mathematics

MDPI

Article

A Rényi-Type Limit Theorem on Random Sums and the Accuracy of Likelihood-Based Classification of Random Sequences with Application to Genomics

Leonid Hanin [1,*] and Lyudmila Pavlova [2]

[1] Department of Mathematics and Statistics, Idaho State University, 921 S. 8th Avenue, Stop 8085, Pocatello, ID 83209-8085, USA
[2] School of Applied Mathematics and Computational Physics, Peter the Great St. Petersburg Polytechnic University, Polytechnicheskaya ul. 29, 195251 St. Petersburg, Russia; lyu0510@gmail.com
* Correspondence: hanin@isu.edu

Abstract: We study classification of random sequences of characters selected from a given alphabet into two classes characterized by distinct character selection probabilities and length distributions. The classification is based on the sign of the log-likelihood score (LLS) consisting of a random sum and a random term depending on the length distributions for the two classes. For long sequences selected from a large alphabet, computing misclassification error rates is not feasible either theoretically or computationally. To mitigate this problem, we computed limiting distributions for two versions of the normalized LLS applicable to long sequences whose class-specific length follows a translated negative binomial distribution (TNBD). The two limiting distributions turned out to be plain or transformed Erlang distributions. This allowed us to establish the asymptotic accuracy of the likelihood-based classification of random sequences with TNBD length distributions. Our limit theorem generalizes a classic theorem on geometric random sums due to Rényi and is closely related to the published results of V. Korolev and coworkers on negative binomial random sums. As an illustration, we applied our limit theorem to the classification of DNA sequences contained in the genome of the bacterium *Bacillus subtilis* into two classes: protein-coding genes and standard noncoding open reading frames. We found that TNBDs provide an excellent fit to the length distributions for both classes and that the limiting distributions capture essential features of the normalized empirical LLS fairly well.

Keywords: Rényi theorem; sequence classification; classification accuracy; random sum; translated negative binomial distribution; Kullback–Leibler distance; Erlang distribution; protein-coding gene; open reading frame

MSC: 60F05; 92D20

Citation: Hanin, L.; Pavlova, L. A Rényi-Type Limit Theorem on Random Sums and the Accuracy of Likelihood-Based Classification of Random Sequences with Application to Genomics. *Mathematics* **2023**, *11*, 4254. https://doi.org/10.3390/math11204254

Academic Editors: Irina Shevtsova and Victor Korolev

Received: 21 August 2023
Revised: 6 October 2023
Accepted: 8 October 2023
Published: 11 October 2023

1. Introduction

This study concerns classification of sequences of characters selected randomly and independently of each other from a given alphabet of $M \geq 2$ characters. The length, N, of any such sequence is assumed to be a random variable (rv) independent of the sequence content. Suppose that there are two models of sequence assembly: one where characters are selected from the alphabet with positive probabilities $p(1), p(2), \ldots, p(M)$ and the sequence length has a certain distribution P (model A), and another where characters are selected with positive probabilities $q(1), q(2), \ldots, q(M)$ and the length has a distribution Q (model B). The two vectors of character selection probabilities (or equivalently, probability measures on $\{1, 2, \ldots, M\}$) are assumed to be distinct and will be denoted by $\mathcal{P}$ and $\mathcal{Q}$. Then, the model-generating probabilities are $P_A = \mathcal{P} \times P$ and $P_B = \mathcal{Q} \times Q$.

The sequence classification problem consists of deciding, for a given sequence of characters $C = (C_1, C_2, \ldots, C_n)$, which model generated this sequence. To simplify our

notation, in what follows we will adopt the following convention: if C_k is the n_k-th character of the alphabet, then we will write $p(C_k) = p(n_k)$ and likewise $q(C_k) = q(n_k)$, $1 \le k \le n$. A natural approach to solving the classification problem is to compare the likelihoods of a sequence C associated with models A and B:

$$L_A(C) = P(n)\Pi_{k=1}^{n} p(C_k) \quad \text{and} \quad L_B(C) = Q(n)\Pi_{k=1}^{n} q(C_k). \tag{1}$$

Specifically, if $L_A(C) > L_B(C)$, then we decide that sequence C is generated by model A, while in the case where $L_B(C) > L_A(C)$, the sequence C is attributed to model B (in the unlikely case where $L_A(C) = L_B(C)$, the sequence C is not assigned to any model). Equivalently, denoting by $\mathcal{L}(C)$ the log-likelihood

$$\mathcal{L}(C) = \log \frac{L_A(C)}{L_B(C)}, \tag{2}$$

we classify sequence C as being generated by model A if $\mathcal{L}(C) > 0$ and by model B if $\mathcal{L}(C) < 0$.

Formulas (1) and (2) suggest that the log-likelihood ratio for a randomly and independently generated sequence $(C_1, C_2, \ldots, C_N)$ of random length N is a rv

$$X = \log \frac{P(N)}{Q(N)} + \sum_{n=1}^{N} \log \frac{p(C_n)}{q(C_n)} = f(N) + U, \tag{3}$$

where $f(N) = \log[P(N)/Q(N)]$ and

$$U = \sum_{n=1}^{N} X_n \tag{4}$$

is a random sum generated by independent and identically distributed (iid) rvs

$$X_n = \log \frac{p(C_n)}{q(C_n)}.$$

The expected value of rvs X_n for sequences generated by models $\mathcal{P}$ and $\mathcal{Q}$ is given by

$$\mu_{\mathcal{P}} = \sum_{m=1}^{M} p(m) \log \frac{p(m)}{q(m)} \quad \text{and} \quad \mu_{\mathcal{Q}} = \sum_{m=1}^{M} q(m) \log \frac{p(m)}{q(m)}\,. \tag{5}$$

It follows from Jensen's inequality [1] that $\mu_{\mathcal{Q}} < 0$, hence also $\mu_{\mathcal{P}} > 0$. Note that $\mu_{\mathcal{P}}$ represents the Kullback–Leibler distance [2] between distributions $\mathcal{P}$ and $\mathcal{Q}$: $\mu_{\mathcal{P}} = d_{KL}(\mathcal{P}, \mathcal{Q})$ and similarly $\mu_{\mathcal{Q}} = -d_{KL}(\mathcal{Q}, \mathcal{P})$. We denote by $\sigma^2_{\mathcal{P}}$ and $\sigma^2_{\mathcal{Q}}$ the corresponding variances.

An alternative way of looking at rv U arises from the following observation. For $1 \le m \le M$, denote by ν_m the number of occurrences of the m-th letter of the alphabet in a random sequence of length N. Then, $\nu_1 + \nu_2 + \ldots + \nu_M = N$ and

$$U = \sum_{m=1}^{M} \nu_m \log \frac{p(m)}{q(m)}. \tag{6}$$

Suppose the sequence is generated by model A. Note that, conditional on $N = n$, the random vector $(\nu_1, \nu_2, \ldots, \nu_M)$ follows the multinomial distribution $Mult(n; p(1), p(2), \ldots p(M))$. In particular, the distribution of rv ν_m is binomial $B(n, p(m))$, $1 \le m \le M$. Then, from Formulas (4)–(6), we obtain the following expression for the expected value of rv U under model A:

$$\mathbb{E}_A U = \sum_{n=1}^{\infty} P(N = n) \sum_{m=1}^{M} np(m) \log \frac{p(m)}{q(m)} = \mu_{\mathcal{P}} \mathbb{E}_P N. \tag{7}$$

Similarly, under model B we have $\mathbb{E}_B U = \mu_Q \mathbb{E}_Q N$. Thus, rv U is closely related to the multinomial process with M outcomes and a random number of replications. The above formulas for the expectation of rv U under models A and B can also be obtained directly by applying Wald's identity [3] to the random sum (4).

Computation of various measures of classification accuracy including important misclassification error rates $P_A(X \leq 0)$ and $P_B(X \geq 0)$ requires the knowledge of the distribution of the log-likelihood score X under models A and B. However, in applications with a large alphabet size, computing these distributions for long sequences, let alone sequences of variable length, in closed form is a daunting task. This motivates studying approximations to the model-specific distributions of rv X arising for very long sequences. In this article, such approximations will be derived from the asymptotic distributions of two normalized versions of rv X,

$$Y = \frac{X}{\mathbb{E}X} \quad \text{and} \quad Z = \frac{X - \mathbb{E}X}{\sigma(X)}, \tag{8}$$

where, as usual, $\mathbb{E}X$ is the expected value of rv X and $\sigma(X)$ is its standard deviation under a given model of sequence assembly. The two asymptotic distributions are identified in Theorem 1 (see below). This theorem implies (see Section 4) that the two misclassification error rates for very long sequences are negligible, i.e., that the likelihood-based classification rule is asymptotically accurate (Theorem 2).

As an illustration of our results, we consider in Section 5 the classification of sequences of triplets of nucleotides contained in the deoxyribonucleic acid (DNA) of a given organism as protein-coding genes or noncoding open reading frames (ORFs). This classification problem is central in computational gene finding for newly sequenced or incompletely annotated genomes [4–6]. One of the most powerful tools used for this purpose is Hidden Markov models, see, e.g., [5–8]. In this setting, triplets of nucleotides (or individual nucleotides) generated by the *same* hidden state are emitted independently and have a random length, i.e., they meet our model assumptions.

In many cases of practical interest, sequences of characters must be sufficiently long. In the case of protein-coding genes, this is due to the fact that, in order to perform various biological functions, e.g., to serve as enzymes, proteins must have certain structural features that can only arise if they contain sufficiently many amino acids. Let $\ell \geq 1$ be the minimum allowed length, then $N \geq \ell$ with probability 1.

The limiting distribution of rvs Y and Z will be obtained in the case where the sequence length in models A and B follows respective translated negative binomial distributions (TNBDs) $NB(a, p) + \alpha$ and $NB(b, q) + \beta$, where $0 < p, q < 1$ and a, b, α, β are integers such that $a, b \geq 1$ and $\alpha, \beta \geq 0$. Recall that there are two closely related kinds of negative binomial distributions $NB(r, p)$. The first is the distribution of the "waiting time" to r-th "success" in a sequence of Bernoulli trials with the success probability p *including* the first r successes, while the second is the distribution of the number of "failures" preceding the r-th success. The latter distribution has a natural extension, sometimes called the Pólia distribution, for any real number $r > 0$ [9]. For compelling biological reasons associated with the structure of genes and elucidated in Section 5, see also [10], we will be modeling the length of DNA segments using negative binomial distributions $NB(r, p)$ of the *first kind* with integer $r \geq 1$. Thus,

$$P(N = n) = \binom{n - \alpha - 1}{a - 1} p^a (1 - p)^{n - \alpha - a}, \quad n \geq a + \alpha, \tag{9a}$$

and similarly

$$Q(N = n) = \binom{n - \beta - 1}{b - 1} q^b (1 - q)^{n - \beta - b}, \quad n \geq b + \beta. \tag{9b}$$

In what follows, the minimum sequence length under the two models will be assumed the same:

$$\ell = a + \alpha = b + \beta. \tag{10}$$

Parameters a, b, α, β of TNBDs, related to each other through Formula (10), will be assumed to be fixed. Therefore, limiting distributions of rvs Y and Z for very long sequences under models A and B will be computed under the conditions $p \to 0$ and $q \to 0$.

Our main goal in Sections 2 and 3 is to prove the following limit theorem. To formulate it, recall that the Erlang distribution $E(a, \lambda)$ is a gamma distribution $G(a, \lambda)$ with an integer shape parameter a. Also, if S is a probability distribution on $\mathbb{R}$ and $\tau \in \mathbb{R}$, then $S + \tau$ denotes the translated distribution and $-S$ stands for the distribution S reflected about the origin.

Theorem 1. *Suppose the sequence length distributions under models A and B are* $P = NB(a, p) + \alpha$ *and* $Q = NB(b, q) + \beta$, *respectively, with* $a + \alpha = b + \beta = \ell$.

(i) If $p,\ q \to 0$ *in such a way that* $p \log q \to 0$, *then under model A, rvs Y and Z converge in distribution to* $E(a, a)$ *and* $E(a, \sqrt{a}) - \sqrt{a}$, *respectively;*

(ii) If $p,\ q \to 0$ *in such a way that* $q \log p \to 0$, *then under model B, rvs Y and Z converge in distribution to* $E(b, b)$ *and* $\sqrt{b} - E(b, \sqrt{b})$, *respectively.*

In the case where rv X is just the random sum U, see Formula (4), the limit theorem for plain (untranslated) negative binomial distributions was known previously. Specifically, for $a = 1$, the fact that the limiting distribution of rv Y is $Exp(1)$ represents a classic theorem due to Rényi [11], see also [12]. A generalization of Rényi's theorem to negative binomial distributions $NB(r, p)$ of the second kind with arbitrary $r > 0$ was obtained by Korolev and Zeifman [13] based on an estimate of the Zolotarev distance [14] between the distributions of the normalized random sum U and $E(r, r)$; for a review of relevant results and methodology, see the article by Korolev [9] and references therein. Although it is probably possible to prove Theorem 1 by reduction to the known limit theorems for the normalized random sum U, we here prefer, for greater insight and the reader's convenience, to give a direct, self-contained, and fairly elementary proof of Theorem 1. In particular, the proof clearly demonstrates that conditions $p \log q \to 0$ and $q \log p \to 0$ in Theorem 1 make the term $f(N)$ in (3) negligible in the limit. The meaning of these conditions is that the expected length of random sequences generated by one model cannot tend to infinity exponentially faster than for sequences generated by the other model.

The article is organized as follows. In Section 2, we study the asymptotic behavior of the expected value and variance of rv X for long sequences generated by models A or B under the conditions of Theorem 1. Section 3 delivers the proof of Theorem 1. In Section 4, we show that, under the conditions of Theorem 1, the likelihood-based classification of random sequences is asymptotically accurate. In Section 5, we delve into genomics and describe in detail the problem of classification of DNA sequences as protein-coding genes or noncoding ORFs using the genome of bacterium *Bacillus subtilis* as an example. In the same section, we estimate the sequence length distributions from data and make a visual comparison of the empirical and theoretical distributions of rvs Y and Z. Finally, in Section 6, we discuss our findings from mathematical and bioinformatics perspectives.

2. Asymptotic Behavior of the Expectation and Variance of the Log-Likelihood Score

Our goal in this section is to establish the following result:

Proposition 1. *Let character selection probabilities under models A and B be governed by the respective probability distributions* $\mathcal{P}$ *and* $\mathcal{Q}$ *with expected values* $\mu_{\mathcal{P}}$ *and* $\mu_{\mathcal{Q}}$. *Also, let the respective sequence length distributions under models A and B be* $P = NB(a, p) + \alpha$ *and* $Q = NB(b, q) + \beta$ *with* $a + \alpha = b + \beta = \ell$. *Suppose that* $p,\ q \to 0$. *Then, for the expected value and variance of the log-likelihood score X under models A and B, we have the following asymptotic relations:*

(i) *If* $p \log q \to 0$, *then* $p\, \mathbb{E}_A X \to a\mu_{\mathcal{P}}$ *and* $p^2 Var_A X \to a\mu_{\mathcal{P}}^2$;
(ii) *If* $q \log p \to 0$, *then* $q\, \mathbb{E}_B X \to b\mu_{\mathcal{Q}}$ *and* $q^2 Var_B X \to b\mu_{\mathcal{Q}}^2$.

The following two lemmas will be instrumental in proving Proposition 1.

Lemma 1. *Let* $\ell \in \mathbb{Z}_+$ *and* $\{P_p(k) : k \geq \ell\}$, $0 < p < 1$, *be a family of probability distributions on* $\{\ell, \ell+1, \ell+2, \dots\}$. *Suppose there is a constant* $C > 0$ *independent of* p *such that* $p\, \mathbb{E}P_p \leq C$ *for all* p. *Then, for any sequence* $\{c(j)\}_{j=0}^{\infty}$ *such that*

$$\frac{c(j)}{j} \to 0 \quad as \quad j \to \infty, \tag{11}$$

we have

$$p \sum_{j=0}^{\infty} c(j) P_p(j+\ell) \to 0 \quad as \quad p \to 0. \tag{12}$$

Proof. We may assume without loss of generality that $c(j) \geq 0$ for all $j \geq 0$. Fix $\varepsilon > 0$. In view of (11), there exists $K \geq 0$, which we will also fix, such that $c(j) \leq j\varepsilon$ for all $j > K$. Let $M_K = \max\{c(j) : 0 \leq j \leq K\}$. Then,

$$p \sum_{j=0}^{\infty} c(j) P_p(j+\ell) = p \sum_{j=0}^{K} c(j) P_p(j+\ell) + p \sum_{j=K+1}^{\infty} c(j) P_p(j+\ell)$$

$$\leq pM_K \sum_{j=0}^{K} P_p(j+\ell) + \varepsilon p \sum_{j=K+1}^{\infty} j P_p(j+\ell)$$

$$\leq pM_K + \varepsilon p \sum_{j=K+1}^{\infty} (j+\ell) P_p(j+\ell) \leq pM_K + \varepsilon p\, \mathbb{E}P_p \leq pM_K + C\varepsilon.$$

Clearly, $pM_K \leq \varepsilon$ for all sufficiently small p. Therefore, for such p, we have

$$p \sum_{j=0}^{\infty} c(j) P_p(j+\ell) \leq (C+1)\varepsilon.$$

□

The second lemma concerns the asymptotic behavior of the Kullback–Leibler distance $d_{KL}(P, Q)$ between two TNBDs.

Lemma 2. *Let* $\ell \in \mathbb{Z}_+$, $P = NB(a, p) + \alpha$ *and* $Q = NB(b, q) + \beta$ *with* $a + \alpha = b + \beta = \ell$. *Suppose that* $p,\ q \to 0$ *in such a way that* $p \log q \to 0$. *Then,* $p\, d_{KL}(P, Q) \to 0$.

Proof. We have

$$d_{KL}(P, Q) = \sum_{n=\ell}^{\infty} P(n) \log \frac{P(n)}{Q(n)} = \sum_{j=0}^{\infty} P(j+\ell) \log \frac{P(j+\ell)}{Q(j+\ell)}. \tag{13}$$

In view of Formulas (9),

$$\frac{P(j+\ell)}{Q(j+\ell)} = \frac{\binom{j+a-1}{a-1} p^a (1-p)^j}{\binom{j+b-1}{b-1} q^b (1-q)^j}$$

independently of ℓ, so that

$$\log \frac{P(j+\ell)}{Q(j+\ell)} = a \log p - b \log q + j \log \frac{1-p}{1-q} + c(j), \tag{14}$$

where the sequence

$$c(j) = \log \frac{\binom{j+a-1}{a-1}}{\binom{j+b-1}{b-1}}, \quad j \geq 0, \tag{15}$$

clearly has property (11). From (13) and (14), we conclude that

$$d_{KL}(P,Q) = a \log p - b \log q + \log \frac{1-p}{1-q} \sum_{j=0}^{\infty} jP(j+\ell) + \sum_{j=0}^{\infty} c(j)P(j+\ell).$$

Recall that the expected value of the distribution $NB(r,p)$ is r/p. Together with (10), this implies

$$\sum_{j=0}^{\infty} jP(j+\ell) = \sum_{j=0}^{\infty} (j+\ell)P(j+\ell) - \ell = \frac{a}{p} + \alpha - \ell = a\left(\frac{1}{p} - 1\right). \tag{16}$$

Therefore,

$$d_{KL}(P,Q) = a \log p - b \log q + a\left(\frac{1}{p} - 1\right) \log \frac{1-p}{1-q} + \sum_{j=0}^{\infty} c(j)P(j+\ell). \tag{17}$$

We now apply Lemma 1 to the family of TNBDs $P_p = NB(a,p) + \alpha$, $0 < p < 1$, and the sequence $\{c(j)\}_{j=0}^{\infty}$ given by (15). Note that the assumption of Lemma 1 regarding distributions P_p is met because $p\, \mathbb{E}P_p = a + \alpha p \leq a + \alpha = \ell$ for all p. Therefore, using (12), we infer from (17) that if

$$p,\ q \to 0 \quad \text{in such a way that} \quad p \log q \to 0, \tag{18}$$

then $p\, d_{KL}(P,Q) \to 0$, which completes the proof of Lemma 2. □

Remark 1. *A similar proof would show that if $p,\ q \to 0$ in such a way that $q \log p \to 0$, then $q\, d_{KL}(Q,P) \to 0$.*

Proof. We now proceed to proving part (i) of Proposition 1 assuming that conditions (18) are met. Recall that the expected value μ_P and variance σ_P^2 of the TNBD (9a) are given by $\mu_P = a/p + \alpha$ and $\sigma_P^2 = a(1-p)/p^2$. In view of (3) and (7), we find that

$$\mathbb{E}_A X = d_{KL}(P,Q) + \mu_{\mathcal{P}}\left(\frac{a}{p} + \alpha\right). \tag{19}$$

Then, according to Lemma 2, we have $p\, \mathbb{E}_A X \to a\mu_{\mathcal{P}}$.

We turn to the asymptotic behavior of the variance of the log-likelihood score X under model A. As a reminder, $X = f(N) + U$, where $f(N) = \log[P(N)/Q(N)]$ and $U = X_1 + X_2 + \ldots + X_N$ is a random sum, see (4). Then, for sequences generated by model A, we have

$$Var_A(X) = Var_P[f(N)] + Var_{\mathcal{P}}(U) + 2Cov_A[f(N), U]. \tag{20}$$

We begin with the term $Var_P[f(N)]$:

$$Var_P[f(N)] = \mathbb{E}_P f^2(N) - [\mathbb{E}_P f(N)]^2 \leq \mathbb{E}_P f^2(N). \tag{21}$$

Using the inequality $(x+y+z)^2 \leq 3(x^2+y^2+z^2)$, we find on account of (14) that

$$\mathbb{E}_P f^2(N) = \sum_{j=0}^{\infty} P(j+\ell) \log^2 \frac{P(j+\ell)}{Q(j+\ell)} \leq 3\,(a\log p - b\log q)^2$$

$$+3\log^2 \frac{1-p}{1-q} \sum_{j=0}^{\infty} j^2 P(j+\ell) + 3\sum_{j=0}^{\infty} c^2(j) P(j+\ell). \tag{22}$$

For the first term after the inequality sign in (22), we have, under the conditions in (18), $p^2(a\log p - b\log q)^2 \to 0$. Regarding the second term, we first estimate the second moment,

$$M_2(P) = \sum_{j=0}^{\infty} (j+\ell)^2 P(j+\ell),$$

of the TNBD $P = NB(a,p) + \alpha$ as follows:

$$M_2(P) = \frac{a(1-p)}{p^2} + \left(\frac{a}{p} + \alpha\right)^2 \leq \frac{a + (a+\alpha)^2}{p^2} = \frac{a+\ell^2}{p^2}.$$

Therefore,

$$p^2 \log^2 \frac{1-p}{1-q} \sum_{j=0}^{\infty} j^2 P(j+\ell) \leq p^2 \log^2 \frac{1-p}{1-q} \sum_{j=0}^{\infty} (j+\ell)^2 P(j+\ell)$$

$$\leq (a+\ell^2) \log^2 \frac{1-p}{1-q} \to 0 \quad \text{as} \quad p,\, q \to 0.$$

Next, since $c^2(j)/j \to 0$ as $j \to \infty$, we conclude from Lemma 1 that

$$p \sum_{j=0}^{\infty} c^2(j) P(j+\ell) \to 0.$$

Combining the above limit relations for the three terms in Formula (22), we obtain $p^2 \mathbb{E}_P f^2(N) \to 0$, which, in view of (21), implies

$$p^2 Var_P[f(N)] \to 0. \tag{23}$$

We now focus on the second term in (20). According to the formula for the variance of a random sum [3],

$$Var_{\mathcal{P}}(U) = \sigma_{\mathcal{P}}^2\, \mathbb{E}_P N + \mu_{\mathcal{P}}^2\, Var_P(N) = \left(\frac{a}{p} + \alpha\right)\sigma_{\mathcal{P}}^2 + \frac{a(1-p)}{p^2}\mu_{\mathcal{P}}^2,$$

then

$$p^2 Var_{\mathcal{P}}(U) \to a\mu_{\mathcal{P}}^2 \quad \text{as} \quad p \to 0. \tag{24}$$

Finally, for the third term in (20), we obtain by the Cauchy–Schwarz inequality

$$p^2 \mid Cov_A[f(N), U] \mid \leq \sqrt{p^2 Var_P[f(N)]}\sqrt{p^2 Var_{\mathcal{P}}(U)}.$$

It follows from (23) and (24) that $p^2 Cov_A[f(N), U] \to 0$.

In summary, Formula (20) and the limit relations for its terms yield $p^2 Var_A X \to a\mu_{\mathcal{P}}^2$.

To prove part (ii) of Proposition 1, notice that, in the case of model B, Formula (19) takes on the form

$$\mathbb{E}_B X = \mu_{\mathcal{Q}}\left(\frac{b}{q} + \beta\right) - d_{KL}(Q,P),$$

which implies that if $p,\ q \to 0$ and $q \log p \to 0$, then $q\, \mathbb{E}_B X \to b\mu_Q$. The proof of the limit relation for the variance of rv X under model B is identical to that for model A. □

Remark 2. *Asymptotic formulas for the expected value of rv X in Proposition 1 hold for any sequence of iid rvs (X_n) with finite expectation and, similarly, asymptotic formulas for the variance of rv X are valid for any such sequence of rvs with a finite second moment.*

3. Proof of Theorem 1

Because models A and B can be treated similarly, we only prove Theorem 1 for model A. As a preliminary, we compute the characteristic function (ch. f.) of rvs X and $Y = X/\mathbb{E}_A X$. To compute the ch. f. of rv X, denote by $\varphi_{\mathcal{P}}$ the ch. f. of rvs X_n for sequences generated by model A. Conditioning on rv N and using its independence of rvs $X_1, X_2, \ldots$, we find that

$$\Phi_X(s) = \mathbb{E}_A e^{isX} = \mathbb{E}_A e^{is[f(N)+U]} = \sum_{n=\ell}^{\infty} P(n) e^{is \log[P(n)/Q(n)]}\, \mathbb{E}_{\mathcal{P}} e^{is(X_1+X_2+\ldots+X_n)}$$

$$= \sum_{n=\ell}^{\infty} P(n) e^{is \log[P(n)/Q(n)]} \varphi_{\mathcal{P}}^n(s). \tag{25}$$

Also, for the ch. f. of rv Y, we have

$$\Phi_Y(t) = \mathbb{E}_A e^{itY} = \Phi_X(t/\mathbb{E}_A X). \tag{26}$$

The following result shows that the presence of the exponential factor $e^{is \log[P(n)/Q(n)]}$ in (25) does not affect the asymptotic behavior of $\Phi_X(t/\mathbb{E}_A X)$.

Lemma 3. *Under the conditions in (18),*

$$\Phi_X(t/\mathbb{E}_A X) - \sum_{n=\ell}^{\infty} P(n) \varphi_{\mathcal{P}}^n(t/\mathbb{E}_A X) \to 0\,, \quad t \in \mathbb{R}.$$

Proof. Recall that $\mid \varphi_{\mathcal{P}}(s) \mid \leq 1$ for all $s \in \mathbb{R}$. Using the inequality $\mid e^{ix} - 1 \mid \leq \mid x \mid$, $x \in \mathbb{R}$, we obtain in view of (25)

$$\mid \Phi_X(s) - \sum_{n=\ell}^{\infty} P(n) \varphi_{\mathcal{P}}^n(s) \mid = \mid \sum_{n=\ell}^{\infty} P(n) \varphi_{\mathcal{P}}^n(s) \Big(e^{is \log[P(n)/Q(n)]} - 1 \Big) \mid$$

$$\leq \sum_{n=\ell}^{\infty} P(n) \mid e^{is \log[P(n)/Q(n)]} - 1 \mid \leq \mid s \mid \sum_{n=\ell}^{\infty} P(n) \mid \log \frac{P(n)}{Q(n)} \mid .$$

We set here $s = t/\mathbb{E}_A X$ and invoke (14) and (16) to find that

$$\mid \Phi_X(t/\mathbb{E}_A X) - \sum_{n=\ell}^{\infty} P(n) \varphi_{\mathcal{P}}^n(t/\mathbb{E}_A X) \mid$$

$$\leq \frac{\mid t \mid}{p\, \mathbb{E}_A X} \Big[ap \mid \log p \mid + bp \mid \log q \mid + a(1-p) \mid \log \frac{1-p}{1-q} \mid + S_p \Big], \tag{27}$$

where

$$S_p = p \sum_{j=0}^{\infty} \mid c(j) \mid P(j+\ell)$$

and sequence $\{c(j)\}$ is defined by (15). By Lemma 1, we have $S_p \to 0$ as $p \to 0$. Also, in view of Proposition 1,under the conditions in (18),

$$p\, \mathbb{E}_A X \to a\mu_{\mathcal{P}} > 0. \tag{28}$$

The conclusion of Lemma 3 now follows immediately from (27). □

Next, we prove Theorem 1 starting with the limiting distribution of rv Y. To identify the latter, we have to find the limit of the ch. f. of rv Y given by (26). According to Lemma 3, we only have to compute the limit of the function

$$\Omega(s) = \sum_{n=\ell}^{\infty} P(n)\varphi_{\mathcal{P}}^{n}(s) = \Psi_P[\varphi_{\mathcal{P}}(s)]$$

evaluated at $s = t/\mathbb{E}_A X$, where

$$\Psi_P(z) = \sum_{n=\ell}^{\infty} P(n)z^n = z^{\alpha}\left[\frac{pz}{1-(1-p)z}\right]^{a}, \quad |\, z\,| < \frac{1}{1-p}\,,$$

is the probability generating function of the TNBD $P = N(a,p) + \alpha$.

Since the distribution of rvs X_n under model $\mathcal{P}$ has a finite first moment, we can use the first-order Taylor expansion of its ch. f.:

$$\varphi_{\mathcal{P}}(s) = 1 + \varphi_{\mathcal{P}}^{'}(0)s + s\rho(s) = 1 + i\mu_{\mathcal{P}}s + s\rho(s),$$

where $\rho(s) \to 0$ as $s \to 0$. Then,

$$\begin{aligned}\Omega(s) &= [\varphi_{\mathcal{P}}(s)]^{\alpha}\left[\frac{p\varphi_{\mathcal{P}}(s)}{1-(1-p)\varphi_{\mathcal{P}}(s)}\right]^{a}\\ &= [1 + i\mu_{\mathcal{P}}s + s\rho(s)]^{\alpha}\left[\frac{p(1 + i\mu_{\mathcal{P}}s + s\rho(s))}{1-(1-p)(1 + i\mu_{\mathcal{P}}s + s\rho(s))}\right]^{a}\\ &= [1 + i\mu_{\mathcal{P}}s + s\rho(s)]^{\alpha}\left[\frac{1 + i\mu_{\mathcal{P}}s + s\rho(s)}{1-(1-p)(i\mu_{\mathcal{P}} + \rho(s))s/p}\right]^{a}.\end{aligned}$$

Setting here $s = t/\mathbb{E}_A X$, we find, due to (28), that under the conditions in (18), s/p has a finite limit $t/(a\mu_{\mathcal{P}})$, which implies that $s \to 0$. Therefore, we conclude from (29) that

$$\Omega(t/\mathbb{E}_A X) \to \left(1 - \frac{it}{a}\right)^{-a}.$$

Thus, by Lemma 3, we also have

$$\Phi_Y(t) \to \left(1 - \frac{it}{a}\right)^{-a}. \tag{29}$$

This limiting function represents the ch. f. of the Erlang distribution $E(a,a)$.

To find the limiting distribution of rv Z under model A, notice that in view of (8)

$$Z = k(X)(Y-1), \tag{30}$$

where $k(X) = \mathbb{E}_A X/\sigma_A(X)$. By Proposition 1, under the conditions in (18)

$$k(X) \to a\mu_{\mathcal{P}}/(\sqrt{a}\mu_{\mathcal{P}}) = \sqrt{a}. \tag{31}$$

Therefore, from (30)–(32),

$$\Phi_Z(t) = e^{-ik(X)t}\Phi_Y[k(X)t] \to e^{-i\sqrt{a}\,t}\left(1 - \frac{it}{\sqrt{a}}\right)^{-a}.$$

Thus, the limiting distribution of rv Z under model A is the Erlang distribution $E(a,\sqrt{a})$ translated by $\sqrt{a}$ to the left, or symbolically $E(a,\sqrt{a}) - \sqrt{a}$.

The limiting ch. f. for rv Y under model B can be computed along similar lines. In this case, one has to take into account that $\mu_{\mathcal{Q}} < 0$, which brings about a change of the sign in the analogs of formulas (28) and (32). As a result, under conditions $p,\ q \to 0$ and $q \log p \to 0$,

$$\Phi_Y(t) \to \left(1 - \frac{it}{b}\right)^{-b} \quad \text{and} \quad \Phi_Z(t) \to e^{i\sqrt{b}\,t}\left(1 + \frac{it}{\sqrt{b}}\right)^{-b}.$$

Therefore, the limiting distributions of rvs Y and Z are, respectively, the Erlang distribution $E(b, b)$ and the reflected Erlang distribution $E(b, \sqrt{b})$, translated by $\sqrt{b}$ to the right, or symbolically $\sqrt{b} - E(b, \sqrt{b})$.

Remark 3. *Theorem 1 holds for any sequence (X_n) of iid rvs with a finite second moment that has a positive expected value under model A and a negative expected value under model B.*

Remark 4. *It follows from (8) that, under model A, the distribution of rv X can be approximated by either an Erlang distribution $E(a, \lambda)$ with $\lambda = a\mathbb{E}_A X$ or by a transformed Erlang distribution $E(a, \gamma) + \tau$ with $\gamma = \sqrt{a}\,\sigma_A(X)$ and $\tau = \mathbb{E}_A X - \sqrt{a}\,\sigma_A(X)$. A similar remark also holds for model B.*

4. The Accuracy of the Likelihood-Based Classification of Random Sequences

Among the many measures of classification accuracy, perhaps the most informative ones are misclassification error rates $P_A(X \le 0)$ and $P_B(X \ge 0)$. The first of them represents the probability that a sequence generated by model A is not assigned to this model by the classification decision rule, i.e., it is either assigned to model B or not assigned to any model. A similar interpretation holds for the other error rate. An important question is whether, for very long sequences, the classification produces the correct result with a probability approaching 1; equivalently, this means that both misclassification error rates approach 0. We will call such a classification *asymptotically accurate*. The following statement about the asymptotic accuracy of the likelihood-based classification of random sequences described in the Introduction follows from Theorem 1.

Theorem 2. *Suppose that character selection probabilities $\mathcal{P}$ and $\mathcal{Q}$ for models A and B are distinct and that the sequence length distributions under these models are $P = NB(a, p) + \alpha$ and $Q = NB(b, q) + \beta$, respectively, with $a + \alpha = b + \beta = \ell$. If $p,\ q \to 0$ in such a way that $p \log q \to 0$ and $q \log p \to 0$, then the likelihood-based classification of such random sequences is asymptotically accurate.*

Proof. It follows from (28) that if p, q and $p \log q$ are all sufficiently small, then $\mathbb{E}_A X > 0$. Therefore, in view of Theorem 1 and due to the fact that the limiting distribution $E(a, a)$ does not have an atom at 0, we obtain

$$P_A(X \le 0) = P_A\left(\frac{X}{\mathbb{E}_A X} \le 0\right) \to P_A(V \le 0) = 0,$$

where V is a rv with Erlang distribution $E(a, a)$. Similarly, if p, q and $q \log p$ are all sufficiently small, then $\mathbb{E}_B X < 0$. Using Theorem 1, we conclude that

$$P_B(X \ge 0) = P_B\left(\frac{X}{\mathbb{E}_B X} \le 0\right) \to P_B(W \le 0) = 0,$$

where W is a rv with Erlang distribution $E(b, b)$. □

Recall that the meaning of the assumptions of Theorem 2 related to parameters p and q is that the expected length of long sequences generated by either model cannot be exponentially larger than that for sequences generated by the other model.

5. An Application to Genomics: Classification of DNA Sequences as Protein-Coding or Noncoding

In this section, we apply Theorem 1 to the classification of DNA sequences of bacterium *Bacillus subtilis* strain 168 as protein-coding or noncoding. *Bacillus subtilis* is a model bacterial organism with a well-annotated genome [15], which was extracted from the open source National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/nuccore/AF012532.1) accessed on 10 October 2023.

The annotated list of *Bacillus subtilis* genes is found at https://www.ncbi.nlm.nih.gov/genome/browse/#!/proteins/665/300274%7CBacillus%20subtilis%20subsp.%20subtilis%20str.%20168/chromosome/, accessed on 21 August 2023.

5.1. Background

Recall that (a) genetic information stored in the DNA can be represented as a sequence of nucleotides, A, C, T, G (adenine, cytosine, guanine, and thymine, respectively); (b) a protein is a sequence of amino acids; (c) each amino acid is encoded by one or several (up to six) triplets of DNA nucleotides, called *codons*; (d) a protein-coding gene is a sequence of codons encoding a protein; (e) the first codon of a gene is a START codon (typically ATG, encoding the amino acid methionine) signaling the start of transcription; (f) every gene is followed by a STOP triplet (TAA, TAG, or TGA) that does not encode an amino acid and signals the termination of the transcription process; (g) each gene belongs to one of the two complementary strands of the DNA; (h) genes of many prokaryotic organisms including all bacteria do not contain noncoding DNA segments, called *introns*. Thus, bacterial genes are contiguous sequences of codons starting with a START codon, followed by one of the three STOP triplets and not containing other in-frame STOP triplets. DNA sequences with these properties are called *open reading frames* (ORFs). DNA of various organisms, including *Bacillus subtilis*, contain numerous ORFs other than protein-coding genes. For more information about DNA, codons, genes, ORFs, amino acids, and proteins, see [16].

In what follows, we compare the Erlang distributions identified in Theorem 1 with the empirical distributions of the normalized log-likelihood scores Y and Z under models A and B associated with two respective classes of DNA sequences extracted from the *Bacillus subtilis* genome: protein-coding genes and a certain natural class, defined below, of protein noncoding ORFs. Parameters of these models of DNA sequence assembly were estimated based on the known membership of *Bacillus subtilis* ORFs in the two classes. A similar comparison can be performed for any other well-annotated prokaryotic genome without introns.

5.2. Protein-Coding Genes and Noncoding Open Reading Frames: Data and Models

The genome of *Bacillus subtilis* was found to contain no repeated genes or those with in-frame internal STOP triplets. A peculiar feature of the *Bacillus subtilis* genome is that only about 77.5% of its 4237 protein-coding genes begin with the standard START codon ATG. The vast majority of the remaining protein-coding genes begin with alternative START codons: TTG (coding for amino acid leucine) or GTG (coding for valine), which occur in 13% and 9% of all protein-coding genes, respectively. Additionally, 15 *Bacillus subtilis* protein-coding genes have nonstandard START codons: CTG encoding leucine and ATT encoding isoleucine.

To identify all protein noncoding ORFs within the *Bacillus subtilis* genome, we deleted all the protein-coding genes from the genome and read the resulting contiguous segments of nucleotides in the $5^{'} \to 3^{'}$ direction on the strand to which they belong. If the number, n, of nucleotides in any such segment was divisible by three then the segment was read in its natural frame; if n was of the form $n = 3k + 1$, then the segment was read in two reading frames (i.e., starting with the first or second nucleotide), while in the case $n = 3k + 2$, it was read in three reading frames (i.e., starting with the first, second, or third nucleotide). From all these reads, sequences of triplets beginning with the main START codon ATG

utilized by *Bacillus subtilis*, followed by one of the STOP triplets, and not containing other in-frame STOP triplets were selected. This resulted in 4571 ORFs beginning with the standard START codon ATG. We will call them *standard noncoding ORFs.*

Note that some of them may actually represent genes encoding various kinds of RNA.

The following idea, borrowed from [10], allows one to view protein-coding genes and standard noncoding ORFs as randomly and independently assembled sequences of triplets of the kind discussed in the Introduction. Recall that any DNA sequence from each of these two classes is followed by a STOP triplet. Proceeding from any such STOP triplet, we move backwards adding new nucleotide triplets other than STOP triplets randomly and independently of each other. The alphabet used for such sequence assembly thus contains $M = 4^3 - 3 = 61$ triplets. The character selection probabilities for protein-coding genes and standard noncoding ORFs can be defined on empirical grounds as the respective frequencies of the 61 triplets found in all 4237 protein-coding genes and all 4571 standard noncoding ORFs, see Table 1. Also note that, under our independent model of DNA sequence assembly, the empirical frequency of a triplet coincides with the maximum likelihood estimate of the class-specific selection probability for the corresponding character given the data [8,17]. Based on the frequencies reported in Table 1, we found that $\mu_{\mathcal{P}} = 0.0709$, $\sigma_{\mathcal{P}} = 0.3575$ and $\mu_{\mathcal{Q}} = -0.0791$, $\sigma_{\mathcal{Q}} = 0.4182$.

Table 1. Observed frequencies of 61 triplets or codons for the two classes of DNA sequences of the *Bacillus subtilis* genome: *A* (protein-coding genes) and *B* (standard noncoding ORFs). Triplets are ordered lexicographically.

Triplets	*A*	*B*	Triplets	*A*	*B*
AAA	0.0496	0.0391	CTT	0.0232	0.0207
AAC	0.0172	0.0158	GAA	0.0493	0.0256
AAG	0.0211	0.0221	GAC	0.0186	0.0121
AAT	0.0223	0.0233	GAG	0.0232	0.0137
ACA	0.0223	0.0185	GAT	0.0332	0.0206
ACC	0.0086	0.0105	GCA	0.0217	0.0168
ACG	0.0145	0.0124	GCC	0.0159	0.0149
ACT	0.0087	0.0102	GCG	0.0202	0.0139
AGA	0.0108	0.0146	GCT	0.0190	0.0178
AGC	0.0142	0.0185	GGA	0.0218	0.0142
AGG	0.0038	0.0113	GGC	0.0235	0.0142
AGT	0.0066	0.0092	GGG	0.0112	0.0091
ATA	0.0094	0.0197	GGT	0.0127	0.0096
ATC	0.0271	0.0228	GTA	0.0134	0.0112
ATG	0.0271	0.0413	GTC	0.0174	0.0131
ATT	0.0372	0.0263	GTG	0.0178	0.0118
CAA	0.0197	0.0179	GTT	0.0193	0.0163
CAC	0.0074	0.0094	TAC	0.0121	0.0104
CAG	0.0187	0.0171	TAT	0.0228	0.0185
CAT	0.0153	0.0172	TCA	0.0148	0.0219
CCA	0.0070	0.0111	TCC	0.0080	0.0149
CCC	0.0033	0.0096	TCG	0.0063	0.0119
CCG	0.0159	0.0161	TCT	0.0129	0.0174
CCT	0.0105	0.0133	TGC	0.0043	0.0143
CGA	0.0040	0.0096	TGG	0.0103	0.0113
CGC	0.0085	0.0111	TGT	0.0036	0.0130
CGG	0.0064	0.0126	TTA	0.0192	0.0177
CGT	0.0074	0.0097	TTC	0.0142	0.0248
CTA	0.0049	0.0065	TTG	0.0155	0.0211
CTC	0.0109	0.0130	TTT	0.0308	0.0382
CTG	0.0233	0.0193			

In what follows, models A and B are used for describing DNA sequences representing protein-coding genes and standard noncoding ORFs, respectively. To specify these models completely, we need to determine their length distributions. Because the first triplet of any ORF is a START codon, random sequences anchored by a given STOP triplet and assembled as described above form a cluster of nested ORFs, each determined by the number, r, of START codons preceding the STOP triplet, as illustrated in Figure 1. Empirically, we found that the number r displays substantial variation, see Figure 2, showing the histogram for

the values of r for all protein-coding genes beginning with the standard START codon ATG found in the *Bacillus subtilis* genome. According to our model of DNA sequence assembly, the length of protein-coding genes and standard noncoding ORFs with a *fixed* number $r \geq 1$ of START triplets ATG would follow respective negative binomial distributions $NB(r, p)$ and $NB(r, q)$, where $p = 0.0271$ and $q = 0.0413$ are the empirical frequencies of the START codon ATG for the two respective classes of ORFs, see Table 1. Therefore, the length distribution for protein-coding genes or standard noncoding ORFs is a mixture of such negative binomial distributions over all observed values of r, whose relative weights can also be determined empirically (for protein-coding genes with the START codon ATG, the absolute weights are given in Figure 2). Additionally, to encode functional proteins, genes have to be sufficiently long. In fact, the shortest protein-coding gene in the *Bacillus subtilis* genome has 20 codons. By comparison, the shortest standard noncoding ORF identified in this genome is 25 triplets long.

To account for such complexity of the length distribution, we assumed it to be TNBD $NB(a, p) + \alpha$ for model A and $NB(b, q) + \beta$ for model B, where $a + \alpha = b + \beta = \ell = 20$ triplets, with adjustable parameters a, b, p, q, to be estimated from the data.

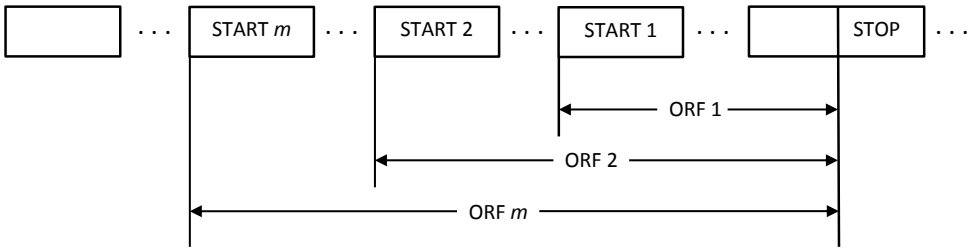

Figure 1. A nested cluster of ORFs anchored by a given STOP triplet. Empty boxes represent triplets of nucleotides other than START codons or STOP triplets. Reproduced with permission from [10].

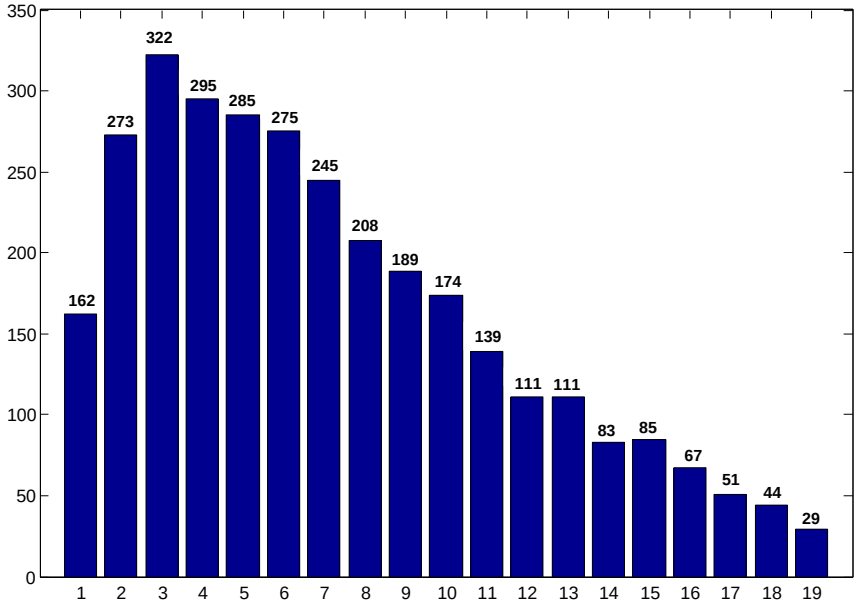

Figure 2. Histogram of the number of START codons ATG preceding a gene's STOP triplet for protein-coding *Bacillus subtilis* genes beginning with the START codon ATG.

5.3. Results

Parameters a, b, p, q of the two TNBDs were estimated by minimizing the total variation distance, d, between the assumed TNBDs and the empirical length distributions for protein-coding genes (model A) and standard noncoding ORFs (model B). The resulting optimal values were $a = 2,\ p = 0.0077$ for model A and $b = 1,\ q = 0.0467$ for model B, while the

respective minimal total variation distances were found to be $d_A = 0.3945$ and $d_B = 0.4707$. Then, the translation parameters are $\alpha = \ell - a = 18$ codons for protein-coding genes and $\beta = \ell - b = 19$ triplets for standard noncoding ORFs. Thus, the best-fitting theoretical length distribution for standard noncoding ORFs is a translated geometric distribution $G(q) + \beta$. The relatively large magnitude of the minimum total variation distance is due to the fact that many lengths of protein-coding genes and standard noncoding ORFs carrying positive probabilities in the theoretical distributions are absent in the genome of *Bacillus subtilis*; yet another reason is the presence in this genome of a large number of anomalously long (in relative terms) sequences of both classes.

The profiles of the total variation distance as functions of parameters p and q for the optimal values $a = 2$ and $b = 1$ are shown in Figure 3. Notice that (i) if $p \to 0$ or $q \to 0$, then the corresponding theoretical length distributions "escape to infinity" so that $d \to 2$; (ii) if $p \to 1$, then $d \to 2[1 - P(20)]$, where $P(20) = 1/4237$ is the frequency of the minimum gene length of 20 codons; and (iii) by contrast, if $q \to 1$, then $d \to 2$ due to the fact that the shortest length of standard noncoding ORFs is 25 triplets rather than 20 triplets. The limiting behaviors (i)–(iii) are clearly seen in Figure 3. Finally, the estimated TNBDs and empirical length distributions approximated by suitable histograms are displayed in Figure 4A,B. We conclude from Figure 4 that TNBDs with the above-specified parameters provide an excellent visual fit to the empirical length distributions for the two classes of DNA sequences.

For the expected model-based lengths of protein-coding genes and standard noncoding ORFs, measured in triplets, we have

$$\mu_P = \frac{a}{p} + \alpha \simeq 278 \quad \text{and} \quad \mu_Q = \frac{b}{q} + \beta \simeq 40$$

while the corresponding standard deviations, also measured in triplets, are

$$\sigma_P = \frac{\sqrt{a(1-p)}}{p} \simeq 183 \quad \text{and} \quad \sigma_Q = \frac{\sqrt{b(1-q)}}{q} \simeq 21,$$

to be compared with their empirical counterparts $\bar{N}_A \simeq 290$, $\bar{N}_B \simeq 54$ and $s_A(N) \simeq 266$, $s_B(N) \simeq 91$. A few comments about the length distributions for the two classes of DNA sequences are in order:

(i) The genome of *Bacillus subtilis* contains a large number of very short standard noncoding ORFs. For example, the number of such ORFs with the length of 25 triplets (the shortest possible) is 273, while the number of those with the length ranging from 25 to 30 triplets is 1331 or 29%;

(ii) On average, protein-coding genes are much longer than standard noncoding ORFs. In fact, the ratio of their observed average lengths is about 5.4 and that of their model-based expected lengths is about 7.0;

(iii) The genome contains a significant number of very long protein-coding genes. The seven longest among them have lengths 3583, 3587, 3603, 4262, 4538, 5043, and 5488 codons, while the eighth longest gene is just 2561 codons long. This explains why the empirical standard deviation of gene length, $s_A(N) \simeq 266$ codons, is substantially larger than its theoretical counterpart, $\sigma_P \simeq 183$ codons. Without the seven longest genes, one would have $s_A(N) \simeq 208$ codons;

(iv) Although the number of anomalously long standard noncoding ORFs is disproportionately smaller than the number of very long protein-coding genes, their effect on the standard deviation of the length distribution is still considerable. For example, the longest standard noncoding ORF has 4428 triplets, while the length of the second longest ORF is 1190 triplets. Removing the longest ORF would reduce the standard deviation of ORF length from $s_B(N) \simeq 91$ to 64 triplets.

We also fitted TNBDs to the empirical length distribution for 3283 protein-coding genes beginning with the standard START codon ATG. This resulted in $a = 2$ (implying

that $\alpha = 18$) and $p = 0.0078$, while the minimum total variation distance was 0.4313. Thus, the best-fitting TNBD is virtually indistinguishable from the same for the entire set of 4237 *Bacillus subtilis* protein-coding genes; however, surprisingly, the goodness of fit for the entire collection of genes is even better than for the seemingly more homogeneous subset of genes with the standard START codon ATG. That is why we used the entire set of protein-coding genes in our analysis.

Once models A and B are completely specified, one can evaluate the log-likelihood score X given by Formula (3) for each DNA sequence from either class. For the above-specified models of sequence length, the first term in (3) for any given sequence of length $N = n$ is

$$\log \frac{P(n)}{Q(n)} = \log \frac{p^2}{q} + \log(n-19) + (n-20) \log \frac{1-p}{1-q},$$

where $p = 0.0077$ and $q = 0.0467$. We then computed the class-specific normalized scores

$$Y_A = \frac{X}{\bar{X}_A}, \quad Y_B = \frac{X}{\bar{X}_B} \quad \text{and} \quad Z_A = \frac{X - \bar{X}_A}{s_A(X)}, \quad Z_B = \frac{X - \bar{X}_B}{s_B(X)},$$

where $\bar{X}_A \simeq 59.50$ and $\bar{X}_B \simeq -2.30$ are sample averages of the log-likelihood score X over all sequences in the two respective classes, while $s_A(X) \simeq 61.50$ and $s_B(X) \simeq 20.59$ are the corresponding sample standard deviations. Because the values of parameters p and q are small and have roughly the same order of magnitude, it would seem reasonable to compare empirical distributions of the samples Y_A, Y_B, Z_A, Z_B with the respective limiting distributions identified in Theorem 1, see Figures 5A,B and 6A,B, where empirical distributions are represented as histograms with appropriately chosen bins. We conclude from Figures 5 and 6 that the plain and transformed Erlang distributions found in Theorem 1 reproduce essential features of empirical distributions of the samples Y_A, Y_B, Z_A, Z_B such as range, shape, and mode fairly well.

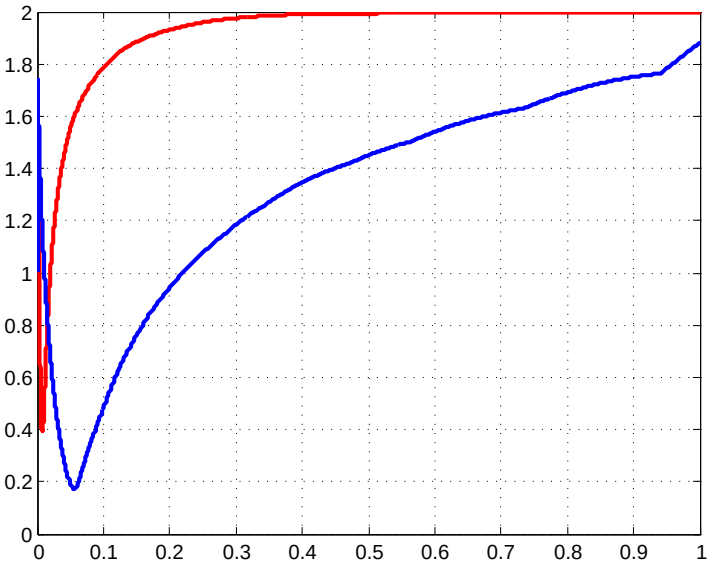

Figure 3. Red curve: plot of the total variation distance, d, between theoretical length distribution $NB(2, p) + 18$ for *Bacillus subtilis* protein-coding genes and its empirical counterpart as a function of parameter p. Blue curve: plot of the distance d between theoretical length distribution $G(q) + 19$ for *Bacillus subtilis* standard noncoding ORFs and its empirical counterpart as a function of parameter q.

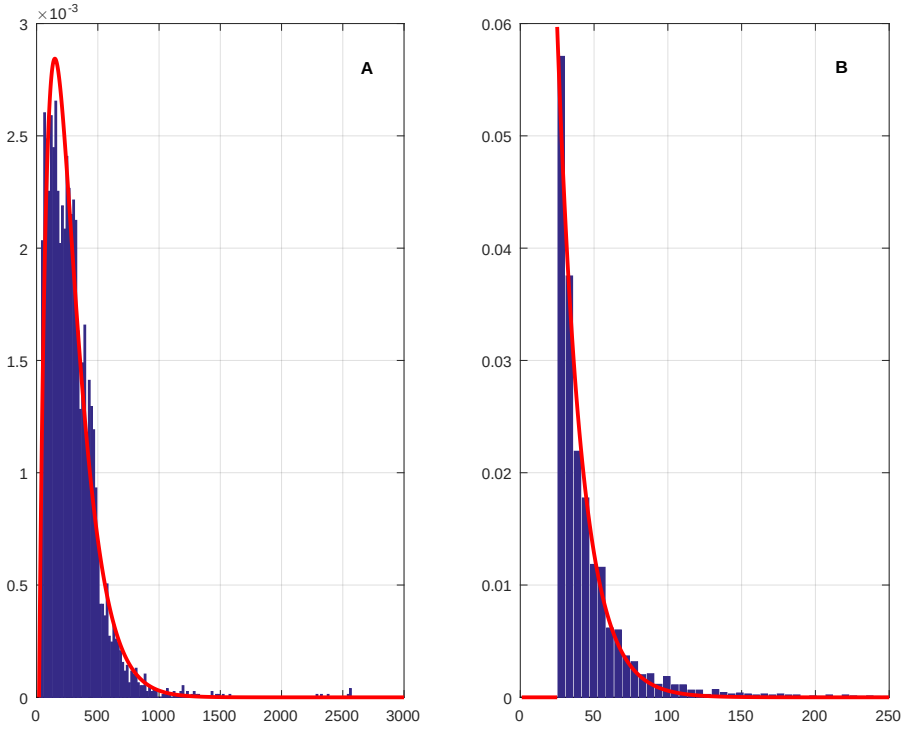

Figure 4. (**A**) Comparison of the empirical length distribution for all *Bacillus subtilis* protein-coding genes with the best-fitting TNBD $NB(2, p) + 18,\ p = 0.0077$. (**B**) Comparison of the empirical length distribution for *Bacillus subtilis* standard noncoding ORFs with the best-fitting translated geometric distribution $G(q) + 19,\ q = 0.0467$.

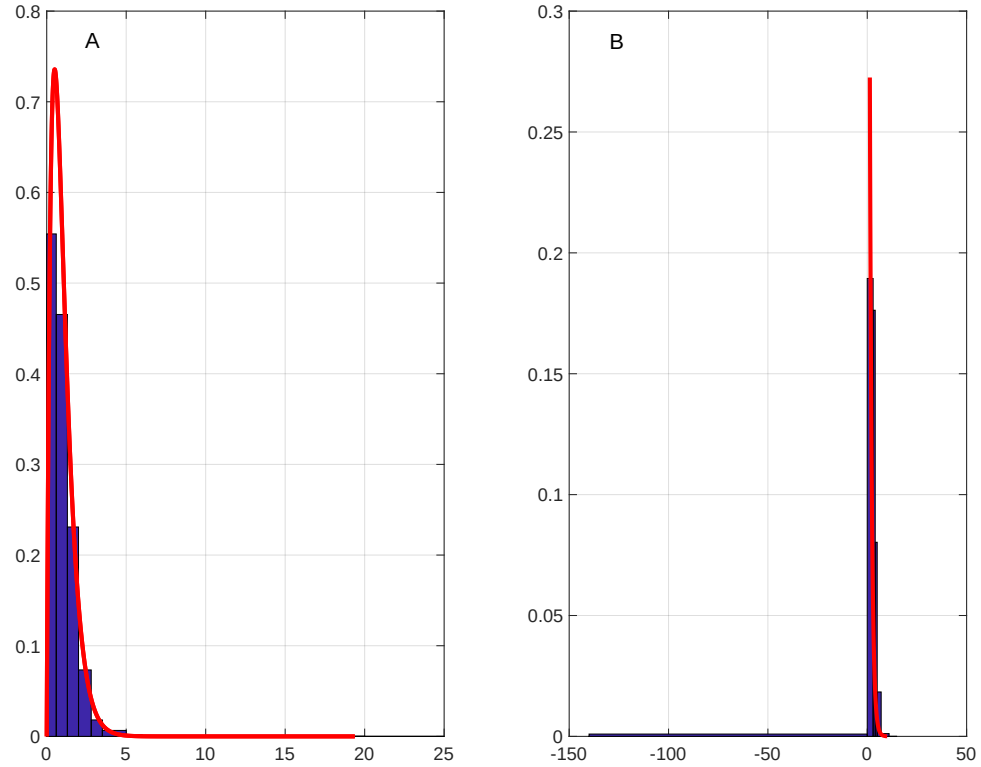

Figure 5. Comparison of the class-specific empirical distribution of the normalized log-likelihood score Y with its theoretical limiting counterpart identified in Theorem 1. (**A**) *Bacillus subtilis* protein-coding genes (class A); (**B**) *Bacillus subtilis* standard noncoding ORFs (class B).

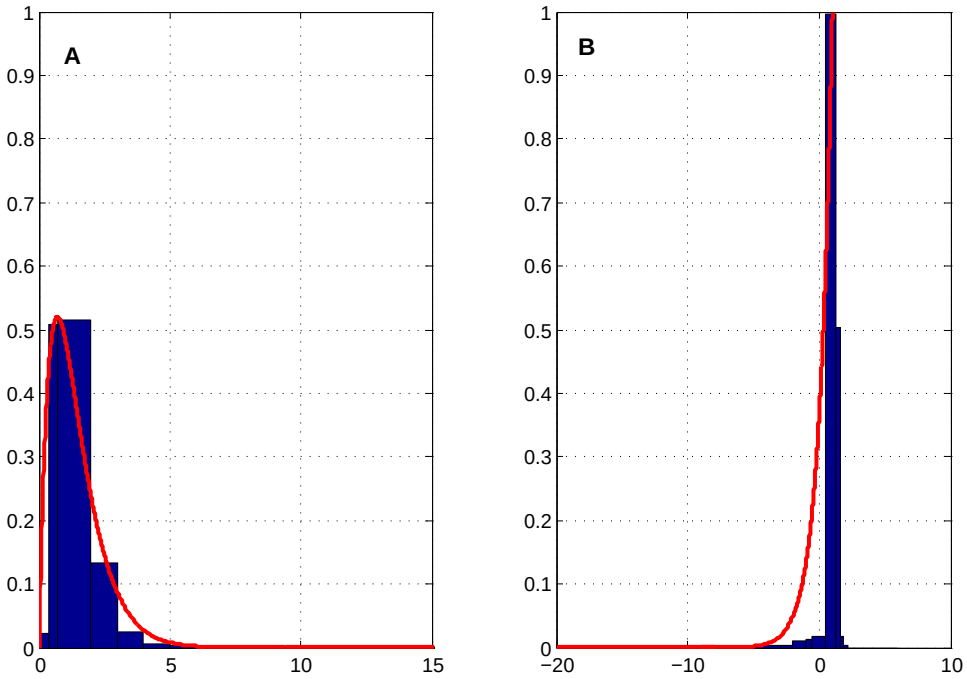

Figure 6. Comparison of the class-specific empirical distribution of the normalized log-likelihood score Z with its theoretical limiting counterpart identified in Theorem 1. (**A**) *Bacillus subtilis* protein-coding genes (class A); (**B**) *Bacillus subtilis* standard noncoding ORFs (class B).

6. Discussion

In this article, we derived a novel limit theorem for two natural normalizations, $Y = X/\mathbb{E}X$ and $Z = (X - \mathbb{E}X)/\sigma(X)$, of the log-likelihood score X, where the expectation and standard deviation are taken relative to either model A or B and it is assumed that the sequence length for these models follows respective TNBDs $NB(a, p) + \alpha$ and $NB(b, q) + \beta$. The limit theorem applies to long sequences ($p, q \to 0$) under the essential additional condition that the expected sequence length for either class is not exponentially larger than for the other class (more precisely, $p \log q \to 0$ and $q \log p \to 0$). The limiting distributions of rv Y under respective models A and B turned out to be Erlang distributions $E(a, a)$ and $E(b, b)$, while for rv Z, they came out as transformed Erlang distributions $E(a, \sqrt{a}) - \sqrt{a}$ and $\sqrt{b} - E(b, \sqrt{b})$, see Theorem 1. It is noteworthy that the limiting distributions depend on integer parameters a and b alone. Thus, the limiting behavior of the normalized log-likelihood score for long sequences represents, under the assumptions of Theorem 1, a fairly crude phenomenon.

Theorem 1 yields an important corollary: the asymptotic accuracy of the likelihood-based classification of random sequences, see Theorem 2.

To test the utility of our limit theorem, we applied it to the classification of open reading frames (ORFs), see Section 4, extracted from the genome of the bacterium *Bacillus subtilis* strain 168, as protein-coding genes (class A) and standard noncoding ORFs (class B). In this case, the alphabet consists of $M = 61$ triplets of DNA nucleotides other than STOP triplets. Since the genome of *Bacillus subtilis* is well annotated, class membership of all ORFs is known with certainty, which allowed us to empirically estimate character selection probabilities and length distributions for both classes of DNA sequences, see Table 1 and Figure 4. As was explained in Section 4, under the model of independent DNA sequence assembly, the length distributions for both classes are mixtures of negative binomial distributions, which we approximated, for each class of sequences, by a single TNBD. The best-fitting distributions from this family provided a surprisingly good fit to the empirical length distributions for both classes of DNA sequences, see Figure 4. This serves as an indirect validation of our model of DNA sequence assembly. This also corroborates earlier findings that the length of protein-coding genes in many organisms

can be approximated by negative binomial distributions [6] or gamma distributions [18], which serve as a continuous analog of negative binomial distributions.

The aforementioned (transformed) Erlang distributions with $a = 2$ and $b = 1$ and their empirical counterparts (i.e., the distributions of the observed normalized log-likelihood scores Y and Z for the two classes of DNA sequences) are compared in Figures 5 and 6. They reveal that the theoretical limiting distributions provide a reasonable fit to the empirical distributions and capture some of their salient features such as range, shape, and mode. This is somewhat unexpected given that (a) the limiting distributions are one-parametric; (b) the model of random independent DNA sequence assembly is quite simplistic; (c) frequencies of the codons immediately following the START codon and immediately preceding the STOP triplet in protein-coding genes are distinct from those for internal codons [6]; and (d) our model disregards various additional features such as the presence in bacterial genomes of short regulatory nucleotide sequences at characteristic distances from the gene's START codon including ribosome binding sites (or Shine–Dalgarno sequences) and binding sites for transcription factors [5,6].

Our results can be applied to the classification of binary sequences ($M = 2$), DNA sequences viewed at the level of individual nucleotides ($M = 4$), and proteins represented as sequences of amino acids ($M = 20$). They may also potentially have applications in the areas of natural language processing and artificial intelligence.

The principal limitation of this work is the use of an independent (or zero-order Markov chain) model of sequence assembly. It was found long ago that DNA sequences are characterized by the presence of substantial short-range [7] and long-range [8] correlations between nucleotides and their triplets. As a result, efficient modern methods of computational gene finding employ higher-order, or even variable-order, Markov chain models and Hidden Markov models at the level of individual nucleotides [4–6]. For example, a gene finder called GeneMark [5] employs a 5th-order Markov chain model, while GLIMMER gene finder [4] combines k-th order Markov chain models for $0 \leq k \leq 8$. Although the accuracy of gene finding generally increases with k (the order of the Markov chain), the use of large values of k is prohibited by the large number, 4^{k+1}, of Markov transition probabilities that have to be estimated from a training set and the sparsity of $(k+1)-$ tuples of nucleotides used for estimation purposes. Thus, to make our limit theorem a better discriminator between protein-coding genes and noncoding ORFs in prokaryotic genomes, it should be extended to higher-order Markov chain models and Hidden Markov models of DNA sequence assembly, and to more general sequence length distributions including translated mixtures of negative binomial distributions.

On the mathematical side, our limit theorem would be more practical if augmented with a tight estimate of the Zolotarev metric [9,14] or another suitable distance [19] between the empirical distribution of the normalized log-likelihood score and its theoretical limiting counterpart.

Author Contributions: Methodology, L.H.; Formal analysis, L.H. and L.P.; Investigation, L.H. and L.P.; Writing—original draft, L.H.; Writing—review and editing, L.H. and L.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We are grateful to Yegor Marin for his help with the extraction and processing of bioinformatics data on *Bacillus subtilis*. We also acknowledge the helpful comments and suggestions by three anonymous reviewers.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hardy, G.H.; Littlewood, J.E.; Polya, G. *Inequalities*, 2nd ed.; Cambridge University Press: Cambridge, UK, 1952.
2. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
3. Ross, S.M. *Introduction to Probability Models*, 6th ed.; Academic Press: San Diego, CA, USA, 1997.
4. Salzberg, S.L.; Delcher, A.L.; Kasif, S.; White, O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **1998**, *26*, 544–548. [CrossRef] [PubMed]
5. Besemer, J.; Lomsadze, A.; Borodovsky, M. GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **2001**, *29*, 2607–2618. [CrossRef]
6. Larsen, T.S.; Krogh, A. EasyGene—A prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinform.* **2003**, *4*, 21. [CrossRef] [PubMed]
7. Almagor, H. A Markov analysis of DNA sequences. *J. Theor. Biol.* **1983**, *104*, 633–645. [CrossRef] [PubMed]
8. Li, W. The study of correlation structures of DNA sequences: A critical review. *Comput. Chem.* **1997**, *21*, 257–271. [CrossRef] [PubMed]
9. Korolev, V. Bounds for the rate of convergence in the generalized Rényi theorem. *Mathematics* **2022**, *10*, 4252. [CrossRef]
10. Hanin, L. A tour of discrete probability guided by a problem in genomics. *Coll. Math. J.* **2020**, *51*, 284–294. [CrossRef]
11. Rényi, A. A characterization of Poisson processes. *Magy. Tud. Akad. Mat. Kut. Int. Kzl.* **1957**, *1*, 519–527.
12. Gnedenko, B.V. Limit theorems for sums of a random number of positive independent random variables. In Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 21 June–18 July 1970 ; pp. 537–549.
13. Korolev, V.Y.; Zeifman, A.I. Bounds for convergence rate in laws of large numbers for mixed Poisson random sums. *Stat. Probab. Lett.* **2021**, *168*, 108918. [CrossRef]
14. Zolotarev, V.M. Properties of and relations among certain types of metrics. *J. Sov. Math.* **1981**, *17*, 2218–2232. [CrossRef]
15. Kunst, F.J.; Ogasawara, N.; Moszer, I.; Albertini, A.M.; Alloni, G.; Azevedo, V.; Bertero, M.G.; Bessières, P.; Bolotin, A.; Borchert, S.; et al. The complete genome sequence of the gram-positive bacterium bacillus subtilis. *Nature* **1997**, *390*, 249–256. [CrossRef] [PubMed]
16. Watson, J.D.; Baker, T.A.; Bell, S.P.; Gann, A.; Levine, M.; Losick, R. *Molecular Biology of the Gene*, 6th ed.; Pearson Education: Cold Spring Harbor, NY, USA, 2008.
17. Ekisheva, S.; Borodovsky, M. Probabilistic models for biological sequences: Selection and Maximum Likelihood estimation. *Int. J. Bioinform. Res. Appl.* **2006**, *2*, 305–324. [CrossRef] [PubMed]
18. Tiessen, A.; Pérez-Rodríguez, P.; Delaya-Oredondo, L.J. Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC Res. Notes* **2012**, *5*, 85. [CrossRef] [PubMed]
19. Rachev, S.T.; Klebanov, L.; Stoyanov, S.V.; Fabozzi, F. *The Methods of Distances in the Theory of Probability and Statistics*; Springer: New York, NY, USA, 2013.