

SYMMETRY RELATIONS ANNOTATED EXAMPLES AND BASIC CONCEPTS

MARLOS VIANA

CONTENTS

1. Introduction	1
2. The Orbit Method	2
2.1. Algebraic Aspects.	2
2.2. Probabilistic Aspects.	2
2.3. Observational Aspects.	4
3. Nucleotide Sequences.	4
3.1. Composing sequences and symmetries.	4
3.2. Cyclic symmetries.	4
3.3. Frequency diversity.	5
3.4. Baseline variation.	5
4. Maxwell-Boltzmann Equilibrium Distribution	9
4.1. The canonical distribution for the Maxwell-Boltzmann model.	11
5. Maxwell-Boltzmann Law for Velocities in a Perfect Gas	13
5.1. Classical derivation of Maxwell-Boltzmann Law.	13
6. Summary	15
7. Related Reading	15
Appendix A. The Human Immunodeficiency Virus Type I	16
Appendix B. Entropy and Information Content	19
Appendix C. Permutations	19
Appendix D. Symmetries in Two-sequences in Length of Four	20
Appendix E. Counting Orbits	21
Appendix F. Orbits and Their Volumes	22
Appendix G. More on Calculus with Orbits of Symmetry- Cross Sections	23
References	25

1. INTRODUCTION

These lecture notes introduce the basic elements of symmetry relations applied to the study of certain structures of interest. These elements are introduced in Section 2 within the context of simple nucleotide sequences, where we distinguish three components of the (orbit) method: one algebraic, one probabilistic and another observational. These components are introduced with the objective of highlighting the symmetry arguments that are common to the examples introduced in these notes, namely, an example in evolutionary molecular biology (Section 3), the classical Maxwell-Boltzmann equilibrium distribution (Section 4) and

Date: January 22, 2003.

Lecture notes prepared for HON 201 *The Physics of Chance*, Fall semester, 2002. Comments, corrections, suggestions to viana@uic.edu. Selected biographic citations were abstracted from <http://www-gap.dcs.st-and.ac.uk/history/BiogIndex.html> or from <http://www.nobel.se/physics/>.

Maxwell-Boltzmann velocity distribution in a perfect gas (Section 5). The geometrical or inferential details of the method are not included in these introductory notes.

2. THE ORBIT METHOD

We have looked at two-sequences in length of four as an example of a set of objects which one wants to simplify, in connection with the appropriate, say molecular or evolutionary, context, e.g., Doi (1991). By simplifying we mean factoring the whole set into smaller, interpretable pieces.

2.1. Algebraic Aspects. Recall that any biological sequence ℓ base pairs long is representable by a function or map

$$s : L \rightarrow \mathcal{A},$$

where $L = \{1, 2, \dots, \ell\}$ is the set for the ordered positions in which the residues in the alphabet \mathcal{A} are located. Typical alphabets are $\mathcal{A} = \{A, G, T, C\}$ in DNA sequences, $\mathcal{A} = \{A, G, T, U\}$ in RNA sequences, or simply a two-letter alphabet $\mathcal{A} = \{u, y\}$ of purine ($u=A$ or $u=G$) and pyrimidine ($y=C$ or $y=T$) residues. Appendix A shows a segment ($\ell = 2586$) bp-long of a global sequence in length of $\ell = 9229$ bps. In Doi (1991) the local sequences of interest were in length of 2, 3, 4, 5 and 6. We indicate by $|\mathcal{A}|$ the number of letters in the alphabet \mathcal{A} . Indicate by V the space of all $|\mathcal{A}|$ -sequences in length of ℓ . There are $|\mathcal{A}|^\ell$ sequences in V . For example, every two-sequence in length of four, with $\mathcal{A} = \{u, y\}$, is a map

$$s : \{1, 2, 3, 4\} \rightarrow \{u, y\}.$$

These $2^4 = 16$ sequences are listed in Appendix D.

2.1.1. The orbits of similarity. We define the symmetries in V as a set of permutations acting on (or applied to), for example, the ordered positions set L . The results of these symmetry operations are summarized in matrix (D.1) in Appendix D, where the symmetries are those of the permutation group of order 4- e.g., Appendix C.

The effect of applying the permutations on the set L is that of removing the order of the positions- equivalently, any two sequences are then equivalent, similar or indistinguishable, when they differ only by reordering the location of the letters or residues. As a result, we obtain another space, called the *quotient space*, in which the elements are the resulting 5 permutation orbits $\mathcal{O}_0, \mathcal{O}_1, \dots, \mathcal{O}_4$. These orbits are characterized by the number of, say, purines. That is, orbit \mathcal{O}_i is composed of those sequences with exactly i purines in it. More precisely, to the orbit \mathcal{O}_i , we associate a distribution of the purine-pyrimidine *levels* that is characteristic of the orbit. That is,

$$(2.1) \quad \mathcal{O}_i \rightarrow (\text{number of purines, number of pyrimidines}) = (i, \ell - i).$$

The total number of sequences in V with the same purine-pyrimidine levels is given by the *volume*

$$(2.2) \quad |\mathcal{O}_i| = \binom{\ell}{i} = \frac{\ell!}{i!(\ell - i)!}$$

of sequences in the corresponding orbit.

2.2. Probabilistic Aspects. The sequences are seen as random variables. Let P indicate a probability model in the space V of sequences. We say that P has the symmetry of a permutation group G if P is constant over each one of the orbits, that is,

$$(2.3) \quad P(s) = P(s\tau)$$

for all sequences s in V and permutations τ in G . Note that the assessment of condition (2.3) may depend on the particular region of a given genome of interest. Because s is now a random variable, the purine-pyrimidine levels

$$(\text{number of purines, number of pyrimidines}) = (i, \ell - i)$$

are also random variables, and consequently, the probability laws

$$(2.4) \quad \mathcal{L}_i = \left(\frac{i}{\ell}, \frac{\ell - i}{\ell} \right), \quad i = 0, 1, \dots, \ell,$$

associated with the orbit of s are also random. Here are the possible probability laws for purine-pyrimidine levels from two-sequences in length of four:

$$\mathcal{L}_0 = (0, 1), \mathcal{L}_1 = \left(\frac{1}{4}, \frac{3}{4}\right), \mathcal{L}_2 = \left(\frac{2}{4}, \frac{2}{4}\right), \mathcal{L}_3 = \left(\frac{3}{4}, \frac{1}{4}\right), \mathcal{L}_4 = (1, 0).$$

The likelihood of each law is therefore determined by the probability of seeing a sequence which is associated with the law- because all sequences in the orbit \mathcal{O}_i lead to the law \mathcal{L}_i and conversely, we see that \mathcal{L}_i occurs with probability $P(\mathcal{O}_i)$; shortly,

$$\text{Probability of law } \mathcal{L}_i = P(\mathcal{O}_i).$$

Clearly, if the law P is such that all sequences are equally likely (P is said to be uniform), then condition (2.3) is satisfied and

$$(2.5) \quad \text{Probability of law } \mathcal{L}_i = P(\mathcal{O}_i) = \frac{|\mathcal{O}_i|}{|V|} = \frac{\binom{\ell}{i}}{|V|}.$$

We have, for two-sequences in length of four,

$$P(\mathcal{O}_0) = \frac{1}{16}, \quad P(\mathcal{O}_1) = \frac{4}{16}, \quad P(\mathcal{O}_2) = \frac{6}{16}, \quad P(\mathcal{O}_3) = \frac{4}{16}, \quad P(\mathcal{O}_4) = \frac{1}{16},$$

so that the most likely distribution of purine-pyrimidine levels, under uniformly distributed sequences, is

$$\mathcal{L}_2 = \left(\frac{2}{4}, \frac{2}{4}\right).$$

2.2.1. *Four-sequences in length of three.* Let $\mathcal{A} = \{A, C, G, T\}$. The space V of all four-sequences in length of three has $|V| = 4^3 = 64$ sequences. The random variables generated by similarity of the positions are

$$(\text{number of adenines, number of cytosines, number of guanines, number of thymines}) = (f_a, f_c, f_g, f_t),$$

with $f_a + f_c + f_g + f_t = 3$. Consequently, the corresponding probability laws

$$\mathcal{L}_\lambda = \left(\frac{f_a}{3}, \frac{f_c}{3}, \frac{f_g}{3}, \frac{f_t}{3}\right)$$

are also random, and we may ask which one is more likely to be present on any given context. The index λ in \mathcal{L}_λ indicates the corresponding orbit, in analogy with expression (2.4). We obtain these indices as the possible *integer partitions* of 3 in length of 4, so that there are three *types* of orbits, and corresponding laws:

$$\begin{aligned} \lambda = 3000 &\rightarrow \mathcal{L}_{3000} = (1, 0, 0, 0), \\ \lambda = 2100 &\rightarrow \mathcal{L}_{2100} = \left(\frac{2}{3}, \frac{1}{3}, 0, 0\right), \\ \lambda = 1110 &\rightarrow \mathcal{L}_{1110} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0\right). \end{aligned}$$

Similarly to expression (2.5) we now obtain

$$(2.6) \quad \text{Probability of law } \mathcal{L}_{3000} = P(\mathcal{O}_{3000}) = \frac{\binom{3}{(3,0,0,0)}}{|V|} = \frac{3!}{3!0!0!0!} \frac{1}{64} = \frac{1}{64},$$

$$(2.7) \quad \text{Probability of law } \mathcal{L}_{2100} = P(\mathcal{O}_{2100}) = \frac{\binom{3}{(2,1,0,0)}}{|V|} = \frac{3!}{2!1!0!0!} \frac{1}{64} = \frac{3}{64},$$

$$(2.8) \quad \text{Probability of law } \mathcal{L}_{1110} = P(\mathcal{O}_{1110}) = \frac{\binom{3}{(1,1,1,0)}}{|V|} = \frac{3!}{1!1!1!0!} \frac{1}{64} = \frac{6}{64},$$

so that, under the assumption that all 4-sequences in length of 3 are equally likely (uniform probability), the most probable distribution by levels of nucleotides comes from the *class* of distribution given by \mathcal{L}_{1110} , each of which has the highest probability, $6/64$. Simple combinatorics show that there are

$$(2.9) \quad \frac{4!}{3!1!} = 4$$

such most probable laws of nucleotide levels, namely,

$$(2.10) \quad \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0\right), \left(\frac{1}{3}, \frac{1}{3}, 0, \frac{1}{3}\right), \left(\frac{1}{3}, 0, \frac{1}{3}, \frac{1}{3}\right), \left(0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right).$$

2.3. Observational Aspects. Now we measure something on each sequence in V , like, for example, its molecular weight. Consequently, starting with the probability laws

$$\mathcal{L}_\lambda = \left(\frac{f_a}{3}, \frac{f_c}{3}, \frac{f_g}{3}, \frac{f_t}{3}\right)$$

we may then derive the mean molecular weight of the sequence. If m_a, m_c, m_g, m_t indicate the molecular weight of the corresponding nucleotides¹, then, the mean molecular weight associated with \mathcal{L}_λ is

$$\bar{m}_\lambda = m_a \frac{f_a}{3} + m_c \frac{f_c}{3} + m_g \frac{f_g}{3} + m_t \frac{f_t}{3}.$$

Example 2.1. Mean molecular weights. Given the molecular weights of the four basic nucleotides.

- Adenine (Amino-6-purine $C_5H_5N_5$) Molecular Weight: 135.128 g/mol;
- Guanine (Amino-2-hydroxy-6-purine $C_5H_4N_5O$) Molecular Weight: 150.12 g/mol;
- Cytosine (2-hydroxy, 4 Amino-pyrimidine $C_4H_5N_3O$) Molecular Weight: 111.1 g/mol;
- Thymine (2,4-dihydroxy-5-methyl pyrimidine $C_5H_6N_2O_2$) Molecular Weight: 126.1 g/mol,

we obtain

$$\bar{m} = 132.116\text{g/mol}, \quad \bar{m} = 137.116\text{g/mol}, \quad \bar{m} = 127.109\text{g/mol}, \quad \bar{m} = 129.106\text{g/mol},$$

corresponding to the laws in (2.10).

3. NUCLEOTIDE SEQUENCES.

We have observed that characterizing a biological 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 the action of a particular group of transformations.

3.1. 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

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

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

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

where C_ℓ is the cyclic group of order ℓ (e.g., Section C in the Appendix). For example,

$$[CGG] = \{CGG, GCG, GGC\}, \quad [uuyuuy] = \{uuyuuy, yuuyuu, uyuyuy\}.$$

¹A molecular weight calculator is available, for example, at <http://www.public.iastate.edu/miller/tables/mwt.htm>.

3.3. Frequency diversity. The work of Doi (1991) on the evolutionary strategy of the HIV-1 virus is based on the study of the *frequency diversity* in each cyclic orbit $[s]$, 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 within a given region of interest.

Example 3.1. Figure 3.1 shows the observed cyclic diversity for the cyclic orbits $[acg]$, $[aac]$, $[atg]$ and $[cgt]$ along the BRU isolate K02013, evaluated at each of 45 intervals in length of 200 residues (the global sequence is approximately 9000 bp-long). Estimated based on two isolates (BRU and OYI) of the same virus are shown. The BRU isolate is 9229 bp (base-pair) long and the OYI is 9102 bp long (accession number M26727)..

3.4. Baseline variation. In Section 2.2 we observed that if P is a probability law in V such that $P(s) = P(s\tau)$ for all permutation $\tau \in S_\ell$, then P is constant in each one of the *orbits* of V by S_ℓ . 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 ℓ . The uniformity within each orbit may serve as a baseline or reference variation e.g., Durbin, Eddy, Krogh and Mitchison (1998). Figures 3.2, 3.3, 3.4 and 3.5 illustrate the relative frequency ratios and diversity within each one of the two orbits

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

The ratios and diversities imply the eventual inadequacy of the independent letters model (which, in particular, satisfies the invariance condition (2.3)).

FIGURE 3.1. Diversity, expressed as the range $\max_{f \in [s]} \hat{f} - \min_{f \in [s]} \hat{f}$, of selected orbits (indicated in the y-axis: [acg], [aac], [atg], [cgt]) along the BRU isolate K02013, evaluated at each of 45 intervals in length of 200 residues.

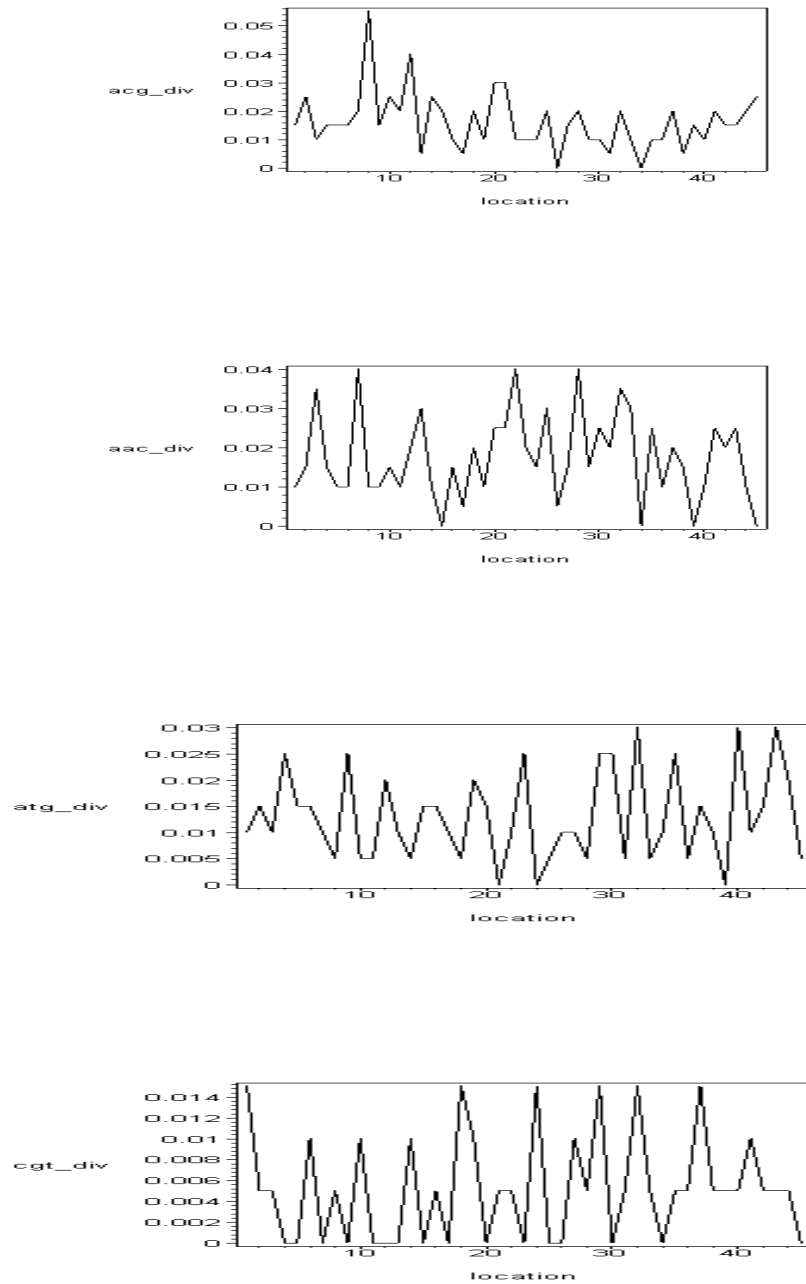


FIGURE 3.2. Relative frequency ratios f_{yyu}/f_{uyy} (red, steady curve) and f_{yyu}/f_{uyy} (green, variable curve) in the single-purine orbit u_1 (top) and corresponding orbit diversity (bottom), along the BRU isolate K02013, evaluated at each of 45 intervals in length of 200 residues.

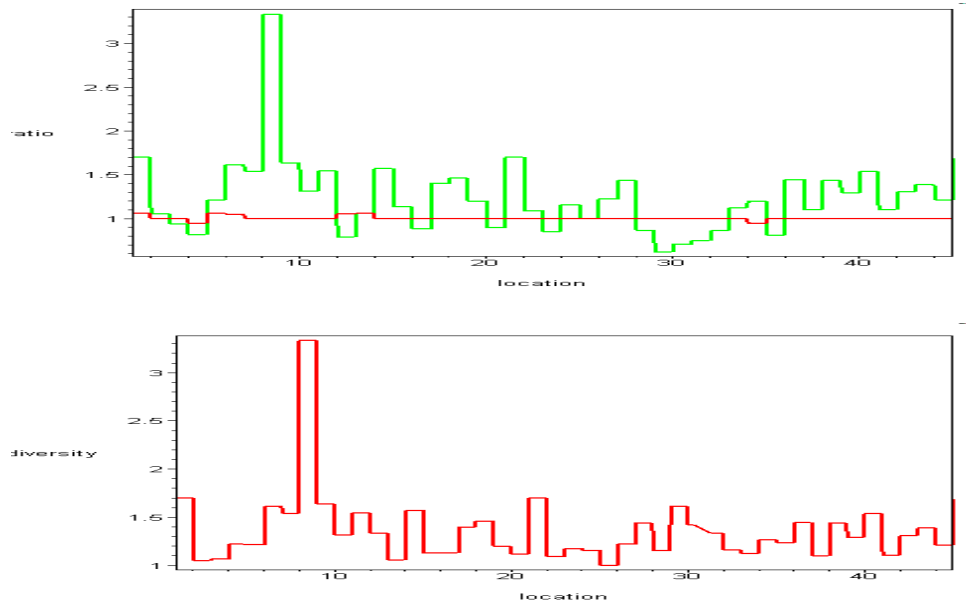


FIGURE 3.3. Relative frequency ratios f_{uuy}/f_{yuu} (red, steady curve) and f_{uuy}/f_{yuu} (green, variable curve) in the single-pyrimidine orbit u_2 (top) and corresponding orbit diversity (bottom), along the BRU isolate K02013, evaluated at each of 45 intervals in length of 200 residues.

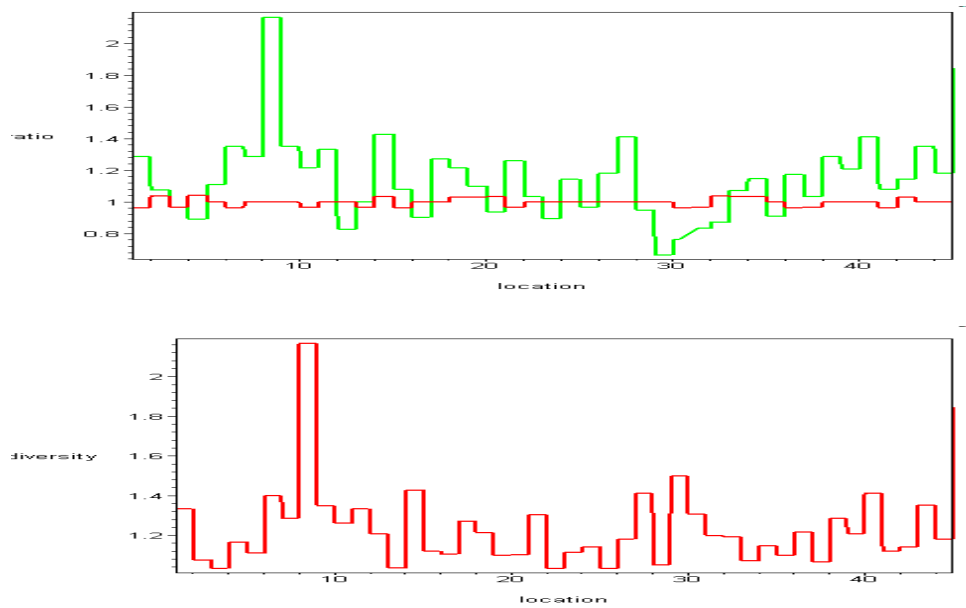


FIGURE 3.4. Relative frequency ratios f_{yyu}/f_{uyy} (green, steady curve) and f_{yyu}/f_{yuy} (red, variable curve) in the single-purine orbit u_1 (top) and corresponding orbit diversity (bottom), along the isolate M26727, evaluated at each of 45 intervals in length of 200 residues.

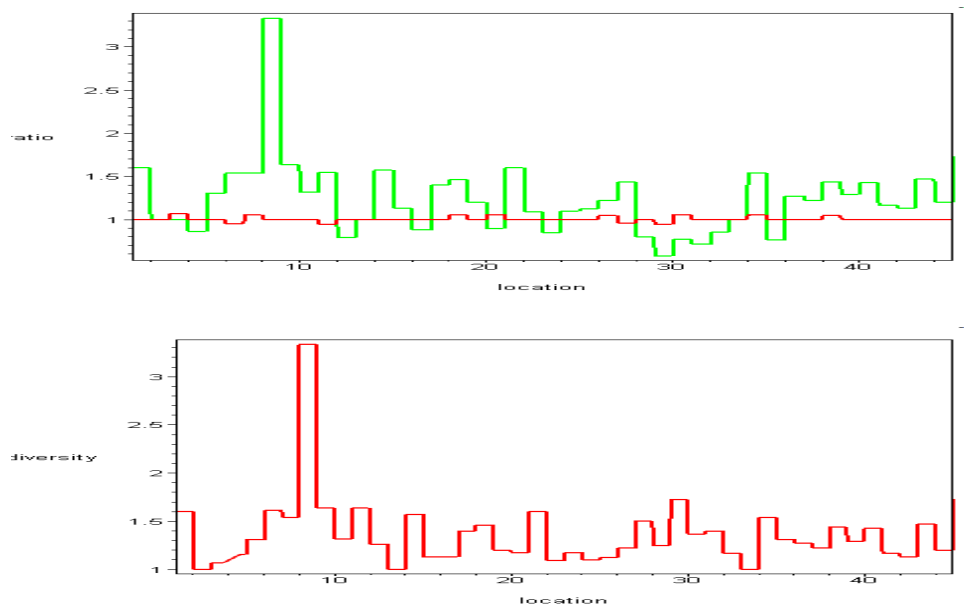
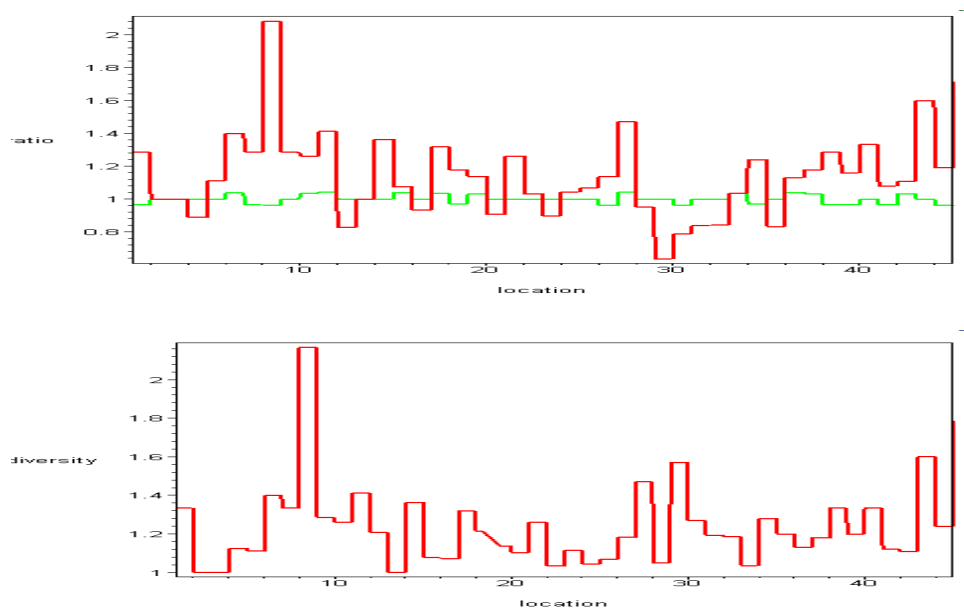


FIGURE 3.5. Relative frequency ratios f_{uuy}/f_{yuu} (green, steady curve) and f_{uuy}/f_{uyu} (red, variable curve) in the single-pyrimidine orbit u_2 (top) and corresponding orbit diversity (bottom), along the isolate M26727, evaluated at each of 45 intervals in length of 200 residues.



4. MAXWELL-BOLTZMANN EQUILIBRIUM DISTRIBUTION

The following is a quote from von Mises (1957, p.200) which describes the context of Boltzmann's arguments. Only the notation was partially adapted to conform with the present one. The *velocity space* is as usually described in the physics literature, e.g., Ruhla (1989, p.79).

The assumption of the classical theory is that equal probabilities are assigned to equal volumes in this velocity space. We will call each element of the volume in the velocity space a possible 'position' or 'place' of the molecule. If we now consider a collective whole elements are distributions of certain number ℓ of molecules over c positions in the velocity space, it follows that all possible c^ℓ distributions have the same probability. For example, imagine two molecules A and B , and three different positions a, b, c . The number of different distributions is 9, since each of the three positions of A , namely Aa, Ab, Ac can be combined with each of B . According to the classical theory, all these distributions have the same probability, $1/9$. A new theory, first suggested by the Indian physicist Bose², and developed by Einstein, chooses another assumption regarding the equal probabilities. Instead of considering single molecules and assuming that each molecule can occupy all positions in the velocity space with equal probability, the new theory starts with the concept of 'repartition'. This is given by the *number* of molecules at each place of the velocity space, without paying attention to the individual molecules. From this point of view, only six 'partitions' are possible for two molecules on three places, namely, both molecules may be together at a , at b , or at c , or they may be separated, one at a and one at b , one at a and one at c , or one at b and one at c . According to the Bose-Einstein theory, each of these six cases has the same probability, $1/6$. In the classical theory, each of these three possibilities would have the probability of $1/9$, each of the other three, however, $2/9$, because, in assuming individual molecules, each of the last three possibilities can be realized in two different ways: A can be in a , and B in b , or vice versa, B can be in a , and A in b .

The Italian physicist Fermi³ advanced still another hypothesis. He postulated that only such distributions are possible- and possess equal probabilities- in which all molecules occupy different places. In our example of two molecules and three positions, there would only be three possibilities, each having the probability $1/3$; i.e., one molecule in a and one in b ; one in a and one in c ; one in b and one in c .

In testing these and other hypothesis it is assumed, according to Boltzmann's entropy theorem, that the probability of the state of a gas is a measure of its entropy, and the object of the investigation is to find which theory best approximates the actually observed dependence of entropy on temperature and mass.

The arguments in von Mises' narrative can be summarized, using the orbit method, as follows: Let $L = \{A, B\}$, $C = \{a, b, c\}$ and V the set of all maps $s : L \rightarrow C$, that is,

$$V = \left[\begin{array}{c|ccccccccc} s & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ \hline s(A) & a & b & c & a & b & a & c & b & c \\ s(B) & a & b & c & b & a & c & a & c & b \end{array} \right].$$

Under the Maxwell-Boltzmann (MB) model, it is assumed that all points or configurations in the space V are equally likely, or *uniformly distributed*, that is:

$$P(s) = \frac{1}{|V|} = \frac{1}{9}, \text{ for all } s \in V.$$

²Satyendranath Bose, Born: 1 Jan 1894 in Calcutta, India Died: 4 Feb 1974 in Calcutta, India

³Enrico Fermi was born in Rome on 29th September, 1901. The Nobel Prize for Physics was awarded to Fermi for his work on the artificial radioactivity produced by neutrons, and for nuclear reactions brought about by slow neutrons. He died in Chicago on 29th November, 1954.

Under the Fermi-Dirac (FD) model, it is assumed that all points in the quotient space V/S_2 of V by the action of shuffling the molecules' labels (in $L = \{A, B\}$) are uniformly distributed. Thus, in the FD model the uniform probability is defined on the *orbits* of V under the symmetries defined by the permutations in $S_2 = \{1, (12)\}$ acting on V . In analogy with the calculations shown in Matrix (D.1), we have

$$\left[\begin{array}{c|cccccccc} \sigma \backslash s & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ \hline 1 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ (12) & 1 & 2 & 3 & 5 & 4 & 7 & 6 & 9 & 8 \end{array} \right],$$

so that the six orbits in the quotient space V/S_2 are

$$\mathcal{O}_{11} = \{1\}, \quad \mathcal{O}_{12} = \{2\}, \quad \mathcal{O}_{13} = \{3\}, \quad \mathcal{O}_{21} = \{4, 5\}, \quad \mathcal{O}_{22} = \{6, 7\}, \quad \mathcal{O}_{23} = \{8, 9\},$$

each one of these having probability of $1/6$. A probability law in V such as

$$P(s) = \begin{cases} 1/6 & \text{when } s \in \{\mathcal{O}_{11}, \mathcal{O}_{12}, \mathcal{O}_{13}\}, \\ 1/12 & \text{when } s \in \{\mathcal{O}_{21}, \mathcal{O}_{22}, \mathcal{O}_{23}\}, \end{cases}$$

would be consistent with the assumptions of the FD model.

The Bose-Einstein (BE) model assumes that only the injective maps

$$V_I \equiv \left[\begin{array}{c|cccccc} s & 4 & 5 & 6 & 7 & 8 & 9 \\ \hline s(A) & a & b & a & c & b & c \\ s(B) & b & a & c & a & c & b \end{array} \right] \subset V$$

are admissible representations of the physical system, and that a uniform probability law is assigned to the resulting orbits in the quotient space of V_I by the action of shuffling the molecules' labels. Therefore, starting with

$$\left[\begin{array}{c|cccccc} \sigma \backslash s & 4 & 5 & 6 & 7 & 8 & 9 \\ \hline 1 & 4 & 5 & 6 & 7 & 8 & 9 \\ (12) & 5 & 4 & 7 & 6 & 9 & 8 \end{array} \right],$$

we obtain the three orbits

$$\mathcal{O}_1 = \{4, 5\}, \quad \mathcal{O}_2 = \{6, 7\}, \quad \mathcal{O}_3 = \{8, 9\}$$

in the quotient space V_I/S_2 . To each of these, a probability of $1/3$ is assigned. In the present example, a probability law in V given by

$$P(s) = \begin{cases} 1/6 & \text{when } s \in \{\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3\}, \\ 0 & \text{otherwise,} \end{cases}$$

would be consistent with the assumptions of the FD model. Thus, in summary:

Model	Domain of the Uniform Law
Maxwell-Boltzmann	V
Fermi-Dirac	V/G
Bose-Einstein	V_I/G

4.1. **The canonical distribution for the Maxwell-Boltzmann model.** Think of a very large set $L = \{1, 2, \dots, \ell\}$ of numbered gas molecules and their possible energy states, represented here by the elements in set $\{\mathcal{E}_1, \mathcal{E}_2, \dots\}$. Each energy configuration corresponds to a map

$$s : L \rightarrow C.$$

These are the *accessible microstates*. If $f_i = |s^{-1}(\mathcal{E}_i)|$ indicates the number of molecules at the energy level \mathcal{E}_i , then the (constant) mean internal energy of the gas is given by

$$(4.1) \quad \bar{\mathcal{E}} = \frac{1}{\ell} \sum_i \mathcal{E}_i f_i.$$

Before we continue, we will discuss an example similar to that in Comment 2.2.1, with the purpose of showing the extent of the analogies. In the context of biological sequences this is an example with four-sequences in length of six. In the present context, we have six numbered molecules indexed by the set $L = \{1, 2, 3, 4, 5, 6\}$ and four energy levels, indicated by the set $\mathcal{E} = \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4\}$. The energy configurations are maps

$$s : L \rightarrow \mathcal{E},$$

so that there is a total of $|\mathcal{E}|^{|L|} = 4^6 = 4096$ accessible microstates. We pass from microstates to measurable macrostates by dividing the space by similarities that result among the molecules when their numbers are erased. This is in analogy to erasing the position of the nucleotides in a biological sequence. Algebraically, this is obtained by letting the (group S_6 of) permutations act on (by shuffling) the molecule labels in the set L . The resulting classes \mathcal{O}_λ of orbits are then the energy macrostates realized by the system. Here are the resulting classes, their volume $|\mathcal{O}_\lambda|$, usually indicated by Ω_λ in the thermodynamics context, and their number Q_λ of quantal states:

λ	Ω_λ	Q_λ	$\Omega_\lambda \times Q_\lambda$
6000	1	4	4
5100	6	12	72
4200	15	12	180
4110	30	12	360
3300	20	6	120
3210	60	24	1440
3111	120	4	480
2220	90	4	360
2211	180	6	1080
total	522	84	4096

The calculations are outlined in Appendix F. There are $Q = 6$ quantal states associated with the most probable ($\Omega = 180$) orbit type, Ω_{2211} .

In Boltzmann model all particles are considered to be distinguishable, so that a uniform probability can be assigned to each one of them. However, the passage from the accessible microstates to macrostates is equivalent to obtaining a partition of the ensemble V of accessible microstates into orbits of symmetry realized by the symmetric group acting on the label set L . Similarly to the molecular weight example, we see that the mean energy level of any configuration in V is an invariant under these permutations and therefore depends only on the orbit (macrostate) realized by the configuration. Boltzmann reasoned that the molecule-energy configurations in V evolved from least probable configurations to most probable configurations, so that the quest for describing the equilibrium energy distribution in the ensemble requires the determination of the most likely configurations in V . This, in turn, requires the determination of the macrostate (orbit) with the largest volume Ω , conditioned on the fact that mean energy $\bar{\mathcal{E}}$ of the isolated ensemble must remain constant.

Given a configuration s with f_1 particles at the energy level \mathcal{E}_1 , f_2 particles at the level \mathcal{E}_2 , f_3 particles at the level \mathcal{E}_3 , etc, then, its orbit \mathcal{O}_s has volume

$$(4.2) \quad |\mathcal{O}_s| = \frac{\ell!}{f_1!f_2!f_3!\dots}$$

We have then a well-defined mathematical problem: find the macrostate identified by f_1, f_2, \dots which maximizes (4.2) for a given mean energy level $\bar{\mathcal{E}}$. The solution is outlined in what follows. Using Stirling's approximation $\ln t! \equiv t \ln t - t$, we have,

$$\begin{aligned} \ln |\mathcal{O}_s| &= \ln \ell! - \sum \ln f_i \\ &= \ell \ln \ell - \ell - \sum (f_i \ln f_i - f_i) \\ &= \ell \ln \ell - \sum f_i \ln f_i. \end{aligned}$$

Equivalently, then, we seek to minimize $\sum f_i \ln f_i$ subject to (4.1). These two conditions lead to

$$\left(\ell + \sum_i \ln f_i\right) df_i = 0, \quad \sum_i \mathcal{E}_i df_i = 0.$$

A sufficient condition for the existence of a solution (using Lagrange multipliers argument) is that there are constants α and β satisfying

$$\sum_i (\mathcal{E}_i + \alpha + \beta \ln f_i) df_i = 0,$$

in which case the solutions take the form $f_i = \alpha e^{-\beta \mathcal{E}_i}$. The condition $\sum_i f_i = \ell$ implies that

$$\alpha = \frac{\ell}{\sum_i e^{-\beta \mathcal{E}_i}},$$

so that

$$(4.3) \quad f_i = \ell \frac{e^{-\beta \mathcal{E}_i}}{\sum_j e^{-\beta \mathcal{E}_j}}.$$

The value of β follows from the condition $\frac{1}{\ell} \sum_i f_i \mathcal{E}_i = \bar{\mathcal{E}}$. That is, β is a solution of

$$\frac{\sum_i e^{-\beta \mathcal{E}_i} \mathcal{E}_i}{\sum_j e^{-\beta \mathcal{E}_j}} = \bar{\mathcal{E}}.$$

From (4.3) we then obtain Maxwell-Boltzmann canonical distribution,

$$(4.4) \quad \boxed{P(\mathcal{E}_i) = \frac{e^{-\beta \mathcal{E}_i}}{\sum_j e^{-\beta \mathcal{E}_j}}}.$$

The canonical distribution is the most likely energy distribution of the ensemble. Similar calculations can be obtained for the models of Fermi-Dirac and Bose-Einstein.

4.1.1. *Moments of the canonical distribution.* Direct calculation shows that the mean ($\bar{\mathcal{E}}$) and variance ($\text{var}(\mathcal{E})$) of the canonical distribution can be expressed in terms of the *partition function* $Z = \sum_j e^{-\beta \mathcal{E}_j}$ as

$$\bar{\mathcal{E}} = -\frac{\partial \ln Z}{\partial \beta}, \quad \text{var}(\mathcal{E}) = \frac{\partial^2 \ln Z}{\partial \beta^2}.$$

4.1.2. *Boltzmann Entropy.* Note that the constrained minimization of $\sum f_i \ln f_i$ is equivalent to the constrained maximization of

$$H = - \sum_i \frac{f_i}{\ell} \ln\left(\frac{f_i}{\ell}\right)$$

which is the *entropy* of the probability law associated with the orbit of f_1, f_2, f_3, \dots . The entropy, indicated by S in thermodynamics, is a physical characteristic (e.g., temperature, mass) of the gas and at the same time, a measure of uniformity in its thermodynamical probability law. The canonical distribution corresponds to an ensemble configured to its maximum entropy. Boltzmann's statistical expression

$$\boxed{S = k \ln \Omega}$$

for the equilibrium entropy relates the equilibrium or limit number of accessible microstates, Ω , and k , (now known as) Boltzmann constant 1.3807×10^{-23} K J/molecule. A volume of gas, left to itself, will almost always be found in the state of the most probable distribution.

In the Appendix B we give an example of the entropy concept applied to nucleotide sequences e.g., Durbin et al. (1998, p.305).

5. MAXWELL-BOLTZMANN LAW FOR VELOCITIES IN A PERFECT GAS

In this section we outline the classical derivation of Maxwell-Boltzmann Law. The interpretation of this derivation in the context of the orbit method is outlined in Example G.1 of Appendix G. Maxwell's assumptions e.g., Ruhla (1989, Ch.4), led to the searching of a probability law, indicated here by F , for the random velocity vector (\mathbf{v}) satisfying the following conditions: First, the component-velocities are statistically independent and identically distributed, so that the law F should have the form

$$F(\mathbf{v}) = f(v_x)f(v_y)f(v_z),$$

where f indicated the common probability law for the component-velocities. The *isotropic condition* states that F should be invariant under all rotations, indicated here by U , in the three-dimensional Euclidian space \mathbb{R}^3 . Denoting by $S(3, \mathbb{R})$ the collection of all such rotations, we write the isotropic condition as,

$$(5.1) \quad F(U\mathbf{v}) = F(\mathbf{v}), \text{ for all } U \in S(3, \mathbb{R}).$$

Note the analogy between the isotropic condition and expression (2.3).

5.1. **Classical derivation of Maxwell-Boltzmann Law.** The isotropic condition implies that

$$\sum_i \frac{\partial F}{\partial v_i} dv_i = 0, \quad \sum_i v_i dv_i = 0,$$

whereas the condition $F(\mathbf{v}) = f(v_x)f(v_y)f(v_z)$ implies that $\ln F(\mathbf{v}) = \sum_i \ln f(v_i)$, so that

$$\frac{\partial F}{\partial v_i} = \frac{F(\mathbf{v})}{f(v_i)} \frac{df(v_i)}{dv_i}, \text{ for all components of } \mathbf{v},$$

and hence

$$\sum_i v_i dv_i = 0, \quad \sum_i \left(\frac{1}{f(v_i)} \frac{df(v_i)}{dv_i} \right) dv_i = 0,$$

must obtain. The Lagrange multiplier argument leads to

$$\sum_i \left(\lambda v_i + \frac{1}{f(v_i)} \frac{df(v_i)}{dv_i} \right) dv_i = 0,$$

of which a sufficient solution is

$$\lambda v_i = - \frac{1}{f(v_i)} \frac{df(v_i)}{dv_i} = - \frac{d}{dv_i} \ln f(v_i),$$

where λ does not depend on \mathbf{v} . Therefore, it follows that $\ln f(v_i) = -\lambda v_i^2/2 + A$, or that each component law has the form

$$f(v_i) = Ae^{-\lambda v_i^2/2},$$

thus leading to

$$F(\mathbf{v}) = \prod_i f(v_i) = A^3 e^{-\lambda \|\mathbf{v}\|^2/2},$$

where $v = \|\mathbf{v}\| = \sqrt{v_x^2 + v_y^2 + v_z^2}$ is the *speed* in the velocity vector \mathbf{v} . Because the integral $\int_{\mathbb{R}^3} F(\mathbf{v}) d\mathbf{v}$ must be finite (and equal to 1), it follows that $\lambda > 0$. Introducing the notation $\mu^2 = \lambda/2$ we then obtain the final form of Maxwell law,

$$(5.2) \quad F(\mathbf{v}) = A^3 e^{-\mu \|\mathbf{v}\|^2}.$$

5.1.1. *Determining the constants A and μ .* The condition $\int_{\mathbb{R}^3} F(\mathbf{v}) d\mathbf{v} = 1$ implies that

$$1 = A^3 \left[\int_{\mathbb{R}} e^{-\mu^2 v_i^2} dv_i \right]^3 = A^3 \left[\frac{\sqrt{\pi}}{\mu} \right]^3,$$

so that the relation $A\sqrt{\pi} = \mu$ must hold.

Secondly, the mean molecular kinetic energy must be constant and equal to $\frac{3}{2}RT$, that is

$$\int_0^\infty \frac{1}{2} m v^2 dn = \frac{3}{2} RT,$$

where $N = 6.02214199 \times 10^{23} \text{ mol}^{-1}$ is Avogadro constant, $k = 1.3806503 \times 10^{-23} \text{ JK}^{-1}$ is Boltzmann constant, and $R = kN$. We now refer to Example G.1 in Appendix G, using the isotropic condition (5.1). Therefore,

$$n = \int_0^\infty n F(\mathbf{v}) 4\pi v^2 dv,$$

so that $dn = n F(\mathbf{v}) 4\pi v^2 dv$ is the mean number of molecules with speed in the orbit with statistical density $F(\mathbf{v}) 4\pi v^2 dv$. It then follows that

$$\begin{aligned} \frac{3}{2} RT &= \frac{3}{2} kNT = \int_0^\infty \frac{1}{2} m v^2 dn = \int_0^\infty \frac{1}{2} m v^2 n F(\mathbf{v}) 4\pi v^2 dv \\ &= \frac{1}{2} mn 4\pi A^3 \int_0^\infty v^4 e^{-\mu v^2} dv = \frac{1}{2} mn 4\pi A^3 \frac{3\sqrt{\pi}}{8\mu^5}, \end{aligned}$$

so that we obtain the relation

$$mA^3 \frac{\pi^{3/2}}{2\mu^5} = kT,$$

which, together with $A\sqrt{\pi} = \mu$ obtained earlier, determine the constants, namely,

$$\mu^2 = \frac{m}{2kT}, \quad A = \left(\frac{m}{2\pi kT} \right)^{1/2}.$$

Thus, Maxwell law:

$$\boxed{\rho(v) = 4\pi \left(\frac{m}{2\pi kT} \right)^{3/2} [e^{-(mv^2)/2kT}] v^2, \quad v \geq 0.}$$

6. SUMMARY

The following table summarizes the symmetry relations introduced in these notes. In each case, a uniform probability law is assigned to the points within the corresponding orbits in the quotient space.

Model	Symmetries	Orbits
DNA sequences	all cyclic permutations, C_ℓ	V/C_ℓ
Maxwell-Boltzmann	1 (the identity)	V
Fermi-Dirac	all permutations, S_ℓ	V/S_ℓ
Bose-Einstein	all permutations, S_ℓ	V_1/S_ℓ
Velocity in a perfect gas	all rotations, $O(3, \mathbb{R})$	$\mathbb{R}^3/O(3, \mathbb{R})$

7. RELATED READING

- (1) Rosen (1995) and Rosen (1975) on symmetry and the symmetry principle;
- (2) Hellige (1993) on hemispherical asymmetry;
- (3) The notion of points as labels identifying potential events appears in modern-day physics, in contrast to Newton's views in which points are essentially indistinguishable. A comment in that direction is found in Cartier (2001).

APPENDIX A. THE HUMAN IMMUNODEFICIENCY VIRUS TYPE I

Here is a fragment of the entire nucleotide sequence. To locate the sequence in the NCBI⁴ data base, use the accession number K02013. Figures A.1 to A.4 show the observed relative frequencies of the indicated (in the vertical axis) residue evaluated at each of 45 intervals in length of 200 residues (the global sequence is approximately 9000 bp-long). Estimated based on two isolates (BRU and OYI) of the same virus are shown. The BRU isolate is 9229 bp (base-pair) long and the OYI is 9102 bp long (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 ctccctggga tattgatgat ctgtagtgtc acagaaaaat tgtgggtcac agtctattat
121 ggggtacctg tgtggaagga agcaaccacc actctatfff gtgcatcaga tgctaaagca
181 tatgatacag aggtacataa tgtttgggcc acacatgcct gtgtaccac agaccccaac
241 ccacaagaag tagtattggt aaatgtgaca gaaaatttta acatgtggaa aaatgacatg
301 gtagaacaga tgcattgagg tataatcagt ttatgggatc aaagcctaaa gccatgtgta
361 aaattaaccc cactctgtgt tagtttaaag tgcactgatt tggggaatgc tactaatacc
421 aatagtagta ataccaatag tagtagcggg gaaatgatga tggagaaagg agagataaaa
481 aactgctctt tcaatatcag cacaagcata agaggtaagg tgcagaaaga atatgcatff
541 tttataaac ttgatataat accaatagat aatgatacta ccagctatac gttgacaagt
601 tgtaacacct cagtcattac acaggcctgt ccaaaggat cctttgagcc aattcccata
661 cattattgtg ccccgctggg ttttgcgatt ctaaaatgta ataataagac gttcaatgga
721 acaggacat gtacaaatgt cagcacagta caatgtacac atggaattag gccagtagta
781 tcaactcaac tgctgttgaa tggcagtcta gcagaagaag aggtagtaat tagactctgc
841 aatttcacag acaatgctaa aaccataata gtacagctga accaatctgt gaaaattaat
901 tgtacaagac ccaacaaca tacaagaaaa agtatccgta tccagagggg accagggaga
961 gcatttgttt caataggaaa aataggaat atgagacaag cacattgtaa cattagtaga
1021 gcaaaatgga atgccacttt aaaacagata gctagcaaat taagagaaca atttggaaat
1081 aataaaacaa taatctttaa gcaatcctca ggaggggacc cagaaattgt aacgcacagt
1141 tttaatgtg gaggggaatt tttctactgt aattcaacac aactgtttaa tagtacttgg
1201 tttaatagta cttggagtac tgaagggtca aataacactg aaggaaatga cacaatcaca
1261 ctcccattga gaataaaaaca atttataaac atgtggcagg aagtaggaaa agcaatgtat
1321 gccctccca tcaggcgaca aattagatgt tcatcaataa ttacagggct gctattaaca
1381 agagatggtg gtaataacaa caatgggtcc gagatcttca gacctggagg aggagatatg
1441 agggacaatt ggagaagtga attatataaa tataaagtag taaaaattga accattagga
1501 gtagcaccca ccaaggcaaa gagaagagtg gtgcagagag aaaaaagagc agtgggaata
1561 ggagctttgt tccttggggt cttgggagca gcaggaagca ctatgggagc acggtcaatg
1621 acgctgacgg tacaggccag acaattattg tctggtatag tgcagcagca gaacaattt
1681 ctgaggcgta ttgaggcgca acagcatctg ttgcaactca cagtctgggg catcaagcag
1741 ctccaggcaa gaatcctggc tgtggaaaga tacctaaagg atcaacagct cctggggatt
1801 tggggttctg ctgaaaaact catttgcacc actgctgtgc cttggaatgc tagttggagt
1861 aataaatctc tggaacagat ttggaataac atgacctgga tggagtggga cagagaaatt
1921 aacaattaca caagcttaat acattcctta attgaagaat cgcaaaacca gcaagaaaag
1981 aatgaacaag aattattgga attagataaa tgggcaagtt tgtggaattg gtttaacata
2041 acaaatggc tgtggtatat aaaaatattc ataagatag taggaggctt ggtaggttta
2101 agaatagttt ttgctgtact ttctatagtg aatagagtta ggcagggata ttcaccatta
2161 tcgtttcaga cccacctccc aacccgagg ggacccgaca ggcccgaagg aatagaagaa
2221 gaagtgagg agagagacag agacagatcc attcgattag tgaacggatc cttagcactt
2281 atctgggac atctgaggag cctgtgcctc ttcagctacc accgctttag agacttactc
2341 ttgattgtaa cgaggattgt ggaactctg ggacgcaggg ggtgggaagc cctcaaatat
2401 tggtggaatc ctctacagta ttggagttag gaaactaaaga atagtctgt tagcttctc
2461 aatgccacag ccatagcagt agctgagggg acagataggg ttatagaagt agtacaagga
2521 gctttagtag ctattcgcca catacctaga agaataagac agggcttggg aaggattttg
2581 ctataa

```

⁴National Center for Biotechnology Information, <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide>

FIGURE A.1. Adenine distribution in the BRU isolate K02013 (top) and OYI isolate M26727 (bottom).

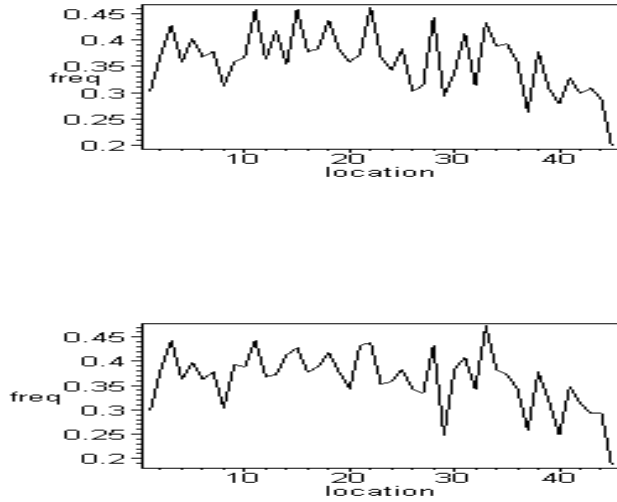


FIGURE A.2. Cytosine distribution in the BRU isolate K02013 (top) and OYI isolate M26727 (bottom).

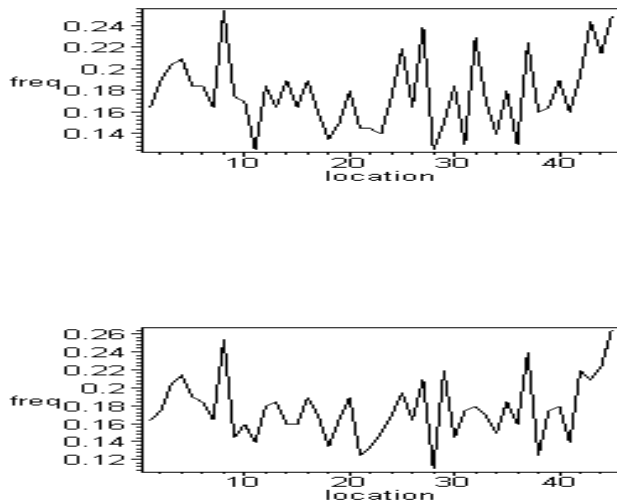


FIGURE A.3. Guanine distribution in the BRU isolate K02013 (top) and OYI isolate M26727 (bottom).

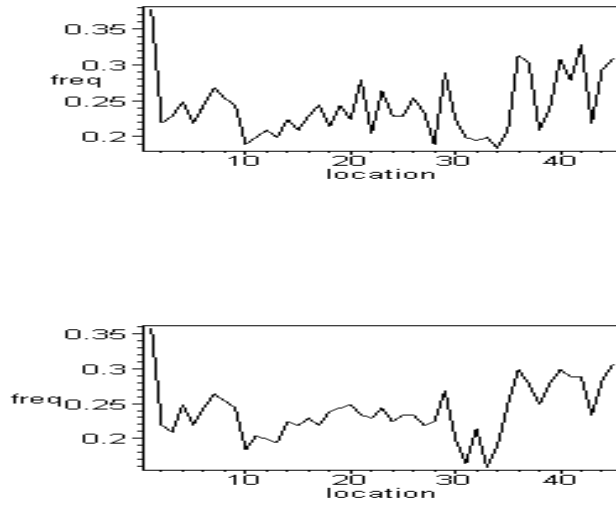
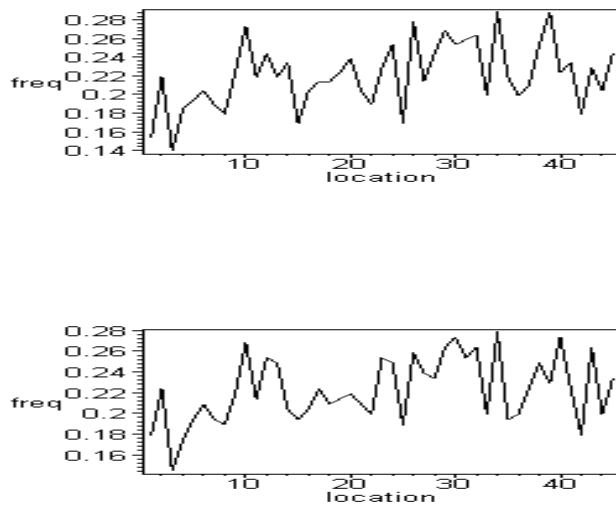


FIGURE A.4. Thymine distribution in the BRU isolate K02013 (top) and OYI isolate M26727 (bottom).



APPENDIX B. ENTROPY AND INFORMATION CONTENT

The *entropy* of the probability law⁵ P is

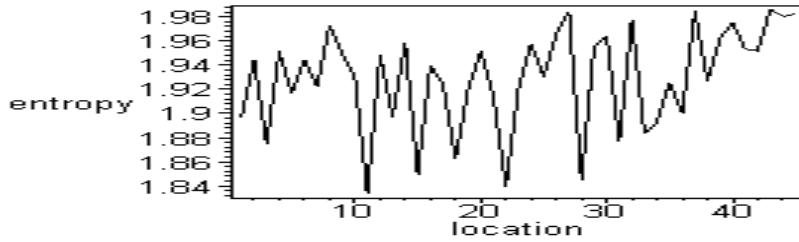
$$H(P) = - \sum_s P(s) \log_2 P(s) \leq \log_2 |\mathcal{A}|,$$

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*⁶. Figure B.1 shows the entropy in the relative frequency distributions

$$\frac{1}{40}(f_a, f_c, f_g, f_t)$$

in 45 consecutive local sequences in length of 200 for the BRU isolate of the human immunodeficiency virus type 1, HIV-1, (K02013). The *information content* at a given 40-base pairs long 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 B.1. Four-base entropy along the BRU isolate K02013.



APPENDIX C. PERMUTATIONS

A permutation τ is a 1-1 map $\tau : L \rightarrow L$. We indicate the set of all permutations on a finite set L with ℓ elements by S_ℓ . We say that a permutation τ *fixes* the element $j \in L$ if $\tau(j) = j$. The *identity* transformation, indicated by 1, fixes all elements in L. Given two permutations τ and σ , the *composite* function $\tau\sigma$ takes the element $j \in L$ to the element $\tau(\sigma(j))$ in L, and is also a permutation. For every permutation τ in L, the equation $\tau(j') = j$ has a unique solution j' for each $j \in L$. The resulting function $j \mapsto j'$ is also a permutation, called the *inverse* permutation, and is indicated by τ^{-1} . It holds that $\tau\tau^{-1} = \tau^{-1}\tau = 1$. Adding the fact that composition of functions is an associative operation, that is $(\tau\sigma)\eta = \tau(\sigma\eta)$, we observe that the set S_ℓ together with the operation of function composition, defines a *permutation group* of order ℓ .

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

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

⁵The base of the logarithm is irrelevant- when the base is 2 the entropy is usually expressed in units called *bits*.

⁶Historically, 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).

three transpositions,

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

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:

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

APPENDIX D. SYMMETRIES IN TWO-SEQUENCES IN LENGTH OF FOUR

The space V of all 2-sequences in length of 4 is represented by

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

with 16 points (the numbers on the first row are labels for each map). The following matrix (D.1) shows all 24 elements (σ) of S_4 , the group of permutations of $\{1, 2, 3, 4\}$, the 16 two-sequences (s), and the sequences ($s\sigma$) that resulted from shuffling the sequence s according to the permutation σ . Also shown are the values

of $|\text{fix}(\sigma)|$ and of the orbit stabilizers $|\mathbf{G}_s|$.

(D.1)

$\sigma \setminus s$	1	16	15	14	12	8	13	11	7	10	6	4	9	5	3	2	$ \text{fix}(\sigma) $
1	1	16	15	14	12	8	13	11	7	10	6	4	9	5	3	2	16
(34)	1	16	15	14	8	12	13	7	11	6	10	4	5	9	3	2	8
(23)	1	16	15	12	14	8	11	13	7	10	4	6	9	3	5	2	8
(24)	1	16	15	8	12	14	7	11	13	4	6	10	3	5	9	2	8
(12)	1	16	14	15	12	8	13	10	6	11	7	4	9	5	2	3	8
(13)	1	16	12	14	15	8	10	11	4	13	6	7	9	2	3	5	8
(14)	1	16	8	14	12	15	6	4	7	10	13	11	2	5	3	9	8
(234)	1	16	15	12	8	14	11	7	13	4	10	6	3	9	5	2	4
(243)	1	16	15	8	14	12	7	13	11	6	4	10	5	3	9	2	4
(123)	1	16	14	12	15	8	10	13	6	11	4	7	9	2	5	3	4
(124)	1	16	14	8	12	15	6	10	13	4	7	11	2	5	9	3	4
(132)	1	16	12	15	14	8	11	10	4	13	7	6	9	3	2	5	4
(134)	1	16	12	14	8	15	10	4	11	6	13	7	2	9	3	5	4
(142)	1	16	8	15	12	14	7	4	6	11	13	10	3	5	2	9	4
(143)	1	16	8	14	15	12	6	7	4	13	10	11	5	2	3	9	4
(12)(34)	1	16	14	15	8	12	13	6	10	7	11	4	5	9	2	3	4
(13)(24)	1	16	12	8	15	14	4	11	10	7	6	13	3	2	9	5	4
(14)(23)	1	16	8	12	14	15	4	6	7	10	11	13	2	3	5	9	4
(1234)	1	16	14	12	8	15	10	6	13	4	11	7	2	9	5	3	2
(1243)	1	16	14	8	15	12	6	13	10	7	4	11	5	2	9	3	2
(1324)	1	16	12	8	14	15	4	10	11	6	7	13	2	3	9	5	2
(1342)	1	16	12	15	8	14	11	4	10	7	13	6	3	9	2	5	2
(1432)	1	16	8	15	14	12	7	6	4	13	11	10	5	3	2	9	2
(1423)	1	16	8	12	15	14	4	7	6	11	10	13	3	2	5	9	2
$ \mathbf{G}_s $	24	24	6	6	6	6	4	4	4	4	4	4	6	6	6	6	

APPENDIX E. COUNTING ORBITS

We denote by $\text{fix}(\sigma) = \{s \in V; s\sigma = s\}$ the set of sequences that remain fixed by the permutation σ . It then follows that

(E.1)
$$\text{Number of orbits of } V = \frac{1}{|\mathbf{G}|} \sum_{\mathbf{G}} |\text{fix}(\sigma)|.$$

This is a well-known result in combinatorics (also known as Burnside's Lemma⁷). We also defined

$$\mathbf{G}_s = \{\tau \in \mathbf{G}; s\tau = s\},$$

which is called the *orbit stabilizer* of s by \mathbf{G} . It then follows that

(E.2)
$$|\mathbf{G}| = |\mathcal{O}_s| |\mathbf{G}_s|.$$

⁷William Burnside, Born: 2 July 1852 in London, England. Died: 21 Aug 1927 in West Wickham, London, England. Among his applied mathematics teachers at Cambridge were Stokes, Adams and Maxwell.

Example E.1. From matrix (D.1), with $G = S_4$, it follows that

$$\text{Number of orbits of } V = \frac{1}{|G|} \sum_G |\text{fix}(\sigma)| = \frac{120}{24} = 5,$$

namely,

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

In addition, because $|\mathcal{O}_i| = |G|/|G_{s_i}|$, we have

$$\begin{aligned} |\mathcal{O}_0| &= 24/24 = 1, \\ |\mathcal{O}_1| &= 24/6 = 4, \\ |\mathcal{O}_2| &= 24/4 = 6, \\ |\mathcal{O}_3| &= 24/6 = 4, \\ |\mathcal{O}_4| &= 24/24 = 1. \end{aligned}$$

APPENDIX F. ORBITS AND THEIR VOLUMES

It is convenient here to include the multiplicities (m_1, \dots, m_k) with which the distinct components a_1, \dots, a_k appear in each partition λ , so that we may write

$$\lambda = (a_1^{m_1}, \dots, a_k^{m_k}).$$

In the six-molecule, four-level energy example ($\ell = 6, c = 4$) we have

λ	Ω_λ	Q_λ	$\Omega_\lambda \times Q_\lambda$
6000 = $6^1 0^3$	1	4	4
5100 = $5^1 1^1 0^2$	6	12	72
4200 = $4^1 2^1 0^2$	15	12	180
4110 = $4^1 1^2 0^1$	30	12	360
3300 = $3^2 0^2$	20	6	120
3210 = $3^1 2^1 1^1 0^1$	60	24	1440
3111 = $3^1 1^3$	120	4	480
2220 = $2^3 0^1$	90	4	360
2211 = $2^2 1^2$	180	6	1080
total	522	84	4096

It then follows that

$$c^\ell = \sum_\lambda \frac{\ell!}{(a_1!)^{m_1} (a_2!)^{m_2} \dots (a_k!)^{m_k}} \frac{c!}{m_1! m_2! \dots m_k!},$$

where λ varies over the (m) different frames, that is, $m_1 a_1 + \dots + m_k a_k = \ell$ and $m_1 + \dots + m_k = c$. Moreover

$$(F.1) \quad \Omega_\lambda = \frac{\ell!}{(a_1!)^{m_1} (a_2!)^{m_2} \dots (a_k!)^{m_k}}, \quad Q_\lambda = \frac{c!}{m_1! m_2! \dots m_k!},$$

and

$$\sum_\lambda Q_\lambda = \binom{c+\ell-1}{\ell}$$

decomposes the Bose-Einstein count $\binom{\ell+c-1}{\ell}$.

APPENDIX G. MORE ON CALCULUS WITH ORBITS OF SYMMETRY- CROSS SECTIONS

Let P be a probability law in V . We will refer to the map space V of two-sequences in length of four, described by the matrix in expression (D.1). We will show how calculus of probabilities in V can be expressed by a *changing of variable* so that we integrate over the entire space by integrating over the orbits in V generated by the symmetries in a group of symmetries, which, in the present example, consists of all permutations in S_4 . To illustrate, we will simply rewrite the equality

$$1 = \sum_{\mathbf{s}} P(\mathbf{s})$$

as outlined above. For that end, recall, from Section E, that

$$V = \mathcal{O}_0 \cup \mathcal{O}_1 \cup \mathcal{O}_2 \cup \mathcal{O}_3 \cup \mathcal{O}_4,$$

where

$$\begin{aligned} \mathcal{O}_0 &= \{s_1\}, \\ \mathcal{O}_1 &= \{s_9, s_5, s_3, s_2\}, \\ \mathcal{O}_2 &= \{s_{13}, s_{11}, s_7, s_{10}, s_6, s_4\}, \\ \mathcal{O}_3 &= \{s_{15}, s_{14}, s_{12}, s_8\}, \\ \mathcal{O}_4 &= \{s_{16}\}, \end{aligned}$$

and that any two of these orbits are disjoint. Consequently, it is clear that

$$(G.1) \quad 1 = \sum_{\mathbf{s}} P(\mathbf{s}) = \sum_{i=0}^4 P(\mathcal{O}_i).$$

Now select a set, Γ , of single representatives of each orbit, say $\Gamma = \{s_1, s_9, s_{13}, s_{15}, s_{16}\}$. Because Γ intercepts each orbit in exactly one point of V , this set is called a *cross section* of V . From matrix (D.1), we see that

$$\sum_{\tau} P(s_1\tau) = 24 \times P(\mathcal{O}_0) = |G_{s_0}| P(\mathcal{O}_0),$$

where $|G_{s_0}|$ indicates the multiplicity with which $s_0 \in \Gamma$ is presented in its orbit by the action of S_4 . Similarly, we observe that

$$\begin{aligned} \sum_{\tau} P(s_9\tau) &= 6 \times P(\mathcal{O}_1) = |G_{s_1}| P(\mathcal{O}_1), \\ \sum_{\tau} P(s_{13}\tau) &= 4 \times P(\mathcal{O}_2) = |G_{s_{13}}| P(\mathcal{O}_2), \\ \sum_{\tau} P(s_{15}\tau) &= 6 \times P(\mathcal{O}_3) = |G_{s_{15}}| P(\mathcal{O}_3), \\ \sum_{\tau} P(s_{16}\tau) &= 24 \times P(\mathcal{O}_4) = |G_{s_1}| P(\mathcal{O}_4). \end{aligned}$$

Consequently, we may rewrite (G.1) in the form

$$1 = \sum_{\mathbf{s}} P(\mathbf{s}) = \sum_{i=0}^4 P(\mathcal{O}_i) = \sum_{i=0}^4 \sum_{\tau} \frac{P(s_i\tau)}{|G_{s_i}|}.$$

Finally, because $|G| = |\mathcal{O}_s| |G_s|$ (see (E.2)), we have

$$(G.2) \quad 1 = \sum_{\mathbf{s}} P(\mathbf{s}) = \sum_{i=0}^4 |\mathcal{O}_i| \sum_{\tau} P(s_i\tau) \frac{1}{|G|},$$

so that, in general, we may write

$$1 = \sum_{\mathbf{s}} P(\mathbf{s}) = \sum_{\mathbf{s} \in \Gamma} |\mathcal{O}_{\mathbf{s}}| \sum_{\tau \in G} P(\mathbf{s}_i \tau) \frac{1}{|G|}.$$

Similarly, for all measurements \mathbf{x} in V with the property that \mathbf{x} is determined by its values on the orbits alone, that is, $\mathbf{x}(\mathbf{s}) = \mathbf{x}(s\tau)$ for all $\tau \in G$, we verify that its mean value can be expressed as

$$(G.3) \quad \bar{\mathbf{x}} = \sum_{\mathbf{s}} \mathbf{x}(\mathbf{s}) P(\mathbf{s}) = \sum_{\mathbf{s} \in \Gamma} \mathbf{x}(\mathbf{s}) |\mathcal{O}_{\mathbf{s}}| \Delta(\mathbf{s}),$$

where

$$\Delta(\mathbf{s}) = \sum_{\tau \in G} P(s\tau) \frac{1}{|G|}, \quad \mathbf{s} \in \Gamma.$$

Example G.1. Here is a simple example. It appears in Maxwell law for the statistical velocity of molecules in a perfect gas e.g., (Ruhla 1989, Ch.4). We make the following identification:

- (1) $V = \mathbb{R}^3$, the three-dimensional Euclidean space with elements indicated by $\mathbf{v} = (v_x, v_y, v_z)$;
- (2) $G = O(3, \mathbb{R})$, the group of all central rotations (τ) in V . That is, of all (Euclidean) distance-preserving linear transformations;
- (3) The orbits in V generated by G are spheres of radius v , that is, $\mathcal{O}_v = \{\mathbf{v} \in V; \|\mathbf{v}\| = v\}$;
- (4) Each orbit has volume $|\mathcal{O}_v| = 4\pi v^2 = \int_{\phi=0}^{\pi} \int_{\theta=0}^{2\pi} \frac{\partial(v_x, v_y, v_z)}{\partial(v, \theta, \phi)} d\theta d\phi = \int_{\phi=0}^{\pi} \int_{\theta=0}^{2\pi} v^2 \sin \theta d\theta d\phi$, where (v, θ, ϕ) are the corresponding spherical⁸ coordinates obtained from $v_x = v \cos \theta \sin \phi$, $v_y = v \sin \theta \sin \phi$ and $v_z = v \cos \phi$;
- (5) The cross section is simply $\Gamma = \{v \in \mathbb{R}, v \geq 0\}$;
- (6) $\Delta(\mathbf{s}) = dv$.

Therefore, if F is any euclidian distance-preserving function in V , we have

$$\int_V F(\mathbf{v}) d\mathbf{v} = \int_{\Gamma} F(v) |\mathcal{O}_v| dv = \int_0^{\infty} F(v) 4\pi v^2 dv.$$

⁸In the spherical coordinate system, v is the radius of the vector \mathbf{v} , which gives distance from the origin, the angle ϕ between \mathbf{v} and the z axis, and the angle θ , measured between the x axis and the projection of \mathbf{v} in the x,y plane.

REFERENCES

- Cartier, P. (2001), 'A mad day's work: from Grothendiek to Connes and Kontsevich- the evolution of concepts of space and symmetry', *Bulletin (New Series) of the American Mathematical Society* **38**(4), 389–408.
- 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.
- Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998), *Biological Sequence Analysis*, Cambridge University Press, Cambridge, UK.
- Hellige, J. B. (1993), *Hemispheric Asymmetry*, Harvard U. Press, Cambridge, MA.
- Rosen, J. (1975), *Symmetry Discovered*, Dover, Mineola, NY.
- Rosen, J. (1995), *Symmetry in Science, An Introduction to the General Theory*, Springer-Verlag, New York.
- Ruhla, C. (1989), *The Physics of Chance*, Oxford Press, New York, NY.
- von Mises, R. (1957), *Probability, Statistics and Truth*, Dover, New York, NY.