 & 8 & 10 & 11 & 4 & 13 & 6 & 7 & 9 & 2 & 3 & 5 \\ (14) & 1 & 16 & 8 & 14 & 12 & 15 & 6 & 4 & 7 & 10 & 13 & 11 & 2 & 5 & 3 & 9 \\ \hline (234) & 1 & 16 & 15 & 12 & 8 & 14 & 11 & 7 & 13 & 4 & 10 & 6 & 3 & 9 & 5 & 2 \\ (243) & 1 & 16 & 15 & 8 & 14 & 12 & 7 & 13 & 11 & 6 & 4 & 10 & 5 & 3 & 9 & 2 \\ (123) & 1 & 16 & 14 & 12 & 15 & 8 & 10 & 13 & 6 & 11 & 4 & 7 & 9 & 2 & 5 & 3 \\ (124) & 1 & 16 & 14 & 8 & 12 & 15 & 6 & 10 & 13 & 4 & 7 & 11 & 2 & 5 & 9 & 3 \\ (132) & 1 & 16 & 12 & 15 & 14 & 8 & 11 & 10 & 4 & 13 & 7 & 6 & 9 & 3 & 2 & 5 \\ (134) & 1 & 16 & 12 & 14 & 8 & 15 & 10 & 4 & 11 & 6 & 13 & 7 & 2 & 9 & 3 & 5 \\ (142) & 1 & 16 & 8 & 15 & 12 & 14 & 7 & 4 & 6 & 11 & 13 & 10 & 3 & 5 & 2 & 9 \\ (143) & 1 & 16 & 8 & 14 & 15 & 12 & 6 & 7 & 4 & 13 & 10 & 11 & 5 & 2 & 3 & 9 \\ \hline (12)(34) & 1 & 16 & 14 & 15 & 8 & 12 & 13 & 6 & 10 & 7 & 11 & 4 & 5 & 9 & 2 & 3 \\ (13)(24) & 1 & 16 & 12 & 8 & 15 & 14 & 4 & 11 & 10 & 7 & 6 & 13 & 3 & 2 & 9 & 5 \\ (14)(23) & 1 & 16 & 8 & 12 & 14 & 15 & 4 & 6 & 7 & 10 & 11 & 13 & 2 & 3 & 5 & 9 \\ \hline (1234) & 1 & 16 & 14 & 12 & 8 & 15 & 10 & 6 & 13 & 4 & 11 & 7 & 2 & 9 & 5 & 3 \\ (1243) & 1 & 16 & 14 & 8 & 15 & 12 & 6 & 13 & 10 & 7 & 4 & 11 & 5 & 2 & 9 & 3 \\ (1324) & 1 & 16 & 12 & 8 & 14 & 15 & 4 & 10 & 11 & 6 & 7 & 13 & 2 & 3 & 9 & 5 \\ (1342) & 1 & 16 & 12 & 15 & 8 & 14 & 11 & 4 & 10 & 7 & 13 & 6 & 3 & 9 & 2 & 5 \\ (1432) & 1 & 16 & 8 & 15 & 14 & 12 & 7 & 6 & 4 & 13 & 11 & 10 & 5 & 3 & 2 & 9 \\ (1423) & 1 & 16 & 8 & 12 & 15 & 14 & 4 & 7 & 6 & 11 & 10 & 13 & 3 & 2 & 5 & 9 \end{array} \right]$$

The resulting orbits may be expressed as

$$\begin{aligned}\mathcal{O}_0 &= \{1\}, \\ \mathcal{O}_1 &= \{9, 5, 3, 2\}, \\ \mathcal{O}_2 &= \{13, 11, 7, 10, 6, 4\}, \\ \mathcal{O}_3 &= \{15, 14, 12, 8\}, \\ \mathcal{O}_4 &= \{16\},\end{aligned}$$

thus showing that the orbit \mathcal{O}_k may be characterized by the number of purines (symbols ‘u’) in the sequences, that is,

$$\mathcal{O}_k = \{s \in V; |s^{-1}(u)| = k\}, \quad k = 0, \dots, 4.$$

Note that

$$|\mathcal{O}_k| = \binom{\ell}{k}.$$

Example 2.2. Order 4 *Cyclic orbits* for 2-sequences in length of 4. Following Example 2.1 we now consider the regular left action of $C_4 = \{1, (1234), (13)(24), (1432)\}$ on the map space V . The resulting actions are shown in matrix (2.2):

$$(2.2) \quad \left[\begin{array}{c|cccccccccccccccc} C_4 \backslash s & 1 & 16 & 15 & 14 & 12 & 8 & 13 & 11 & 7 & 10 & 6 & 4 & 9 & 5 & 3 & 2 \\ \hline 1 & 1 & 16 & 15 & 14 & 12 & 8 & 13 & 11 & 7 & 10 & 6 & 4 & 9 & 5 & 3 & 2 \\ (13)(24) & 1 & 16 & 12 & 8 & 15 & 14 & 4 & 11 & 10 & 7 & 6 & 13 & 3 & 2 & 9 & 5 \\ (1234) & 1 & 16 & 14 & 12 & 8 & 15 & 10 & 6 & 13 & 4 & 11 & 7 & 2 & 9 & 5 & 3 \\ (1432) & 1 & 16 & 8 & 15 & 14 & 12 & 7 & 6 & 4 & 13 & 11 & 10 & 5 & 3 & 2 & 9 \end{array} \right]$$

The resulting orbits may be expressed as

$$\begin{aligned}\mathcal{O}_0 &= \{1\}, \\ \mathcal{O}_1 &= \{9, 5, 3, 2\}, \\ \mathcal{O}_{21} &= \{13, 7, 10, 4\}, \quad \mathcal{O}_{22} = \{11, 6\}, \\ \mathcal{O}_3 &= \{15, 14, 12, 8\}, \\ \mathcal{O}_4 &= \{16\},\end{aligned}$$

thus showing that C_4 splits the original \mathcal{O}_2 under S_4 into two new orbits, \mathcal{O}_{21} and \mathcal{O}_{22} , that is, $\mathcal{O}_{21} \cup \mathcal{O}_{22} = \mathcal{O}_2$ of Example 2.1. There are many more cyclic symmetries and orbit configurations (see Exercise nnn).

Example 2.3. *Dihedral orbits* for 2-sequences in length of 4. Following Examples 2.1 and 2.2 we now consider the regular left action of the group D_4 on the map space V . Recall that D_4 may be realized as the group of rotational and axial symmetries of the regular rectangle. The resulting actions are shown in matrix (2.3):

$$(2.3) \quad \left[\begin{array}{c|cccccccccccccccc} D_4 \backslash s & 1 & 16 & 15 & 14 & 12 & 8 & 13 & 11 & 7 & 10 & 6 & 4 & 9 & 5 & 3 & 2 \\ \hline 1 & 1 & 16 & 15 & 14 & 12 & 8 & 13 & 11 & 7 & 10 & 6 & 4 & 9 & 5 & 3 & 2 \\ (24) & 1 & 16 & 15 & 8 & 12 & 14 & 7 & 11 & 13 & 4 & 6 & 10 & 3 & 5 & 9 & 2 \\ (13) & 1 & 16 & 12 & 14 & 15 & 8 & 10 & 11 & 4 & 13 & 6 & 7 & 9 & 2 & 3 & 5 \\ (12)(34) & 1 & 16 & 14 & 15 & 8 & 12 & 13 & 6 & 10 & 7 & 11 & 4 & 5 & 9 & 2 & 3 \\ (13)(24) & 1 & 16 & 12 & 8 & 15 & 14 & 4 & 11 & 10 & 7 & 6 & 13 & 3 & 2 & 9 & 5 \\ (14)(23) & 1 & 16 & 8 & 12 & 14 & 15 & 4 & 6 & 7 & 10 & 11 & 13 & 2 & 3 & 5 & 9 \\ (1234) & 1 & 16 & 14 & 12 & 8 & 15 & 10 & 6 & 13 & 4 & 11 & 7 & 2 & 9 & 5 & 3 \\ (1432) & 1 & 16 & 8 & 15 & 14 & 12 & 7 & 6 & 4 & 13 & 11 & 10 & 5 & 3 & 2 & 9 \end{array} \right],$$

which shows that D_4 and C_4 generate the same set of orbits.

Example 2.4. Orbits generated by cyclic permutations of order 2, for 2-sequences in length of 4. Following Example 2.1 we now consider the regular left action of $G = \{1, (13)(24)\}$ on the map space V . The resulting actions are shown in matrix (2.4):

$$(2.4) \quad \left[\begin{array}{c|cccccccccccccccc} G \backslash s & 1 & 16 & 15 & 14 & 12 & 8 & 13 & 11 & 7 & 10 & 6 & 4 & 9 & 5 & 3 & 2 \\ \hline 1 & 1 & 16 & 15 & 14 & 12 & 8 & 13 & 11 & 7 & 10 & 6 & 4 & 9 & 5 & 3 & 2 \\ (13)(24) & 1 & 16 & 12 & 8 & 15 & 14 & 4 & 11 & 10 & 7 & 6 & 13 & 3 & 2 & 9 & 5 \end{array} \right]$$

The resulting orbits may be expressed as

$$\begin{aligned} \mathcal{O}_0 &= \{1\}, \\ \mathcal{O}_{11} &= \{9, 3\}, \quad \mathcal{O}_{12} = \{5, 2\}, \\ \mathcal{O}_{211} &= \{13, 4\}, \quad \mathcal{O}_{212} = \{7, 10\}, \quad \mathcal{O}_{221} = \{11\}, \quad \mathcal{O}_{222} = \{6\}, \\ \mathcal{O}_{31} &= \{14, 8\}, \quad \mathcal{O}_{32} = \{15, 12\}, \\ \mathcal{O}_4 &= \{16\}, \end{aligned}$$

thus showing that G further splits the original order-4 cyclic orbits into additional, smaller orbits.

3. JOINTLY DEPENDENT PROBABILITY MODELS

3.1. Given probability models, π_1, \dots, π_ℓ , in $\mathcal{A} \times \mathcal{A}$, then

$$P(\mathbf{s}, \mathbf{f}) = \pi_1(\mathbf{s}(1), \mathbf{f}(1)) \times \dots \times \pi_\ell(\mathbf{s}(\ell), \mathbf{f}(\ell))$$

defines a bivariate (jointly dependent or matched) probability model⁸ in $\mathcal{A}^\ell \times \mathcal{A}^\ell$, for $|\mathcal{A}|$ -sequences in length of ℓ , with generating function

$$(x_{11} + \dots + x_{1c})^2 \times \dots \times (x_{\ell 1} + \dots + x_{\ell c})^2,$$

where $c = |\mathcal{A}|$.

3.2. *Example.* Generating a joint model for two 2-sequences in length of two: The generating function is

$$(x + y)^2 \times (u + v)^2 = x^2u^2 + 2x^2uv + x^2v^2 + 2xyu^2 + 4xyuv + 2xyv^2 + y^2u^2 + 2y^2uv + y^2v^2,$$

which has 16 distinct coefficients, in correspondence with the pairs (\mathbf{s}, \mathbf{f}) of 2-sequences in length of two in $\mathcal{A}^2 \times \mathcal{A}^2$. Here is how the probabilities are generated for this model (we think of x and u as different labels for the base a and y and v as different labels for the base b , with $\mathcal{A} = \{a, b\}$.) :

- $x^2u^2 = P(\mathbf{aa}, \mathbf{aa}) = \pi_1(a, a) \times \pi_2(a, a)$;
- $x^2uv = P(\mathbf{aa}, \mathbf{ab}) = P(\mathbf{ab}, \mathbf{aa}) = \pi_1(a, a) \times \pi_2(a, b)$;
- $x^2v^2 = P(\mathbf{ab}, \mathbf{ab}) = \pi_1(a, a) \times \pi_2(b, b)$;
- $xyu^2 = P(\mathbf{aa}, \mathbf{ba}) = P(\mathbf{ba}, \mathbf{aa}) = \pi_1(a, b) \times \pi_2(a, a)$;
- $xyuv = P(\mathbf{aa}, \mathbf{bb}) = P(\mathbf{bb}, \mathbf{aa}) = P(\mathbf{ab}, \mathbf{ba}) = P(\mathbf{ba}, \mathbf{ab}) = \pi_1(a, b) \times \pi_2(a, b)$;
- $xyv^2 = P(\mathbf{ab}, \mathbf{bb}) = P(\mathbf{bb}, \mathbf{ab}) = \pi_1(a, b) \times \pi_2(b, b)$;
- $y^2u^2 = P(\mathbf{ba}, \mathbf{ba}) = \pi_1(b, b) \times \pi_2(a, a)$;
- $y^2uv = P(\mathbf{ba}, \mathbf{bb}) = P(\mathbf{bb}, \mathbf{ba}) = \pi_1(b, b) \times \pi_2(a, b)$;
- $y^2v^2 = P(\mathbf{bb}, \mathbf{bb}) = \pi_1(b, b) \times \pi_2(b, b)$.

⁸More generally, there is a family of jointly dependent probability models, $P(\mathbf{s}, \mathbf{f} \mid \tau, \sigma) = \pi(\mathbf{s}(\tau 1), \mathbf{f}(\sigma 1)) \times \dots \times \pi(\mathbf{s}(\tau \ell), \mathbf{f}(\sigma \ell))$, one for each pair (τ, σ) of permutations in S_ℓ .

3.3. **Example.** A joint model for two 4-sequences in length of two. The generating function is

$$(a + b + c + d)^2 \times (x + y + z + w)^2,$$

which has 256 distinct coefficients, each one generating the probability of the corresponding pair (s, f) of 4-sequences in length of two in $\mathcal{A}^2 \times \mathcal{A}^2$ (of dimension $16 \times 16 = 256$). Start with a probability model for $\mathcal{A} \times \mathcal{A}$, such as

$$\pi = \begin{bmatrix} 0.04239 & 0.05148 & 0.05753 & 0.05450 \\ 0.05148 & 0.06250 & 0.06985 & 0.06618 \\ 0.05753 & 0.06985 & 0.07806 & 0.07396 \\ 0.05450 & 0.06618 & 0.07396 & 0.07007 \end{bmatrix},$$

obtained here, for illustration, as the product $p \times p$ of two independent and identical ($\pi_1 = \pi_2$) laws

$$p = \frac{1}{68} [14, 17, 19, 18]$$

described in Comment 1.7. From the product $\mathcal{L} = \pi \otimes \pi$, it results the law

$$\mathcal{L} = \frac{1}{100} \begin{bmatrix} 0.1797 & 0.2182 & 0.2439 & 0.2310 & 0.2182 & 0.2649 & 0.2962 & 0.2806 & 0.2439 & 0.2962 & 0.3309 & 0.3134 & 0.2310 & 0.2806 & 0.3134 & 0.2970 \\ 0.2182 & 0.2650 & 0.2962 & 0.2806 & 0.2650 & 0.3218 & 0.3596 & 0.3407 & 0.2962 & 0.3596 & 0.4019 & 0.3806 & 0.2806 & 0.3407 & 0.3806 & 0.3607 \\ 0.2439 & 0.2962 & 0.3311 & 0.3137 & 0.2962 & 0.3596 & 0.4018 & 0.3806 & 0.3311 & 0.4018 & 0.4491 & 0.4254 & 0.3137 & 0.3806 & 0.4254 & 0.4031 \\ 0.2310 & 0.2805 & 0.3135 & 0.2970 & 0.2805 & 0.3406 & 0.3805 & 0.3605 & 0.3135 & 0.3805 & 0.4254 & 0.4031 & 0.2970 & 0.3605 & 0.4031 & 0.3819 \\ 0.2182 & 0.2650 & 0.2962 & 0.2806 & 0.2650 & 0.3218 & 0.3596 & 0.3407 & 0.2962 & 0.3596 & 0.4019 & 0.3806 & 0.2806 & 0.3407 & 0.3806 & 0.3607 \\ 0.2649 & 0.3218 & 0.3596 & 0.3407 & 0.3218 & 0.3906 & 0.4364 & 0.4135 & 0.3596 & 0.4364 & 0.4879 & 0.4622 & 0.3407 & 0.4135 & 0.4622 & 0.4379 \\ 0.2961 & 0.3595 & 0.4018 & 0.3806 & 0.3595 & 0.4366 & 0.4878 & 0.4622 & 0.4018 & 0.4878 & 0.5452 & 0.5167 & 0.3806 & 0.4622 & 0.5167 & 0.4894 \\ 0.2805 & 0.3408 & 0.3808 & 0.3608 & 0.3408 & 0.4136 & 0.4621 & 0.4378 & 0.3808 & 0.4621 & 0.5166 & 0.4894 & 0.3608 & 0.4378 & 0.4894 & 0.4637 \\ 0.2439 & 0.2962 & 0.3311 & 0.3137 & 0.2962 & 0.3596 & 0.4018 & 0.3806 & 0.3311 & 0.4018 & 0.4491 & 0.4254 & 0.3137 & 0.3806 & 0.4254 & 0.4031 \\ 0.2961 & 0.3595 & 0.4018 & 0.3806 & 0.3595 & 0.4366 & 0.4878 & 0.4622 & 0.4018 & 0.4878 & 0.5452 & 0.5167 & 0.3806 & 0.4622 & 0.5167 & 0.4894 \\ 0.3309 & 0.4018 & 0.4490 & 0.4254 & 0.4018 & 0.4879 & 0.5454 & 0.5167 & 0.4490 & 0.5454 & 0.6093 & 0.5773 & 0.4254 & 0.5167 & 0.5773 & 0.5470 \\ 0.3135 & 0.3808 & 0.4255 & 0.4031 & 0.3808 & 0.4622 & 0.5166 & 0.4894 & 0.4255 & 0.5166 & 0.5773 & 0.5469 & 0.4031 & 0.4894 & 0.5469 & 0.5182 \\ 0.2310 & 0.2805 & 0.3135 & 0.2970 & 0.2805 & 0.3406 & 0.3805 & 0.3605 & 0.3135 & 0.3805 & 0.4254 & 0.4031 & 0.2970 & 0.3605 & 0.4031 & 0.3819 \\ 0.2805 & 0.3408 & 0.3808 & 0.3608 & 0.3408 & 0.4136 & 0.4621 & 0.4378 & 0.3808 & 0.4621 & 0.5166 & 0.4894 & 0.3608 & 0.4378 & 0.4894 & 0.4637 \\ 0.3135 & 0.3808 & 0.4255 & 0.4031 & 0.3808 & 0.4622 & 0.5166 & 0.4894 & 0.4255 & 0.5166 & 0.5773 & 0.5469 & 0.4031 & 0.4894 & 0.5469 & 0.5182 \\ 0.2970 & 0.3608 & 0.4032 & 0.3820 & 0.3608 & 0.4379 & 0.4895 & 0.4638 & 0.4032 & 0.4895 & 0.5470 & 0.5183 & 0.3820 & 0.4638 & 0.5183 & 0.4910 \end{bmatrix}$$

Each entry of \mathcal{L} has the form $\pi(i, j) \times \pi(i', j')$, for i, j, i', j' letters in the alphabet \mathcal{A} , thus corresponding to the joint probability of the sequences $s = ii'$ and $f = jj'$, that is

$$P(ii', jj') = \pi(i, j) \times \pi(i', j').$$

3.4. *Multivariate* models are defined similarly. Given probability models, π_1, \dots, π_ℓ , in \mathcal{A}^p , then

$$P(s_1, s_2, \dots, s_p) = \pi_1(s_1(1), s_2(1), \dots, s_p(1)) \times \dots \times \pi_\ell(s_1(\ell), s_2(\ell), \dots, s_p(\ell))$$

defines a joint probability model for p $|\mathcal{A}|$ -sequences in length of ℓ , with generating function

$$(x_{11} + \dots + x_{1c})^p \times \dots \times (x_{\ell 1} + \dots + x_{\ell c})^p.$$

4. PAIRWISE SEQUENCE ALIGNMENT

A sequence f becomes distinct from an ancestor sequence s by *mutations* such as *substitutions* and *gaps* (insertions and deletions), which may be modulated by natural selection. Mutations are typically site-independent in DNA and protein sequences and site-dependent in RNA sequences.

4.1. A *scoring* for two sequences s and f in \mathcal{A}^ℓ is a log-likelihood ratio

$$S(s, f) = \log \frac{P(s, f)}{Q(s, f)},$$

comparing two joint probability models, P and Q , descriptives of the two sequences. Typically, Q is the independent letters model (Comment 1.9) and P a jointly dependent model (Comment 3.1), so that

$$S(s, f) = \sum_{j=1}^{\ell} \log \frac{\pi(s(j), f(j))}{\pi(s(j)) \times \pi(f(j))},$$

which depends only on a *scoring* or *substitution* matrix, s , with entries

$$s(a, b) = \log \frac{\pi(a, b)}{\pi(a) \times \pi(b)},$$

which may be obtained, in the evolutionary context, as

$$s(a, b) = \log \frac{\pi(b | a)}{\pi(b)},$$

with the transition probabilities $\pi(b | a)$ obtained from recent branches of phylogenetic trees and taken to near-stationary states;

4.2. *Example.* Given the probability law in $\mathcal{A} \times \mathcal{A}$ obtained by symmetrizing the law for four-sequences in length of two, described in Comment 1.12)

$$p_2 = \begin{bmatrix} & \text{A} & \text{C} & \text{G} & \text{T} \\ \text{A} & 0.119 & 0.0702 & 0.0939 & 0.0735 \\ \text{C} & 0.0702 & 0.0413 & 0.0277 & 0.0403 \\ \text{G} & 0.0939 & 0.0277 & 0.0684 & 0.0519 \\ \text{T} & 0.0734 & 0.0403 & 0.0519 & 0.0568 \end{bmatrix},$$

and the law in \mathcal{A} , again from Comment 1.12,

$$p_1 = [0.355 \quad 0.179 \quad 0.241 \quad 0.221],$$

the scoring matrix

$$\begin{bmatrix} & \text{A} & \text{C} & \text{G} & \text{T} \\ \text{A} & -0.093 & 0.138 & 0.124 & -0.105 \\ \text{C} & 0.138 & 0.367 & -0.644 & 0.014 \\ \text{G} & 0.124 & -0.644 & 0.226 & -0.046 \\ \text{T} & -0.106 & 0.014 & -0.046 & 0.202 \end{bmatrix},$$

would be obtained, with entry (a, b) given by $\log_2[p_2(a, b)/p_1(a)p_1(b)]$.

4.3. *Example* Consider sequences 3 (f) and 4 (s), from Comment 5.1.

$f = \text{ttttcgtatggaacctgggatcttttagtttgaatgggagagccattccgcttgaaaaaattagataaggtaag},$

$s = \text{ttccgtcatggaacctggaatagttgctcaaaaagtgaggagcaaccgcttaggtttgaaaaaattagataagggcgg}.$

Here are their pairwise concordance frequencies:

$$\begin{bmatrix} & f: \text{A} & \text{C} & \text{G} & \text{T} \\ s: \text{A} & 17 & 2 & 4 & 0 \\ & \text{C} & 3 & 5 & 1 & 3 \\ & \text{G} & 3 & 1 & 15 & 1 \\ & \text{T} & 1 & 4 & 2 & 15 \end{bmatrix}.$$

By symmetrization, we obtain the joint model, in $\mathcal{A} \times \mathcal{A}$,

$$p_2 = \begin{bmatrix} 0.220 & 0.0325 & 0.0455 & 0.00649 \\ 0.0325 & 0.0648 & 0.0130 & 0.0455 \\ 0.0455 & 0.0130 & 0.195 & 0.0195 \\ 0.00649 & 0.0455 & 0.0195 & 0.195 \end{bmatrix}.$$

Here we take as p_1 in \mathcal{A} the marginal model, namely

$$p_1 = [0.304 \quad 0.156 \quad 0.274 \quad 0.266].$$

From p_1 and p_2 we obtain the scoring matrix

$$\begin{bmatrix} & \text{A} & \text{C} & \text{G} & \text{T} \\ \text{A} & 1.25 & -0.544 & -0.873 & -3.64 \\ \text{C} & -0.544 & 1.42 & -1.72 & 0.138 \\ \text{G} & -0.873 & -1.72 & 1.38 & -1.91 \\ \text{T} & -3.64 & 0.138 & -1.91 & 1.46 \end{bmatrix},$$

and scores

$$(4.1) \quad \begin{bmatrix} & 1 & 2 & 3 & 4 & 5 \\ 1 & 105.0 & -38.7 & -14.5 & -14.9 & -38.7 \\ 2 & -38.7 & 103.0 & -3.2 & 14.8 & 103.0 \\ 3 & -14.5 & -3.2 & 105.0 & 50.3 & -3.2 \\ 4 & -14.9 & 14.8 & 50.3 & 104.0 & 14.8 \\ 5 & -38.7 & 103.0 & -3.2 & 14.8 & 103.0 \end{bmatrix}$$

among the five sequences.

4.4. *Example* The following is the BLASTn alignment output for the sequences 3 and 4 discussed in Comment 4.2:

Score = 47.5 bits (29), Expect = 0.001

Identities = 51/73 (69\%) Strand = Plus / Plus

```
Query: 1  ttttcgctatggaacctgggatctttagtttgaaatgggagagccattccgcctggaaa 60
          ||| || ||||| || || | || ||||| || | | ||||
Sbjct: 1  tttccgtcatggaacctggaatagttgctcaaaagtgggagcaaccgcttaggtttgaaa 60
```

```
Query: 61 aaattagataagg 73
```

```
|||||
```

```
Sbjct: 61 aaattagataagg 73
```

CPU time: 0.04 user secs. 0.05 sys. secs 0.09 total secs.

Lambda K H
1.10 0.333 0.549

Gapped Lambda K H
1.10 0.333 0.549

Matrix: blastn matrix:1 -1

4.5. *Bayesian Assessment* of alignment scores. Based on the probability models introduced in Comments 1.9 and 3.1, we may define two models describing the joint alignment of two sequences s and f , namely, the independent probabilities model,

$$P(s, f | R) = \pi(s(1))\pi(f(1)) \times \pi(s(2))\pi(f(2)) \times \dots \times \pi(s(\ell))\pi(f(\ell)),$$

and the matched probability model,

$$P(s, f | M) = \pi(s(1), f(1)) \times \dots \times \pi(s(\ell), f(\ell)).$$

Now apply Bayes Formula to obtain the posterior odds on the matching model, that is,

$$\frac{P(M | s, f)}{P(R | s, f)} = \frac{P(s, f | M)}{P(s, f | R)} \times \frac{P(M)}{P(R)}.$$

Taking the log, from Comment 4.1, we obtain the posterior probability for the matched model,

$$P(M | s, f) = \frac{\mathcal{O}e^{S(s,f)}}{1 + \mathcal{O}e^{S(s,f)}},$$

where $S(s, f)$ is the scoring for the two sequences and \mathcal{O} is the prior odds $\frac{P(M)}{P(R)}$ on the matching model.

4.6. *Example* The posterior probabilities for matching sequences, based on the matrix of scores shown in (4.1), relative to uniform prior probability, are (see Comment 4.5):

$$\begin{bmatrix} & 1 & 2 & 3 & 4 & 5 \\ 1 & 1.0 & 1.72 \times 10^{-17} & 0.000000504 & 0.000000374 & 1.72 \times 10^{-17} \\ 2 & 1.72 \times 10^{-17} & 1.0 & 0.0396 & 1.0 & 1.0 \\ 3 & 0.000000504 & 0.0396 & 1.0 & 1.0 & 0.0396 \\ 4 & 0.000000374 & 1.0 & 1.0 & 1.0 & 1.0 \\ 5 & 1.72 \times 10^{-17} & 1.0 & 0.0396 & 1.0 & 1.0 \end{bmatrix}.$$

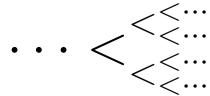
5. TREES

The *nodes* $s(1), s(2), \dots, s(\ell)$ (points in a set C) defining the *branches*

$$r \rightarrow s(1) \rightarrow s(2) \rightarrow \dots \rightarrow s(\ell)$$

of a tree of *height* ℓ and *root* $r \in C$ may be represented by the space C^L of all maps from $L = \{1, \dots, \ell\}$ to

$C = \{1, \dots, c\}$. Here is a diagram illustrating a 3-generation segment of a binary tree ($\ell = 3, c = 2$):



For each root, the tree has c^ℓ branches (s), in each of which an event $x(s)$ may be annotated, such as its *leaf* $s(\ell)$. We may imbed a tree into any context as follows: suppose that the context of interest is the space V of all 4-sequences in length of 3, and consider a tree $C^{\{1,2\}}$ of height 2. Let C be the set S_3 of permutations in the set $\{1, 2, 3\}$ indexing the sequences. To each $f \in C^{\{1,2\}}$ and root τ in C , we associate the branch

$$\text{root } \tau \rightarrow f(1) = \sigma_1 \rightarrow f(2) = \sigma_2$$

of a tree $C^{\{1,2\}}$, which now can be used to generate a tree $V^{\{1,2\}}$ of height 2 in V by properly reading the original tree within the context of interest. For each root 4-sequence (s) in length of 3, we obtain the branch

$$\text{root } s\tau \xrightarrow{\sigma_1} s\tau\sigma_1 \xrightarrow{\sigma_2} s\tau\sigma_1\sigma_2.$$

Here is one branch of a tree with root AGC:

$$\text{root } \text{AGC} \xrightarrow{(123)} \text{CAG} \xrightarrow{(12)} \text{ACG} \text{ leaf.}$$

Similarly, we may imbed the original tree *from the right* by choosing C the set S_4 of permutation in the 4-letter alphabet. The branches of a tree $V^{\{1,2\}}$ of height 2 in V now have the form

$$\text{root } \pi s \xrightarrow{\pi_1} \pi_1 \pi s \xrightarrow{\pi_2} \pi_2 \pi_1 \pi s \text{ leaf.}$$

For example, let

$$f(1) = \begin{bmatrix} A & \rightarrow & A \\ C & \rightarrow & G \\ G & \rightarrow & C \\ T & \rightarrow & T \end{bmatrix}, \quad f(2) = \begin{bmatrix} A & \rightarrow & C \\ C & \rightarrow & A \\ G & \rightarrow & T \\ T & \rightarrow & G \end{bmatrix},$$

so that the resulting branch with root AGC is

$$\text{root } AGC \xrightarrow{f(1)} ACG \xrightarrow{f(2)} CAT \text{ leaf.}$$

Each node in the tree may be interpreted as an aligned nucleotide sequence with common ancestor πs and substitution process dictated by probability laws in S_4 . We observe that these trees have the property that roots and leaves are always members of the same orbit. Also note that a given tree may be embedded from the right and from the left during its evolution, giving rise to more complex orbits.

5.1. A *cladogram* (or dendrogram) is a diagram showing the evolutionary relationships among leaves sharing a common root. The *distances* among leaves indicate their degree of phylogenetic similarity.

Example 5.1. Here are 5 fragments of the HIV virus from different isolates⁹.

Sample 1: African Green Monkey (HIVagm)

ttttcgctatgggtgctggagatagtgctatagcgggcaatatcggcgacaaaaagacaggtttatattaacgaaggaatgg

Sample 2: Human (HIVbru)

ttatcggcatgggttaaaggaataaagagtgctaagtttagaagcaatcttaggtttgaaaaattagataacgatgcccc

Sample 3: Human (HIVsc)

ttttcgctatggaaacctgggatcttttagtttgaaatgggagagccattccgcctggaaaaattagataaaggtaaggccgc

Sample 4: Human (HIVcdc)

ttccgctcatggaaacctggaatagttgctcaaaagtgggagcaaccgcttaggtttgaaaaattagataaaggcgcg

Sample 5: Chimpanzee (HIVcpz)

ttatcggcatgggttaaaggaataaagagtgctaagtttagaagcaatcttaggtttgaaaaattagataacgatgcccc

Here we take as distance between sequences their observed concordance frequencies:

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 1.0 & 0.403 & 0.468 & 0.481 & 0.403 \\ 2 & 0.403 & 1.0 & 0.533 & 0.598 & 1.0 \\ 3 & 0.468 & 0.533 & 1.0 & 0.676 & 0.533 \\ 4 & 0.481 & 0.598 & 0.676 & 1.0 & 0.598 \\ 5 & 0.403 & 1.0 & 0.533 & 0.598 & 1.0 \end{bmatrix},$$

from which the cladogram, shown in Figure 5.1, is obtained.

REFERENCES

- Doi, H. (1991), 'Importance of purine and pyrimidine content of local nucleotide sequences (six bases long) for evolution of human immunodeficiency virus type 1', *Evolution* **88**(3), 9282–9286.
- Evans, S. N. and Speed, T. P. (1993), 'Invariants of some probability models used in phylogenetic inference', *Ann. Statist.* **21**(1), 355–377.

⁹<http://smccd.net/accounts/case/CPS/400.html>

FIGURE 5.1. Cladogram for five sample sequences.

