that fits the asymptotics of the problem.

References

[1] D. Aldous and P. Shields. A diffusion limit for a class of randomly growing binary trees. *Probability Theory*, 79:509–542, 1988.

[2] R. Breathnach, C. Benoist, K. O'Hare, F. Gannon, and P. Chambon. Ovalbumin gene: Evidence for leader sequence in mRNA and DNA sequences at the exon-intron boundaries. *Proceedings of the National Academy of Science*, 75:4853–4857, 1978.

[3] S. Brunak, J. Engelbrecht, and S. Knudsen. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *Journal of Molecular Biology*, 220:49, 1991.

[4] Jack Cophen and Ian Stewart. The information in your hand. *The Mathematical Intelligencer*, 13(3), 1991.

[5] R. G. Gallager. *Information Theory and Reliable Communication.* John Wiley & Sons, Inc., 1968.

[6] Ali Hariri, Bruce Weber, and John Olmstead. On the validity of Shannon-information calculations for molecular biological sequence. *Journal of Theoretical Biology*, 147:235–254, 1990.

[7] W. B. Davenport Jr. and W. L. Root. *An Introduction to the Theory of Random Signals and Noise.* McGraw-Hill, 1958.

[8] Andrzej Knopka and John Owens. Complexity charts can be used to map functional domains in DNA. *Gene Anal. Techn.*, 6, 1989.

[9] S.M. Mount. A catalogue of splice-junction sequences. *Nucleic Acids Research*, 10:459–472, 1982.

[10] H.M. Seidel, D.L. Pompliano, and J.R. Knowles. Exons as microgenes? *Science*, 257, September 1992.

[11] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948.

[12] Peter S. Shenkin, Batu Erman, and Lucy D. Mastrandrea. Information-theoretical entropy as a measure of sequence variability. *Proteins*, 11(4):297, 1991.

[13] R. Staden. Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. *Nucleic Acids Research*, 12:551–567, 1984.

[14] J.A. Steitz. Snurps. *Scientific American*, 258(6), June 1988.

[15] H. van Trees. *Detection, estimation and modulation theory.* Wiley, 1971.

[16] J. D. Watson, N. H. Hopkins, J. W. Roberts, J. Argetsinger Steitz, and A. M. Weiner. *Molecular Biology of the Gene.* Benjamin/Cummings, Menlo Park, CA, fourth edition, 1987.

[17] A.D. Wyner and A.J. Wyner. An improved version of the Lempel-Ziv algorithm. *Transactions of Information Theory.*

[18] A.J. Wyner. *String Matching Theorems and Applications to Data Compression and Statistics.* PhD thesis, Stanford University, 1993.

[19] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, IT-23(3):337–343, 1977.

38839 non-intron beginning **GT** pairs which are surrounded by the seven-letter patterns of interest. We have seen before that the letters surrounding an arbitrary **GT** pair are close in an entropy sense to being independent and equiprobably taking on the values from the set {**A**,**C**,**G**,**T**}. Hence, we will assume that $Pr(z_0|H_0) \approx Pr(z_1|H_0)$ for all z_0 and $z_1 \in U^{(7)}$. Under this assumption, we choose our discriminating set by greedily including the z which maximizes $\frac{Pr(z|H_1)}{Pr(z|H_0)}$. If there is more than one such z, pick one which maximizes $Pr(z|H_1)$. We procede until we achieve a false positive rate $p_F > \alpha$. Such a heuristic produces a decision rule which is sometimes optimal and always "good" in the Neyman-Pearson sense.

The currently existing consensus rules suggest that there are four patterns which differentiate the beginning of an intron from an arbitrary **GT** pair. These are **AAG<u>GT</u>AAGT**, **AAG<u>GT</u>GAGT**, **CAG<u>GT</u>AAGT**, and **CAG<u>GT</u>GAGT**. According to our data set, if Z_1 is selected to be the set of these four patterns, then $p_D = 0.041812$, and $p_F = 0.000232$. Clearly, if this Z_1 were used in a parsing rule, the performance of the rule would be extremely poor, at least for this data set. One can do much better by taking advantage of the information collected by the method we describe, and which is summarized in Figure 4. For example, we can find Z_1 which will satisfy $p_D = 0.212544$ and $p_F = 0$, or if we are willing to let p_F be as large as 0.000232, we can use a decision rule with $p_D = 0.297909$ and $p_F = 0.000206$. For this data set, if we use all 265 patterns specified above, we have $p_D = 1.0$ and $p_F = 0.023173$, and this value of p_F may be acceptable for many parsing applications. A list of patterns to achieve any point on the curve is available upon request.

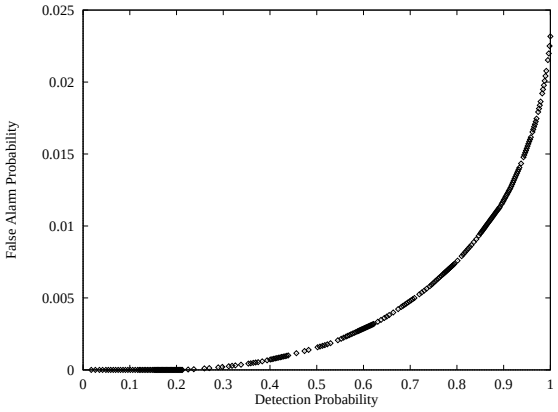

Figure 4: Neyman-Pearson Criterion for Splice Junction Detection

5 Conclusion

While entropy measures information, most attempts to provide meaningful or useful entropy based characterization of DNA have ended in failure. We suggest that this is because the wrong entropy estimators were used in previous studies. In particular, the most well known entropy estimators have notoriously slow convergence rates. The main result of this paper is that we find that the entropy of exons is higher than that of introns. This seems surprising in that introns are presumed to be the mechanism by which many random changes can accumulate without being subjected to restorative survival forces and thereby produce an entirely new gene, without each small incremental change being more fit for survival.

The natural explanation which occurs for our observation, based on using a new estimator of entropy, more suitable for estimating the entropy of a short string, is that (a large) part of the introns are also subject to restorative forces, and that some or all of these parts may serve to define and determine the splice juctions, *i.e.* the intron-exon boundaries. If this is the case, there must be rules which are coded into the introns and should be inferrable, given enough data. If these regions are the only parts of the introns subject to restorative forces then the rules must be complicated – which seems to be the case. Indeed, we attempt to find these (presumed) coding rules for the splice junctions in the introns, but unfortunately conclude that we will need much more data to infer the rules and solve this fascinating and important problem.

We believe that designers of algorithms which rely on some entropic property of a source must be careful not to misuse entropy estimators. In particular, in practical settings in which an algorithm relies on an estimation of the entropy of a source, one must be innovative in choosing an estimator

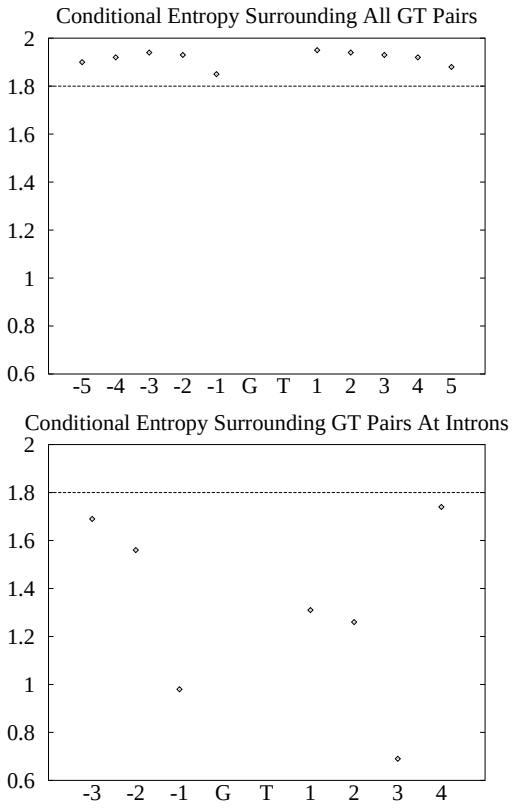

Figure 3: Estimates of Conditional Entropy Profiles

ference in $H(\tilde{L}_n|\tilde{L}_1^{n-1})$ from the value it previously assumed indicates that the n^{th} letter preceding a **GT** pair is information that may be used by the splicing mechanism.

To summarize our findings, we see that the beginning of an intron appears to be linked to certain patterns of the form $\{xxx\mathbf{GT}xxxx\}$, where the **GT** marks the beginning of the intron. We look at these patterns in the following section. Similar estimates of entropy profiles have been attempted at the end of introns. Unfortunately, the results of the tests do not give clear indications of the type of pattern which differentiates the ends of introns from other **AG** pairs. This quantifies earlier observations that the ends of introns are more difficult to predict than the beginings [9, 13].

4.2 Patterns We have reduced our search-space for good discriminators to the 7-tuples surrounding a **GT** . We implement a discriminator by applying the well-known statistical test called the *Neyman-Pearson criterion* [15, 7] to produce a useful rule to decide whether or not a **GT** pair should be classified as an intron beginning.

We let z be the seven-letter pattern of interest surrounding a certain **GT** pair. Upon observing z, we wish to choose one of the following two hypotheses:
H_0: The **GT** pair does not mark the beginning of an intron.
H_1: The **GT** pair does mark the beginning of an intron.
If $U^{(7)}$ denotes the set of $4^7 = 16384$ possible seven-letter patterns, then a decision rule will partition $U^{(7)}$ into two sets Z_0 and Z_1 with the following properties:

1. Every element $z \in U^{(7)}$ will be an element in exactly one of the sets Z_0 and Z_1.

2. If $z \in Z_0$, we select hypothesis H_0 and otherwise we select hypothesis H_1.

The performance of a decision rule can be measured by two criteria known as the *detection probability* and the *false alarm probability*, which are denoted by p_D and p_F, respectively. p_D is the probability that the decision rule correctly decides that a **GT** pair which begins an intron is an intron beginning; p_F is the probability that the decision rule incorrectly decides that a **GT** pair which does not begin an intron is an intron beginning. In other words,

$$p_D = \sum_{z \in Z_1} Pr(z|H_1); \qquad p_F = \sum_{z \in Z_1} Pr(z|H_0)$$

In general, we would like to pick a decision rule which makes p_D as large as possible and p_F as small as possible. These objectives conflict, so there is a trade-off between these goals. However, we can consider the following problem: if we are given a constraint on the maximum acceptable value of p_F, we would like to select a decision rule which maximizes p_D while satisfying the constraint on p_F. The solution to this problem is known as a *Neyman-Pearson criterion.* As usual, for any $z \in U^{(7)}$, we approximate $Pr(z|H_0)$ and $Pr(z|H_1)$ by their sample values in our data set. For the 143 genes we considered, there are 574 intron beginnings and

introns is identical to that of the exons (or more strongly, that they are stochastically equivalent), $E[Y_{ij}] = 0$. Using this data to perform the signed rank test on the paired comparisons of adjacent exon/intron sequences, we found that of the 303 comparisons, 73% found the average match length to be larger for the intron. Our conclusion is that the data does not support the equivalence hypothesis (with power $P \approx 10^{-5}$).

Since $\bar{L}$ is a difficult quantity to pin down mathematically, we ran a control test on a randomly selected test sequence chosen with equal probability among the 4 characters {**A**,**C**,**G**,**T**}. As expected the paired comparison test showed no significant difference between the groups.

3.2 Variability Measure We performed the identical signed rank test on the paired comparisons, using the variance of the match lengths instead of the mean. One would expect the match lengths to have a greater variability for the lower entropic sequences – and this was observed. Surprisingly, the variance measure proved to be an even more sensitive discriminator. We found that the variability of the intron match lengths, in 80% of the paired comparisons, to be higher than that of its neighboring exon. In fact, in a large number of pairs with lower *exon* entropy, the intron variability was still greater.

4 Detecting Splice Junctions

4.1 Entropy Tests Recall from Section 1.1 that each intron begins with a **GT** and ends with an **AG** . We now consider the task of discriminating between an arbitrary **GT** or **AG** from one that signals a splice junction. We will focus for now on the **GT** discriminator. Our data set provides 39439 strings which begin with **GT** . It contains 579 introns. Our method will be to use conditional entropies to find which bits flanking the **GT** pairs are significantly correlated with an occurrence of **GT** .

Let R_n denote the ensemble of n^{th} letters following all **GT** 's, and let R_1^{n-1} denote the set of first $n-1$ letters following all **GT** 's. From the properties of entropy we mentioned in the introduction, $H(R_n|R_1^{n-1}) \leq H(R_n) \leq 2$. Hence, if we observe that $H(R_n|R_1^{n-1}) \approx 2$, we can make two statements about the n^{th} letter in in a **GT** string. The first is that the n^{th} letter is chosen nearly equiprobably from the set of characters and the second is that the n^{th} letter in a **GT** string is essentially independent of the previous letters in the string. We will arbitrarily say that $H(R_n|R_1^{n-1}) \approx 2$ if $H(R_n|R_1^{n-1}) > 1.8$. We can compute a symmetric profile for characters to the left of each **GT** . The are shown on the left plot of Figure 3. For $n > 5$, we do not have enough data to make statistically significant estimates since $4^n \geq 4^6 = 4096$. Because of the size of our data set, our heuristic is to estimate $H(R_n|R_1^{n-1})$ by $H(R_n|R_1^5)$; this quantity measures the dependence of the n^{th} letter of a string to the right of a **GT** with the first five letters of the string. With this approximation, we find that $H(R_n|R_1^{n-1}) \approx 2$ for all $n < 31$, and similarly to the left. Symmetric experiments were performed for the symbols to the left of the **GT** boundaries with similar results.

Now, we can restrict ourselves to the characters flanking the **GT** 's at the begining of introns. The right plot of Figure 3 give this results. The $\tilde{R}$ variables are defined analogously to the R variables of the preceding paragraph. Here, for $n \geq 4$, the size of the data set prevents the sample values of $H(\tilde{R}_n|\tilde{R}_1^{n-1})$ from being statistically significant. To be consistent with our earlier estimation technique, we resort to approximating $H(\tilde{R}_n|\tilde{R}_1^{n-1})$ with $H(\tilde{R}_n|\tilde{R}_1^3)$. For $5 \leq n \leq 30$, we have that $H(\tilde{R}_n|\tilde{R}_1^{n-1}) \approx 2$, which suggests that we can ignore the $5^{th}, 6^{th}, \ldots, 30^{th}$ letters in the intron in our study of the splicing mechanism. The currently existing theories about splicing suggest that we can also ignore letters that are even further away from the beginning of the intron. For $n = 4$, we have that $H(E_4^r) \simeq 1.74$, so the fourth letter of an intron may or may not be helpful in differentiating the intron from an arbitrary string which begins with **GT** . We have decided not to ignore it because the splice junction consensus indicates that it might be important. The $\tilde{L}$ variables are defined symmetrically to the $\tilde{R}$ variables (i.e., they look at the patterns immediately to the left of an intron beginning). For $n \leq 3$, the fairly significant dif-

bias term of $O(1/\log N_w)$ from Theorem 3. This error comes from the fact that we do not know the length of the memory of the process. If the memory is greater than N_w we are losing a source for predictivity. However, the error term is only a bound on how extensive that source of error can be, and indeed we have no prior reason to suspect bias at all.

There are no clear rules to follow in selecting the size of the window. For a data set of n samples, there are roughly n/N_w "independent" match lengths. This implies that a prudent choice would be to let $\log N_w$ be approximately equal to the standard error of $\bar{L}$. This can surely not be a hardfast rule since there is some reason to believe that the $O(1)$ bound in Theorem 3 may be, in many cases, quite small.

We point out several assumptions in our analysis:

- The entropy measure we are using only approximates an entropy measure, due to the fact that the memory of the "process" is longer than our N_w.
- DNA is not stationary. This is probably a reason why the entropy estimator does better — it is more robust to weaker conditions and non-stationary processes cannot be characterized by entropy.
- DNA is not a random process. Mathematics serves only as a guide in looking for a useful statistic, we do not suggest that we have characterized the "entropy of DNA".

3 Results of Entropy Estimation

In our preliminary studies, the use of Lempel-Ziv compression suggested different entropy estimates for exons versus introns, and indicated that the variability of introns was higher that that of exons. However, due to the fact that the new entropy estimation methods described below converge faster, we now have more reliable evidence. This follows from the fact that Lempel-Ziv carries out a string-matching operation only once per phrase while the new method does it for every letter.

We will use the sliding window entropy estimate on the genetic material of introns and exons. To do this, we will consider DNA to be a stochastic source defined on sequences of letters from the alphabet $U = \{\mathbf{A},\mathbf{C},\mathbf{G},\mathbf{T}\}$. To this end, let $X_1, X_2, \ldots$ be a sequence of base pairs representing a pure sequence of either exon or intron material. We will apply the sliding window entropy estimator to our experimental set of exon or intron material. Recall that we define, for any choice of N_w:

$$L_i = \min\{k : X_{i+1}^{i+k+1} \not\subseteq X_{i-N_w+1}^{i}\}.$$

This defines the *longest match* of the sequence at position i with respect to the N_w neighbors to the left. We let $\bar{L}$ be the average match length. It is hoped that the difference in the makeup of the exon and the introns will be reflected in the average sliding match length statistic, $\bar{L}$. The usefulness of this statistic is established by Theorem 3, and the Ergodic Theorem, which implies the convergence of $\bar{L}$ to $E[L]$. We must establish, however, the assumption that DNA has a fixed memory, so that Equation 2.2 will hold.

3.1 Entropy Estimate This first experiment presupposes the knowledge of the boundaries. This information (which we assume to be accurate) allows us to create sequences of pure exon and pure intron material. We were then able to form the sliding window entropy estimate $\bar{L}$ for each exon and intron along 66 different genes (see Section 1.4). This process produces the random variables L_{ij}^{type}; where *type* is either exon or intron, i is the index of the gene, and j is index of occurrence within the gene, that is, L_{ij}^{type} is the length of the j^{th} segment of type *type* in gene i. For this experiment we let N_w, the window size, be 16.

We performed two tests on the data. The difference between the entropies is small (for each were nearly maximal entropy). We chose to perform a signed rank test, selecting, with out loss of generality, the event of interest to be when the estimated entropy of the intron was larger than the estimated entropy of the exon.

To this end let,

$$Y_{ij} = sgn[L_{ij}^{exon} - L_{ij}^{intron}].$$

Then, under the hypothesis that the entropy of the

an entropy estimate. Indeed, it is a popular choice since it is easy to implement and is universally applicable (empirically it performs very robustly in situations that are neither stationary nor ergodic, but we will return to this point later). Furthermore, the string matching concept that lies behind this approach is intuitively appealing as a complexity measure since it quantitatively captures repetitive structure. The drawbacks to this scheme include a notoriously slow rate of convergence due to the large number of observations needed to build the dictionary of patterns. Furthermore, there are very few known distributional properties (necessary to any statistical application). It has been shown [1] that C_n is asymptotically Normal, but this was proved only for memoryless sources with $p = .5$ (this case appears later as a pathological example).

2.1 A Different Approach Another technique is based on the Fixed-Database LZ algorithm, a method that very closely resembles practical versions that are widely in use, and that is very similar in spirit to all other versions of the algorithm (through their common usage of string matching). Let X_1^∞ be the data source, which we wish to compress or estimate its entropy. We assume that we have a "database", D_n, of n observations X_{-n+1}^0. We define the longest match into D_n of the incoming sequence $X_1, X_2, \ldots$ by

$$L = \inf\{k : X_1^{k+1} \not\subseteq D_n\}$$

where $\not\subseteq$ means "as a contiguous substring". For example, if

$$D_n = \{0011010011000100\}$$

with $n = 16$ and $X_1, X_2, \ldots = \{0100100\ldots\}$ then $L = 5$ since the string $\{01001\} \subseteq D_n$ but the extension $\{010010\} \not\subseteq D_n$. The following theorems (see [18] for proofs) provide insight as to how L, the longest match, can be used to estimate entropy (an alternative proof of theorem 3 appears in [17].

THEOREM 1. *If $\{X_k\}$ is Uniform[A] i.i.d., then for any positive integers l and n*

$$\Pr\{L < l + \log n\} \approx \exp(-2^l).$$

A surprising result shows the uniform case to be a remarkable pathology:

THEOREM 2. *Let $\{X_k\}$ be a stationary, ergodic source with finite memory that is not uniform i.i.d. Thus, $Pr\{X_k = x_k | X_{-\infty}^{k-1} = x_{-\infty}^{k-1}\} = \Pr\{X_k = x_k | X_{k-M}^{k-1} = x_{k-M}^{k-1}\}$. In such a situation the asymptotic distribution of L is normal with mean $\mu = \frac{\log n}{H}$ and variance $\sigma^2 = \frac{\log n}{H}$ where σ^2 is the variance of the $\log P(X)$ defined as:*

$$\sigma^2 = \lim_{n\to\infty} \frac{Var(-\frac{\log P(X_1^n)}{\log |\mathcal{A}|})}{n}.$$

The last theorem provides us with the first moment of L as well as the relationship of L to the entropy H:

THEOREM 3. *As $n \to \infty$, $|E[L] - \frac{\log n}{H}| = O(1)$.*

In light of these theorems we form an estimate of the entropy of the process based on the length of repeated patterns. This scheme is potentially better than the the data-compression scheme that is based on the length of repeated patterns when used as an entropy estimator, since we apply it repeatedly to each incoming letter.

2.2 A Sliding Window Entropy Estimate Choose a positive integer N_w. This parameter will be the size of a "window" of observations that will serve as the database into which we will reference incoming data to find the longest match. Given the sequence $X_1, X_2, \ldots$, define, for every index i, the longest match of the string of observation to the right of i, into the string of observations in a window of size N_w to the left of i. Formally, let

$$L_i = \min\{k : X_{i+1}^{i+k+1} \not\subseteq X_{i-N_w+1}^{i}\} \qquad (2.1)$$

This defines a sequence of random variables $\{L_i\}$. Theorem 3 suggests the following entropy estimator:

$$\hat{H} = \frac{\log_2 N_w}{\bar{L}}, \qquad (2.2)$$

where $\bar{L}$ is the average of L_i. Theorems 1 and 2 suggest that this convergence takes place with an error that is $O(\frac{1}{\log N_w})$. In fact, there are therefore two sources of statistical error in the estimate; (1) the standard error of $\bar{L}$ for fixed N_w (ameliorated by large datasets) and (2) the

The results in Sections 2 and 3 do not, however, imply a method for finding intron/exon boundaries, and as noted above, the false positive rate of the consensus sequences method (as well as methods based on neural nets [3]) make them virtually useless for this task. One of the main features of consensus sequences is that they are *memoryless* models of canonical sequences, that is, they fail to take into account correlations between positions surrounding the splice junction.

- In Section 4, we use conditional entropy as a tool for deciding which bits of information surrounding a splice junction are relevant, and we proceed to derive an algorithm for splice-junction detections. Our experiments show that our method is significantly superior, in terms of sensitivity/specificity tradeoff, to previous methods for detecting intron/exon boundaries.

1.4 Data Sequence data were obtained from GenBank release 80.0, and included only human genes which were described as "complete coding sequences". Within these sequences were 1275147 bases, with 659 introns and 669 exons. The average exon size was 184 bases with a standard deviation of 96, and the average intron size was 867 bases with a standard deviation of 583. The median lengths, however, are much shorter. Exons and introns have median lengths of 139 and 434, respectively. This fact is important in the methods described below with respect to choosing the size of the window size (N_w). Since most exons are quite a bit shorter than the average the window size must be kept small in order to make sure that data is not lost.

2 Methods of Entropy Estimation

There are several common methods of estimating the entropy of a random process; we will describe several. The most straightforward would be to attempt a direct computation of the expected log of the empirical distribution function. Using this approach the entropy estimate would be only as accurate an estimate as the estimate of the probability of n-tuples for n large. It should be clear that this technique is generally impractical, since in most cases the amount of data is insufficient to achieve a good estimate of all but the marginal (first order) distribution and perhaps the distribution of pairs. If, however, it is possible to determine *a priori* that the process is of small Markov order, then such a scheme may provide satisfactory results.

Another popular choice involves data compression schemes. Such a procedure would involve compressing the data, measuring the total compression, and thereby determining an upper bound for the entropy of the process. If the algorithm is universal, then the compression ratio will approach the entropy as the size of the dataset increases. This method provides a startling advantage over the previous one: since it is not based on a model, no specific underlying structure of the process need be assumed. However, the approach, although very efficient in compressing data, is seriously impaired by a slow rate of convergence when used to estimate entropies.

For example (and for context) we will explain how a version of the the Lempel-Ziv data compression algorithm can be used to estimate entropy. Although there are many modifications of the original algorithm, they are all sufficiently alike in spirit to consider a representative. In fact, all versions reflect a common usage of string matching and pattern frequency; we demonstrate in the following example:

EXAMPLE 1. (LZ ALGORITHM [19])
Consider a sequence of binary data, e.g. $\{0010111000101\}$. *We parse the sequence into unique phrases by placing commas after every contiguous substring completes a "new" pattern. This pattern forms a phrase which becomes part of the "dictionary" of patterns, with new phrase formed by searching left to right down the sequence to find the shortest contiguous substring that is not already in the dictionary. For example, the sequence above would be parsed into* $\{0, 01, 011, 1, 00, 010, ..\}$. *Notice that every phrase is unique, and that new phrases are formed by extending previous phrases by exactly one symbol.*
Let C_n be the number of commas formed in a LZ parse of a sequence of length n. Ziv and Lempel [19] have shown that the quantity $\frac{C_n \log C_n}{n} \to H$ as $n \to \infty$, which makes it an obvious candidate for

entropy is understood to be a measure of randomness. However, little work has been done to examine differences between entropies of introns and exons. Knopka and Owens [8] estimated a value they termed "Local Compositional Complexity", which corresponds to the one-dimensional entropy of a DNA sequence. This parameter, based on frequency counts of nucleotides from a large number of different sources, showed a maximum value for introns.

1.2 Entropy Information Entropy was introduced by Shannon [11]. We first provide some definitions. Let $(X_1, X_2, \ldots) = X_1^n$ be a stochastic process with probability law $\mathcal{P}$. For every positive integer l, and every possible sequence of outcomes (from the alphabet $\mathcal{A}$) $x_1^l \in \{\mathcal{A}\}^l$, define the probability or likelihood function to be $P(x_1^l) = \Pr\{X_1^l = x_1^l\}$. Thus P maps every sequence to its probability under $\mathcal{P}$. The Entropy, $H(P)$, is defined as the following limit,

$$H(P) = \lim_{n\to\infty} \frac{E[-\log_2 P(X_1^n)]}{n}.$$

We have chosen to consider entropy because it is a natural measure of the following phenomena: complexity, compressibility, predictivity, and randomness. A common feature of all these qualities is that they are all properties of individual sequences, even those that are not thought of as outcomes of a random process. It was the genius of Shannon to consider "messages" to be random quantities, which allowed him to proceed with a general theory with far-reaching consequences, even though the theory was truly inapplicable in most situations. Accordingly, when we now consider the entropy of sequences, it becomes important to point out that entropy is a property of distributions, not, at least directly, a property of individual sequences.

Confusion about the notion of information theoretic entropy has led to incorrect conclusions in previous work. Hariri, et al. [6] examine the efficacy of several poor entropy estimators, and misapply the methods to individual DNA sequences. Furthermore, they draw an invalid conclusion from a non-existent correspondence between the chemical entropy and the information theoretic entropy of a DNA molecule. Cophen and Stewart [4] confuse entropy and information content. This is a fairly usual mistake that occurs when "context" is considered, and Shannon realized that we could only have a mathematical theory of information in a context-free universe. The appropriate approach is to consider, as Shannon did, sequences to be outcomes of stochastic processes and then to *estimate* the entropy of the distribution (or distributions) from which the observed data would be typical.

The notion of entropy can also be expanded to account for memory. For example, let us consider the chance variables U and V. The *conditional entropy of U given that $V = v$* is defined by

$$H(U|v) = -\sum_{u\in U} Prob(u|v) * \log_2 Prob(u|v)$$

and the *conditional entropy of U with respect to V* is defined by

$$H(U|V) = E_v[H(U|v)]$$

where E_v denotes the operation of taking an expectation with respect to the elements of V. It is known that for any random variable V, $H(U|V) \leq H(U)$ with equality if and only if U and V are statistically independent (see, for example, [5]).

1.3 Our Results Our results are two-fold. First, we consider the question of whether there is an information theoretic difference between introns and exons. Having noted the previous failure by researchers to find such a distinction, we nonetheless tried standard methods for estimating the entropies of sequences, i.e. compression, and character singleton and tuple distributions. Not surprisingly, we found no statistically significant difference between these measures for introns and exon. We noted, however, that exons are quite short, and our estimators may not have enough time to converge to the correct entropy.

- In Section 2, we give theorems which show that another entropy estimator, the *match-length* estimator, has a significantly faster convergence rate. Indeed, we show that a significant difference exists, both qualitatively and quantitatively, in the intron and exon entropies using this scheme. The experimental results are described in Section 3.

rates than previously known methods.

Our work suggests that entropy is a useful tool in exploring DNA. We suggest some reasons why previous researchers failed to detect significant variance in entropy of genomic regions, and offer a cautionary tail to those interested in using entropy based tools in settings where asymptotic complexity breaks down. Finally, we provide a discussion of the role of a proper mathematical treatment of entropy in a biological setting.

1.1 Biology DNA and proteins are polymers, constructed of subunits known as nucleotides and amino acids, respectively. The sequence of each protein is a function of a DNA sequence which serves as the "gene" for that protein. The cellular expression of proteins proceeds by the creation of a "message" copy from the DNA template into a closely related molecule known as RNA (Figure 1). This RNA is then translated into a protein.

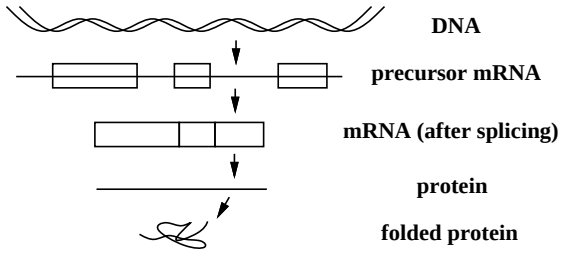

Figure 1: The flow of information in a eukaryotic cell.

One of the most unexpected findings in molecular biology is that large pieces of the RNA are removed before it is translated further [2]. The majority of eukaryotic[1] genes display a complex structure in which sequences which code for protein are interrupted by intervening, non-coding sequences. Initial transcription of these genes results in a pre-message RNA molecule from which segments must be accurately removed to produce a translatable message.

The retained sequences (represented by boxes in Figure 1) are known as *exons*, while the removed sequences are known as *introns*. Exons tend to be no more than 200 characters long, while introns can be many tens of thousands of characters long. Thus the majority of a typical eukaryotic gene will consist of intron regions. Since the discovery of such "split genes" over a decade ago, the nature of the splicing event has been the subject of intense research (for a recent study see [10, 14]). RNA precursors contain patterns similar to those in the Figure 2. The points at which RNA is removed (the boundaries of the boxes in Figure 2) are known as *splice-junctions*. Evolutionary conservation for splice junctions is commonly observed by the construction of a "consensus" sequence – a process by which many sequences are aligned and a composite subsequence is created by taking the majority base at each position. Such a composite is used both to support biological inferences, and as a discriminant tool for the recognition of intron/exon boundaries in raw sequence data now being produced by the various genome projects.

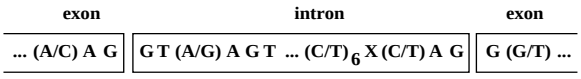

Figure 2: Splice junction consensus.

One feature which will be important in our detection algorithm (Section 4) will be the fact that introns *almost always* begin with a **GT** and end with an **AG** . However, numerous other locations can resemble these canonical patterns. As a result, these patterns do not by themselves reliably imply the presence of a splice-junction. Evidently, if junctions are to be recognized on the basis of sequence information alone, longer-range sequence information will have to be included in the decision-making criteria. A central problem is therefore to determine the extent to which sequences surrounding splice-junctions differ from sequences surrounding spurious analogues.

Finally, while mutations are thought to occur with approximately the same frequency in introns and exons, exons are subject to greater selective pressure than introns. Thus, it makes good intuitive sense that their entropy may be different, since they are subject to different random processes, and

[1] Eukaryotic cells contain nuclei, unlike prokaryotic cells such as bacterial and viruses. See [16] for a general introduction to molecular biology

On the Entropy of DNA: Algorithms and Measurements based on Memory and Rapid Convergence*

Martin Farach[†]
Dept of CS
Rutgers Univ.

Michiel Noordewier[‡]
Dept of CS
Rutgers Univ.

Serap Savari[§]
Dept. of EE
MIT

Larry Shepp[¶]
AT&T Bell Labs

Abraham Wyner[‖]
Dept. of Stat.
Stanford Univ.

Jacob Ziv[**]
Dept. of EE
Technion Inst.

November 1, 1994

Abstract

We have applied the information theoretic notion of *entropy* to characterize DNA sequences. We consider a genetic sequence signal that is too small for asymptotic entropy estimates to be accurate, and for which similar approaches have previously failed. We prove that the *match length* entropy estimator has a relatively fast converge rate and demonstrate experimentally that by using this entropy estimator, we can indeed extract a meaningful signal from segments of DNA. Further, we derive a method for detecting certain signals within DNA – known as *splice junctions* – with significantly better performance than previously known methods.

The main result of this paper is that we find that the entropy of genetic material which is ultimately expressed in protein sequences is higher than that which is discarded. This is an unexpected result, since current biological theory holds that the discarded sequences ("introns") are capable of tolerating random changes to a greater degree than the retained sequences ("exons").

1 Introduction

DNA carries the instructions for the operation of living organisms. The simple combinatorial structure of the DNA molecule has been an obvious lure to theorists interested in studying the way information is transmitted in living organisms. Most such attempts have had little or no success [12], to the point where some researchers have denied the utility of information theory, and more specifically information theoretic entropy, in the study of DNA [6].

In this paper, we examine the utility of information theoretic tools in a classic setting, the *intron/exon boundary problem* (described below). Prediction of these boundaries in the sequence of DNA is an essential task if we are to predict the product of a gene. We give the *first experimental verification* of an entropic difference between introns and exons, based on novel entropy estimation methods with very fast convergence times. The convergence time is important since exons tend to be quite small. Thus, the fact the methods exit that approximate entropy *in the limit* has proven to be useless for finding the entropy of genetic regions. Further, we give an entropy-based algorithm for intron/exon boundary detection. Our studies show that *we achieve better false positive/false negative*

[†] farach@cs.rutgers.edu; Supported by DIMACS (Center for Discrete Mathematics and Theoretical Computer Science), a National Science Foundation Science and Technology Center under NSF contract STC-8809648.

[‡] noordewi@cs.rutgers.edu

[§] ayse@mit.edu

[¶] las@research.att.com

[‖] ajw@playfair.stanford.edu

[**] jz@ee.technion.ac.il