The Genomic Structure: Proof of the Role of Non-Coding DNA

Nidhal Bouaynaya and Dan Schonfeld

Abstract— **We prove that the introns play the role of a decoy in absorbing mutations in the same way hollow uninhabited structures are used by the military to protect important installations. Our approach is based on a probability of error analysis, where errors are mutations which occur in the exon sequences. We derive the optimal exon length distribution, which minimizes the probability of error in the genome. Furthermore, to understand how can Nature generate the optimal distribution, we propose a diffusive random walk model for exon generation throughout evolution. This model results in an alpha stable exon length distribution, which is asymptotically equivalent to the optimal distribution. Experimental results show that both distributions accurately fit the real data. Given that introns also drive biological evolution by increasing the rate of unequal crossover between genes, we conclude that the role of introns is to maintain a genius balance between stability and adaptability in eukaryotic genomes.**

I. INTRODUCTION

The unexpected discovery of the intron-exon structure of eukaryotic genomes in 1977 struck the molecular biology community. The genes of eukaryotic genomes contain protein-coding sequences, called *exons*, separated by non-coding sequences, called *introns*. Thus, introns are excluded from the main gene function: making proteins. What is more intriguing is that introns make up a large portion of eukaryotic DNA. In humans, for example, approximately 30% of the human genome is made up of introns. Questions and speculations about the evolutionary origins and function of introns appeared immediately after their discovery. Finding the role of introns is critical in understanding the function and evolution of genomes. More than 25 years later, the subject is still an active area of research. The extra energy needed to maintain and process the introns, throughout evolution, seems to defy evolutionary logic; "The cell puts a huge amount of its energy into the creation of these introns, then discards them ... Nature would not go to all that trouble without a reason" [1].

On year after their discovery, Gilbert [2] advanced that recombination in intronic regions of genes increases the rate of creation of new genes by forming novel combinations of exons. Such shuffling must have speeded up evolution by accelerating the diversity of proteins and so of living things. However, the exon shuffling hypothesis does not explain the lack of introns in prokaryotic genomes. Information theory and coding techniques, which are well known to communication engineers, offer an alternative explanation to the nature of introns: Battail [3] hypothesized that error-correcting codes are used in the replication process of the genome. A consequence of this hypothesis is the existence of redundant DNA. The genes in the DNA are viewed as the encoded messages composed of the information symbols (i.e., exons) and the redundant symbols (i.e., introns) needed by the error-correction process. It is well known that DNA replication and protein synthesis involve error repair mechanisms [4]. However, no linkage has been found between these repair mechanisms and the intron sequences in the genes. An algebraic interpretation of the role of introns appeared in [5]. Using pure mathematics and specifically a branch of number theory called sequence algebra, Huen showed that the role of introns is to stabilize an ever changing genome. Moreover, the corrective action of the introns is never sufficient to bring the genome to the full equilibrium state, which spells stagnation and ultimately extinction.

We propose that introns control the balance between stability and adaptability in eukaryotic genomes. In this paper, we focus on the stability role of introns. The role of introns in driving evolution by increasing the rate of recombination of exons is inspired by Gilbert's exon shuffling hypothesis. However, unlike Gilbert, we do not necessarily claim that exons represent functionally and/or structurally important subunits of proteins nor do we adopt his intron-early view. All we claim is that the long sequences of introns make them hot spots for genetic recombination via unequal crossover. The role of introns in increasing the rate of unequal crossovers must be tempered in order to prevent excessive evolutionary adaptability. Rapid changes in the genomic code must not occur too frequently, or else we would experience evolutionary jumps in each generation. Based on probability of error analysis and optimization, we show that introns play the role of a decoy in absorbing many mutations modelled as the transmission errors of the biological communication channel presented in [6]. Introns protect coding regions in the DNA sequence from frequent errors in the same way hollow uninhabited structures are used by the military to protect important installations, such as aircraft hangars and missile launching facilities, from a bomb attack by serving as a dummy target that resembles the protected structure. The stability role attributed to introns accounts for at least two biological facts: (i) The absence of introns in prokaryotic genomes translates, according to our view, to a high mutability rate of these primitive organisms. It is widely known today that many bacteria and viruses rely on mutations for diversification. (ii) The decoy role for introns predicts that coding sequences should be more conserved among organisms than non-coding sequences. Studies in comparative genomics showed that functional DNA sequences tend to undergo mutation at a slower rate than nonfunctional sequences [7]. For example, the coding sequence of a human protein-coding gene is typically about 80% identical to its mouse ortholog, while their genomes as a whole are much more widely divergent. Nevertheless, one can legitimately ask: Why wouldnt Nature invest in more error correction mechanisms rather than carrying this enormous decoy luggage? We argue that several reasons lie behind this choice: First, if nature had to design error correction codes to control the exact rate of mutation required to simultaneously maintain life and encourage evolution, it would need to know the exact distribution and form of all possible mutations which occurred in the past and will occur in the future. Designing complex error correcting codes for a given noise model might be completely useless in the face of dynamic noise characteristics. Second, a reduction in the error rate comes at the price of an increase in complexity. Nature might have preferred to spend more energy in carrying the decoy sequences rather than investing in complex and costly error repair mechanisms.

This paper is organized as follows: In Section II, we presume a deterministic analysis of the optimal exon lengths, which minimize the probability of error in the genome. This analysis will motivate the need to consider a stochastic model for the exon length distribution. In Section III, we readdress the probability of error optimization problem assuming a stochastic distribution of the exon lengths. First, we derive the optimal exon length distribution, which minimizes the probability of error. Second, we address the question of a plausible physical realization of the optimal exon length distribution. Experimental results on real data are discussed in Section IV. Finally, Section V summarizes the main results of this paper and discusses future work.

II. Genomic Structure: Deterministic Analysis

A. Why not a random mutation model?

Proposition 1: Assume that the point mutation rate is randomly distributed in the genome, i.e., the occurrence of mutations is independent and identically distributed in all regions of the genome. Then, the probability of error is a decreasing function of the length of introns and is independent of the distribution of introns in the genome.

The proof of Proposition 1 follows immediately from the Binomial distribution characteristics. Hence, we see that a binomial error model does not account for the biological exon (or intron) length distribution inside the genome. In other words, the biological intron-exon distribution would be equivalent, from an error robustness criterion, to the distribution which groups all exons in the beginning of the gene and all introns at its end. Therefore, we need to consider a different mutation model, which can account for the observed intron-exon structure in eukaryotic genomes. We propose a Poisson mutation model. This choice is justified by numerous arguments. First, the Poisson distribution is the limiting distribution of the binomial when the probability of error is small and the genome size is large such that the rate of point mutation in a unit interval is held constant (De Moivre-Laplace theorem). Second, many rare random phenomena in nature follow a Poisson distribution, e.g., the number of winning tickets in a large lottery, the number of printing errors in a book, etc. In the remainder of this paper, we assume that the mutations are Poisson distributed in the genome.

B. Error robustness analysis

Assume that there are K exons of total length M in a gene of T nucleotides. Let l_k be the length of exon k. In this subsection, we answer the question: “What are the optimal exon lengths, $l_k^*,\ k = 1, \cdots, K$, which minimize the probability of error in the gene?”.

Proposition 2: Assume that the mutations are Poisson distributed with rate λ. Consider a genome of length T nucleotides including K exons having total length M. Let l_k be the length of the k^{th} exon. Then, the probability of error is given by

$$P_e = 1 - e^{-\lambda K T} \prod_{k=1}^{K} \sum_{n=0}^{T-l_k} \frac{\lambda^n (T-l_k)^n}{n!}. \tag{1}$$

Since $l_k \leq M$ for all $k = 1, \cdots, K$, we obtain an upper bound on the probability of error by truncating the summation in Eq. (1) to $T-M$ instead of $T-l_k$. Minimizing the maximum probability of error, $P_e^{\max}$, is more tractable analytically than minimizing the probability of error in Eq. (1). Using the Lagrange multiplier technique, with constraint $\sum_{k=1}^K l_k = M$, and taking the derivative of $P_e^{\max}$ with respect to l_k, we obtain the following coupled system for the optimal exon lengths:

$$l_{i_0} = M \frac{[\prod_{k\neq i_0} \sum_{n=0}^{T-M} \frac{\lambda^n (T-l_k)^n}{n!}][\frac{\sum_{n=1}^{T-M} \lambda^n (T-l_{i_0})^{n-1}}{(n-1)!}]}{\sum_{j=1}^{K}[\prod_{k\neq j} \sum_{n=0}^{T-M} \frac{\lambda^n (T-l_k)^n}{n!}][\frac{\sum_{n=1}^{T-M} \lambda^n (T-l_j)^{n-1}}{(n-1)!}]}. \tag{2}$$

An obvious solution to the system in Eq. (2) is obtained when $l_k^* = \frac{M}{K}$ for all $k = 1, \cdots, K$. This surprising simple result states that the optimal exon lengths are distributed according to a delta function centered at the mean value $\frac{M}{K}$. But, in nature, the exon lengths are not uniformly distributed in the genome (see Fig. 1). The reason this deterministic analysis fails in capturing the intron-exon distribution is that the genome is not a deterministic entity but rather a continuously evolving one. Therefore, a stochastic model for the exon lengths would be more appropriate to correctly describe the genome’s dynamic nature. The deterministic analysis does, however, capture some characteristics of the biological data in the following sense:

Proposition 3: Let $\delta_{\frac{M}{K}}$ be the delta function centered at $\frac{M}{K}$. For every $\rho > 0$, consider the measure d_ρ between a continuous unimodal probability density function f_X and $\delta_{\frac{M}{K}}$ given by

$$d_\rho(\delta_{\frac{M}{K}}, f_X) = 1 - Pr(X \in [\frac{M}{K} - \rho, \frac{M}{K} + \rho]). \tag{3}$$

Let x_0 be the mode of f_X. Then, $\text{argmin}_{x_0}\, d_\rho = \frac{M}{K}$. That is the mode of f_X, which minimizes the measure d_ρ, is equal to $\frac{M}{K}$.

The biological exon distribution is asymmetric given that its support is $[0, \infty]$. The mode of asymmetric distributions is always less or equal than their mean. From proposition 3, the distribution, which best approximates $\delta_{\frac{M}{K}}$ in the

d_ρ measure sense, would have its mode very close to its mean. Amazingly, the exon length distribution of the human genome has its mode almost equal to its mean obtained at about 170 nucleotides (see Fig. 1)!

Even though the deterministic analysis gave some insights on the optimality of the biological exon length distribution from an error minimization criterion, a stochastic model for the exon distribution is needed to capture the dynamics of the evolving genome.

III. Genomic Structure: Stochastic Analysis

A. Error Robustness Analysis

In this subsection, we readdress the probability of error optimization problem formulated in Section II assuming a stochastic distribution of the exon lengths. The following proposition establishes the new expression for the probability of error assuming an infinite genome length, i.e., $T = \infty$.

Proposition 4: Let $p(l)$ be the continuous distribution of the length of exons. Assume that there are K exons in a genome infinitely long. The mutations are assumed to be Poisson with parameter λ. Then the probability of error is given by

$$P_e = 1 - (\int_0^\infty e^{-\lambda l} p(l)\ dl)^K. \tag{4}$$

We want to determine the optimal exon length distribution, $p^*(l)$, which minimizes the probability of error subject to $\int_0^\infty p^*(l)\ dl = 1$. It can be easily shown that the delta function centered at 0, δ_0, satisfies this optimization problem. This solution is somehow intuitive: no exons implies no error! In order to get a meaningful solution to this optimization problem, we need to impose more constraints on the exon length distribution. For instance, the mean exon length should be larger than a pre-specified number l_0 or, in general, the α^{th} moment of $p(l)$ should be larger than l_0. Consequently, the stochastic optimization problem is reformulated as follows:

$$p^*(l) = \underset{p(l)}{\operatorname{argmax}} \int_0^\infty e^{-\lambda l} p(l)\ dl, \quad \text{subject to}$$

$$\begin{aligned} &1) \int_0^\infty p(l)\ dl = 1; \\ &2) \int_0^\infty l^{1+\alpha} p(l)\ dl \geq l_0, \ \text{for some } \alpha \geq 0. \end{aligned} \tag{5}$$

the optimization problem formulated in Eq. (5) is solved using the Euler-Lagrange equation. We obtain:

$$p^*(l) = \frac{p_0(1+\mu)}{e^{-\lambda l} + \gamma l^{1+\alpha} + \mu}, \tag{6}$$

where μ and γ are the Lagrange multipliers, which are determined numerically. Taking the derivative of p^*, it is easy to show that it has a unique maximum. Observe that the $(1+\alpha)^{\text{th}}$ moment of $p(l)$ is infinite; thus satisfying condition 2) in Eq. (5). This infinite moment agrees with the heavy tail characteristic of the biological exon length distribution (see Fig. 1). The parameter α determines the tail decay of the distribution for a given mutation rate λ.

At this point, it is interesting to ask ourselves: "How can Nature generate such a distribution? Is there a simple enough model for exon generation, which leads to the distribution p^*?" The answer is investigated in the next subsection.

B. A Random Walk model

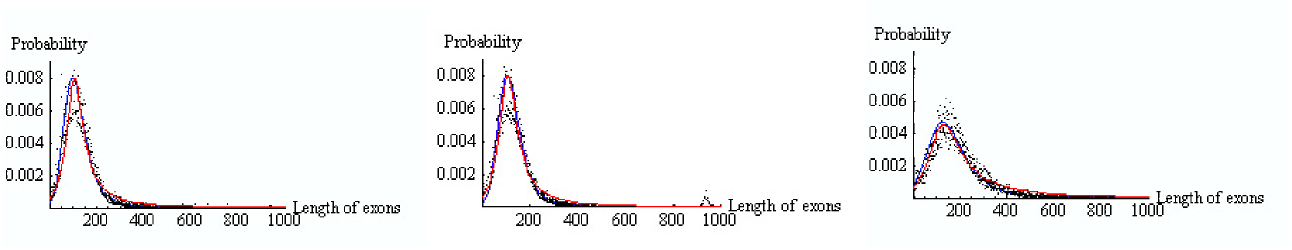
Insertion and deletion of exon nucleotides have been confirmed biologically for many primitive organisms. If, during evolution, exons were formed by insertion and deletion mechanisms, their lengths would follow some kind of a random walk. The length of the exon at any time corresponds to the position of the random walk. We assume that the sub-exons are formed independently by a stochastic process according to a distribution $f(l)$. So, the length of the final exon after N steps, X_N, is the sum of N independent displacements distributed according to $f(l)$, i.e., $X_N = \sum_{i=1}^N l_i$. Given the heavy tail characteristic of the biological exon length distribution, we assume that the sub-exons are generated by a distribution of the form:

$$f(l) = \alpha\ l^{-(\alpha+1)}, \quad l \geq 1. \tag{7}$$

where $0 < \alpha < 2$. We want to determine the limiting distribution of X_N as $N \to \infty$ or as the time $t \to \infty$. By the Generalized Central Limit Theorem [9], the density of X_N tends towards an alpha-stable distribution $p(l|\alpha, \beta, \sigma, \xi)$, where $-1 \leq \beta \leq 1$ is the skewness parameter, $\sigma > 0$ is the scale and $\xi \in \mathbb{R}$ is the location. Alpha-stable distributions do not have a closed form expression in general. They are defined by their characteristic function. Some of their prominent properties are: heavy tail, skewness of the distribution (when $\beta \neq 0$), and smooth unimodal density. Their asymptotic behavior is described by: $\lim_{|x|\to\infty} p(x|\alpha, \beta, \sigma, \xi) = \frac{C}{|x|^{1+\alpha}}$, where C is some constant [9]. Hence, from Eq. (6), we see that the optimal distribution p^* is asymptotically equivalent to an alpha-stable distribution. Nature would prefer to generate a simple random walk rather than solve the Euler-Lagrange equation!

IV. Experimental Results

The data files used were obtained from the NCBI web site: "ftp://ftp.ncbi.nih.gov/ genomes". In this paper, we choose to display the Homo Sapiens (Human), the Rattus Norvegicus (Rat) and the Apis Mellifera (Honey bee) exon length distributions. We extracted 281975 exons from the Homo Sapiens genome, 185769 from the Rat genome and 32753 exons from Apis Mellifera genome. Figure 1 shows the biological data, the optimal density and the alpha-stable distribution of the above organisms. For alpha-stable density fitting, we used the Mathematica package for stable distributions available from J. P. Nolan's website: "academic2.american.edu/∼jpnolan". The parameter α was estimated by plotting the data on a log-log scale and estimating the slope. The 1.5-stable distributions $p(l|1.5, 0.9, 35, 135)$, $p(l|0.85, 35, 140)$ and $p(l|1.5, 0.9, 60, 190)$ fit the exon length distributions of Homo Sapiens, Rat Norvegicus and Apis Mellifera, respectively. The same $\alpha = 1.5$ was used to display the optimal density $p^*(l)$ for these organisms. We experimentally



Homo Sapiens

Rattus norvegicus (Rat)

Apis Mellifera (Honey bee)

Fig. 1. Exon length distribution: The data points represent the biological data; the red curve is the optimal density, which minimizes the probability of error; and the blue curve is the fitted alpha-stable distribution. The graphs of the densities are truncated at exon lengths of 1000 nucleotides.

observed that, for each eukaryotic organism, there exists a couple (λ, α) which accurately fits the optimal distribution to the biological data. For $\alpha = 1.5, \lambda = 0.024$ for the three considered organisms. The mutation rate λ can be interpreted as the average rate of accepted mutations since the beginning of life on earth.

V. Conclusion

In this paper, we proved that the introns temper the effervescence of the ever-changing genome, under the chemical, physical and environmental conditions, by playing the role of a decoy for mutations. We also maintain that introns increase the rate of evolutionary adaptation by providing a mechanism for unequal crossover. The proposed dual role of introns serves to provide a balance between stability and adaptability. In our future work, we will investigate the intron length distribution in eukaryotic genomes by using a stochastic model of gene creation by means of unequal crossover.

Appendix

Proof: [Proof of Proposition 2] Let x_k denote the start position of the k^{th} exon in the genome. We have

$$P_e = 1 - \prod_{k=1}^{K} \; Pr \; (\text{“}0 \;\; \text{error in exon } k\text{”}), \tag{8}$$

where Pr (“0 error in exon k”)

$$\begin{aligned}
&= \sum_{n=0}^{T-l_k} e^{-\lambda l_k} \; Pr \; (\text{“}n \text{ errors outside } l_k) \\
&= \sum_{n=0}^{T-l_k} e^{-\lambda l_k} \{ \sum_{i=0}^{n} Pr \; (\text{“}i \text{ errors } \; \in [1, x_k - 1]\text{”}) \\
&\quad Pr \; (\text{“}(n-i) \text{ errors } \; \in [x_k + l_k, T]\text{”}) \} \\
&= \sum_{n=0}^{T-l_k} e^{-\lambda l_k} (\sum_{i=0}^{n} e^{-\lambda (x_k - 1)} \frac{(\lambda (x_k - 1))^i}{i!} \\
&\quad e^{-\lambda (T - x_k - l_k + 1)} \frac{(\lambda (T - x_k - l_k + 1))^i}{(n-i)!}) \\
&= \sum_{n=0}^{T-l_k} e^{-\lambda T} \sum_{i=0}^{n} \frac{\lambda^n}{n!} \binom{n}{i} (x_k - 1)^i (T - x_k - l_k + 1)^{n-i} \\
&= e^{-\lambda T} \sum_{n=0}^{T-l_k} \frac{\lambda^n}{n!} (T - l_k)^n. \qquad (9)
\end{aligned}$$

From Eqs (9) and (8), we obtain the desired expression of P_e. ■

Proof: [Proof of Proposition 3] Let f_X be a unimodal density which reaches its mode at x_0. Then $f_X(x - x_0)$ reaches its mode at 0.

$$x_0^* \;\; = \;\; \underset{x_0}{\text{argmax}} \int_{\frac{M}{K} - \rho}^{\frac{M}{K} + \rho} f_X(x - x_0) dx.$$

By continuity of f_X, $|(x - x_0) - (\frac{M}{K} - x_0)| < \rho \rightarrow |f_X(x - x_0) - f_X(\frac{M}{K} - x_0)| < \epsilon$, for some $\epsilon > 0$. So, we have

$$|\, \underset{x_0}{\text{argmax}} \{ 2\rho f_X(\frac{M}{K} - x_0) \} - x_0^* | \leq 2\rho \; \epsilon.$$

Since $f_X(x - x_0)$ reaches its mode at 0, we obtain $x_0^* = \frac{M}{K}$. ■

Proof: [Proof of Proposition 4]

$$\begin{aligned}
P_e &= 1 - \prod_{k=1}^{K} Pr(\text{“}0 \text{ error in exon } k\text{”}) \\
&= 1 - \prod_{k=1}^{K} \int_0^\infty Pr(\text{“}0 \text{ error in exon } k | l\text{”}) p(l) dl \\
&= 1 - \prod_{k=1}^{K} \int_0^\infty e^{-\lambda l} p(l) \; dl = 1 - (\int_0^\infty e^{-\lambda l} p(l) \; dl)^K.
\end{aligned}$$

■

References

[1] C. C. Kopezynski and M. A. T. Muskavitch, “Introns excised from the delta primarv transcript are localized near sites of delta transcription,” *The Journal of Cell Biology*, vol. 119, p. 503, 1992.

[2] W. Gilbert, “Why genes in pieces?” *Nature*, vol. 271, p. 501, February 1978.

[3] G. Battail, “Does information theory explain biological evolution,” *Europhysics Letters*, vol. 40, no. 3, pp. 343–348, 1997.

[4] R. D. Wood, M. Mitchell, J. Sgouros, and T. Lindahl, “Human dna repair genes,” *Science*, vol. 291, no. 5507, pp. 1284–1289, 2001.

[5] Y. Huen, “Brief comments on junk dna: is it really junk?” *Complexity International*, vol. 9, February 2002.

[6] N. Bouaynaya and D. Schonfeld, “Analysis of protein evolution as a communication system,” in *IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, May 2006.

[7] “Mouse genome sequencing consortium: Initial sequencing and comparative analysis of the mouse genome,” *Nature*, vol. 420, no. 6915, pp. 520–562, December 2002.

[8] S. E. Ptak and D. A. Petrov, “How intron splicing affects the deletion and insertion profile in drosophila melanogaster,” *Genetics*, vol. 162, pp. 1233–1244, 2002.

[9] B. Gnedenko and A. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley, 1954.