Abstract—We demonstrate the use of information-theoretic tools for the task of identifying segments of biomolecules (DNA or RNA) that are statistically correlated. We develop a precise and reliable methodology, based on the notion of mutual information, for finding and extracting statistical as well as structural dependencies. A simple threshold function is defined, and its use in quantifying the level of significance of dependencies between biological segments is explored. These tools are used in two specific applications. First, for the identification of correlations between different parts of the maize zmSRp32 gene. There, we find significant dependencies between the 5' untranslated region and its alternatively spliced exons. This observation may indicate the presence of as-yet unknown alternative splicing mechanisms or structural scaffolds. Second, using data from CODIS, we demonstrate that our approach is well suited for the problem of discovering short tandem repeats (STRs).

I. INTRODUCTION

Questions of quantification, representation, and description of the overall flow of information in biosystems are of central importance in the life sciences. We develop statistical tools based on information-theoretic ideas, and demonstrate their use in identifying informative parts in biomolecules. Specifically, our goal is to detect statistically dependent segments of biosequences, hoping to reveal potentially important biological phenomena. It is well-known [5], [18], [20] that various parts of biomolecules, such as DNA, RNA, and proteins, are significantly (statistically) correlated, although formal measures and techniques for quantifying these correlations are topics of current investigation. The biological implications of these correlations are deep, and they themselves remain unresolved. For example, statistical dependencies between exons carrying protein coding sequences and noncoding introns may indicate the existence of as-yet unknown error correction mechanisms or structural scaffolds. Thus motivated, we propose to develop precise and reliable methodologies for quantifying and identifying such dependencies, based on the information-theoretic notion of mutual information.

Mutual information has already been employed in several applications for finding regions of statistical dependence in biological sequences: For example, comparative analysis of the 5' untranslated regions of DNA coding sequences in [14], which resulted in the detection of a novel translational initiation signal for eukaryotic genes; estimation of the average mutual information content of the core promoter regions from different organisms to verify the importance of TATA-boxes and transcriptional initiation [15], [16]; and patterns of sequence conservation at the 3' untranslated regions of orthologous genes from human, mouse, and rat genomes reported in [19]. Statistical approaches to estimation of mutual information from gene expression datasets have been investigated in [20]. Reliable predictions of protein secondary structures based on long-range dependencies may enhance functional characterizations of proteins [2], so analyses that incorporate mutual information estimates may also provide more accurate predictions in the area of protein engineering.

In this work we focus on developing reliable information-theoretic methods for determining whether two biosequences are likely to be statistically dependent. Another motivating factor, which is more closely related to ideas from information theory, is the question of determining whether there are error correction mechanisms built into large molecules, as argued by Battail; see, e.g., [3]. We choose to work with protein coding, well conserved exons and non-coding introns which are parts of genes with greater variability. They are dispersed on strings of biopolymers and still they have to be precisely identified in order to produce biologically relevant information. It seems that there is no external source of information, but the structure of RNA molecules themselves to generate functional templates for protein synthesis. Determining potential mutual relationships between exons and introns may justify additional search for still unknown factors affecting RNA processing.

The complexity and importance of the RNA processing system is emphasized by the largely unexplained mechanisms of alternative splicing, which provide a source of substantial diversity in gene products. The same sequence may be recognized as an exon or an intron, depending on a broader context of splicing reactions. The information that is required for the selection of a particular segment of RNA molecules is very likely embedded into either exons or introns, or both. It seems that the splicing outcome is determined by structural information carried by RNA molecules themselves, unless the fundamental dogma of biology (the unidirectional flow of information from DNA to proteins) is to be questioned.
Finally, the constant evolution of genomes introduces certain polymorphisms, such as tandem repeats, which are an important component of genetic profiling applications. We also study these forms of statistical dependencies in biological sequences using mutual information.

II. THEORETICAL BACKGROUND

In this section we outline the theoretical basis for the mutual information estimators applied to biological sequences.

Suppose we have two strings, \( X^n_1 = (X_1, X_2, \ldots, X_n) \) and \( Y_{1}^{M} = (Y_1, Y_2, \ldots, Y_{M}) \) taking values in a common finite alphabet \( A \). In most of our experiments, \( M \) is significantly larger than \( n \); typical values of interest are \( n \approx 50 \) and \( M \approx 200 \). Our main goal is to determine whether or not there is some form of statistical dependence between them. Specifically, we assume that the string \( X^n_1 \) consists of independent and identically distributed (i.i.d) random variables with common distribution \( P \) on \( A \), and that the random variables \( Y_i \) are also i.i.d. with a possibly different distribution \( Q \). Let \( \{W(y|x)\} \) be a family of conditional distributions, or "channel," with the property that, when the input distribution is \( P \), the output has distribution \( Q \), that is, \( \sum_x P(x)W(y|x) = Q(y) \), for all \( y \).

To differentiate between the following two scenarios:

(I) Independence. \( X^n_1 \) and \( Y^n_1 \) are independent.

(II) Dependence. First \( X^n_1 \) is generated. Then an index \( J \in \{1, 2, \ldots, M\} \) is chosen in an arbitrary way, and \( Y_{j+1}^{j+n-1} \) is generated as the output of the discrete memoryless channel \( W \) with input \( X^n_1 \). Finally the rest of the \( Y_i's \) are generated i.i.d. according to \( Q \).

To distinguish between scenarios (I) and (II) we compute the empirical per-symbol mutual information between the two strings. Recall that if \( X \) and \( Y \) are two random variables taking values in \( A \), where \( X \) has distribution \( P \), \( Y \) has distribution \( Q \), and the conditional distribution of \( Y \), given \( X = x \) is \( W(y|x) \), the mutual information between \( X \) and \( Y \) is,

\[
I(X;Y) = \sum_{x,y} P(x)W(y|x) \log \frac{P(x)W(y|x)}{P(x)Q(y)},
\]

where logarithms are taken to base 2, \( \log = \log_2 \).

For each \( j = 1, 2, \ldots, M - n + 1 \), let \( \hat{p}_j(x,y) \) denote the joint empirical distribution of \( (X^n_1, Y_{j+1}^{j+n-1}) \), and let \( \hat{p}(x) \) and \( \hat{q}(y) \) denote the empirical distributions of \( X^n_1 \) and \( Y_{j+1}^{j+n-1} \), respectively. The empirical (per-symbol) mutual information \( \hat{I}_j(n) \) between \( X^n_1 \) and \( Y_{j+1}^{j+n-1} \) is defined as,

\[
\hat{I}_j(n) = \sum_{x,y} \hat{p}_j(x,y) \log \frac{\hat{p}_j(x,y)}{\hat{p}(x)\hat{q}(y)},
\]

where the two random hypotheses. To see this, observe that we can simply rewrite,

\[
\hat{I}_j(n) = \frac{1}{n} \sum_{x,y} \sum_{i=1}^{n} \hat{p}_j(x,y) \log \frac{\hat{p}_j(x,y)}{\hat{p}(x)\hat{q}(y)}
\]

\[
\approx \frac{1}{n} \sum_{i=1}^{n} \log \frac{\hat{p}_j(X_{i}, Y_{j+i-1})}{\hat{p}(X_i)\hat{q}(Y_{j+i-1})}
\]

where, for any probability distribution \( p \) on \( A \), \( p^n \) denotes the induced product distribution on strings in \( A^n \).

To avoid the trivial case where both scenarios are identical, we assume that the rows of \( W \) are not all equal to \( Q \) so that in the second scenario \( X^n_1 \) and \( Y_{j+1}^{j+n-1} \) are actually not independent.

θ for some \( j \), declare that the strings \( X^n_1 \) and \( Y_{j+1}^{j+n-1} \) are dependent; otherwise, declare that they are independent.

Before examining the issue of selecting the value of the threshold \( \theta \), we note that the statistic we consider is identical to the (normalized) log-likelihood ratio between the above two hypotheses. To see this, observe that we can simply rewrite,
\( \sigma^2 \) is given by, \( \sigma^2 = \text{Var} \left( \log \frac{W(Y|X)}{Q(Y)} \right) \). Therefore, for any fixed threshold \( \theta < I \), the probability of error here satisfies,

\[
P_{e,2} = \Pr \{ \text{declare independence} | W \text{-dependent strings} \} \\
= \Pr \{ I(n) \leq \theta | W \text{-dependent strings} \} \\
\approx \Pr \{ V \leq \theta - I \sqrt{n} \} \approx \Theta \left( \exp \left\{ -\frac{(I-\theta^2)}{2\sigma^2 n} \right\} \right).
\]

Thus, \( P_{e,2} \) also decays exponentially.

c) Discussion: First note that, in order for both probabilities of error to decay to zero for large \( n \), the threshold \( \theta \) needs to be strictly between 0 and \( I \). For that, we need to have some prior information about the value of \( I \), i.e., of the level of dependence we are looking for. If the value of \( I \) were actually known and a fixed threshold \( \theta \in (0, I) \) was chosen independent of \( n \), then both probabilities of error would decay exponentially fast, but with typically very different exponents:

\[
P_{e,1} \approx \exp \left\{ -(\theta \ln 2) n \right\} \quad \text{and} \quad P_{e,2} \approx \exp \left\{ -\left( \frac{1-\theta^2}{\sqrt{2\sigma^2}} \right) n \right\}.
\]

Clearly, balancing the two exponents also requires knowledge of the value of \( \sigma^2 \) in the case when the two strings are dependent, which, in turn, requires full knowledge of the marginal distribution \( P \) and the channel \( W \). Of course this is unreasonable, a-priori, since we cannot specify in advance the exact kind and level of dependence we are actually trying to detect in the data.

A more practical approach is as follows: Since the probability of error of the first kind \( P_{e,1} \) only depends on \( \theta \) (at least for large \( n \)), and since in practice declaring false positives is much more undesirable than overlooking potential dependence, in our experiments we decide on an acceptably small false-positive probability \( \epsilon \), and then select \( \theta \) based on the above approximation by setting \( P_{e,1} \approx \epsilon \).

### III. Experimental Results

In this section we apply the mutual information test described above to biological data.

A. Detecting DNA Sequence Dependencies

All of our experiments were performed on the maize zmSRp32 gene [6]. The gene zmSRp32 is coded by 4735 nucleotides and has four alternative splicing variants. Two of these are due to splicing of this gene, between positions 1–369 and 3243–4220, respectively, as shown in Figure 1. The results given here are primarily from experiments on these segments of zmSRp32.

In order to understand and quantify the amount of correlation between different parts of this gene, we computed the mutual information between all functional elements including exons, introns, and the 5' untranslated region. As before, we denote the shorter sequence of length \( n \) by \( X^n \) and the longer one of length \( M \) by \( Y^n \). We apply the simple mutual information estimator \( I_j(n) \) defined in (1) to estimate the mutual information between \( X^n \) and \( Y^n \) for each \( j = 1, 2, \ldots, M - n + 1 \), and we plot the “dependency graph” of \( I_j = I_j(n) \) versus \( j \); see Figure 2. The threshold \( \theta \) is computed according to (2), by setting \( \epsilon \), the probability of false positives, equal to 0.001; it is represented by a (red) straight horizontal line in the figures.

In order to “amplify” the effects of regions of potential dependency in various segments of the zmSRp32 gene, we computed the mutual information estimates \( I_j \) on the original strings over the regular four-letter alphabet \{A, C, G, T\}, as well as on transformed versions of the strings where pairs of letters were grouped together, using either the grouping \{AT, CG\} or \{AG, CT\}. In our results we observed that such groupings are often helpful in identifying dependency, as clearly illustrated in Figure 2.

Our early experiments strongly suggested that there is significant dependence between the bases in positions 1–369 and certain substrings of the bases in positions 3243–4220. Therefore, we focused our attention on these two segments. While the 1–369 region contains the 5' untranslated sequences, an intron, and the first protein coding exon, the 3243–4220 sequence encodes an intron that undergoes alternative splicing. After narrowing down the mutual information calculations to the 5' untranslated region (5'UTR) in positions 1–78 and the 5'UTR intron in positions 78–268, we found that the initially identified dependency was still present; see Figure 2. A close inspection of the resulting graphs indicates that the dependency is restricted to the alternative exons embedded into the intron sequences, in positions 3688–3800 and 3884–4254.

These findings suggest that there might be a deeper connection between the 5'UTR DNA sequences and the DNA sequences that undergo alternative splicing. The UTRs are multifunctional genetic elements that control gene expression by determining mRNA stability and efficiency of mRNA translation. Like in the zmSRp32 maize gene, they can provide multiple alternatively spliced variants for more complex regulation of mRNA translation [9]. They also contain a number of regulatory motifs that may affect many aspects of mRNA metabolism. Our observations can therefore be interpreted as suggesting that the maize zmSRp32 5'UTR contains information that could be utilized in the process of alternative splicing, yet another important aspect of mRNA metabolism. The fact that the value of the empirical mutual information between 5'UTR and the DNA sequences that encode alternatively spliced elements is significantly greater than zero clearly points in that direction.

B. Application to Tandem Repeats

Here we further explore the utility of the mutual information statistic, and we examine its performance on the problem
of detecting Short Tandem Repeats (STRs) in genomic sequences. STRs, usually found in non-coding regions, are made of back-to-back approximate repetitions of a pattern which is at least two bases long and generally shorter than 15 bases. The period of an STR is defined as the length of the pattern in it. Some STRs are polymorphic; the number of repetitions at any specific locus varies significantly among individuals. These properties make polymorphic STRs suitable tools for determining genetic profiles, and this has become a prevalent method in forensic investigations.

Several algorithms have been proposed for detecting STRs in long DNA strings [4], [10], [17] with no prior knowledge about the size and the pattern of repetition. These algorithms are mostly based on pattern matching, and they all have high time-complexity. Here we propose a statistical approach using an adaptation of the method described in the previous sections.

In the U.S., the FBI has decided on 13 loci to be used as the basis for genetic profile analysis, and they continue to be the standard in this area. To demonstrate how our approach can be used for STR detection, we chose to use sequences from the FBI’s Combined DNA Index System (CODIS). The SE33 locus contained in the GenBank sequence V00481, and a modified (as explained below) version of VWA locus contained in the GenBank sequence M25858 (see http://www.cstl.nist.gov/biotech/strbase/ for details).

Let $Y_1^M$ denote the DNA sequence in which we are looking for STRs. The gist of our approach is simply to choose a periodic probe sequence of length $n$, say, $X_1^n$ (typically much shorter than $Y_1^M$), and then to calculate the empirical mutual information $\hat{I}_j = \hat{I}_j(n)$ between $X_1^n$ and each of its possible alignments with $Y_1^M$. In order to detect the presence of STRs, the values of the empirical mutual information in regions where STRs do appear should be significantly larger than zero, where “significantly” means larger than the corresponding estimates in ordinary DNA fragments containing no STRs. Obviously, the results will depend heavily on the exact form of the probe sequence. Therefore, it is critical to decide on the method for selecting: (a) the length, and (b) the exact contents of $X_1^n$. The length of $X_1^n$ is crucial; if it is too short, then $X_1^n$ itself is likely to appear often in $Y_1^M$, producing many large values of the empirical mutual information and making it hard to distinguish between STRs and ordinary sequences. Moreover, in that case there is little hope that the analysis of the previous section (which was carried out of long sequences $X_1^n$) will provide useful estimates for the probability of error. If, on the other hand, $X_1^n$ is too long, then any alignment of the probe $X_1^n$ with $Y_1^M$ will likely also contain too many irrelevant base pairs. This will produce negligibly small mutual information estimates, again making impossible to detect STRs. These considerations are illustrated by the results in Figure 3. As a result of our experiments with CODIS sequences, we have found out that the ideal length of probe $X_1^n$ for detecting STRs in a given DNA sequence should be between 80–100 base pairs which is, not surprisingly, the average sequence length in CODIS.

As for the contents of the probe sequence $X_1^n$, the best choice would be to take a segment $X_1^n$ containing an exact match to an STR present in $Y_1^M$. But in most of the interesting applications, this is of course unavailable to us. A “second best” choice might be a sequence $X_1^n$ that contains a segment of the same “pattern” as the STR present in $Y_1^M$, where we say that two sequences have the same pattern if each one can be obtained from the other via a permutation of the letters in the alphabet; cf. [1], [13]. For example, if $X_1^n$ contains the exact same pattern as the periodic part of the STR to

Fig. 3. Dependency graph of the GenBank sequence $Y_1^M = V00481$, for a probe sequence $X_1^n$ which is a repetition of $AGGT$, of length: (a) 12, or (b) 60. The sequence $Y_1^M$ contains STRs that are repetitions of the pattern $AAAG$, in the following regions: (i) there is a repetition of $AAAG$ between bases 62–108; (ii) $AAAG$ is intervened by $AG$ and $AAGG$ until base 138; (iii) again between 138–294 there are repetitions of $AAAG$, some of which are modified by insertions and substitutions. In (a) our probe is too short, and it is almost impossible to distinguish the SE33 locus from the rest. However, in (b) the location SE33 is singled out by the two big peaks in the mutual information estimates; the shorter peak between the two larger ones is due to the interventions described above.
The proposed methodology is very effective at detecting the presence of STRs of interest. Due to a very noisy STR again with a 4 base period, in (a), the probe \( \tilde{X}_1^m \) produces the opposite result: The artificial STR is clearly identified, but there is no indication of the STR present at the VWA locus.

In the dependency graph, for example, by feeding the relevant parts separately into one of the standard string matching-based tandem repeat algorithms. Thus, our method can serve as an initial filtering step which, combined with an exact pattern matching algorithm, provides a very accurate and efficient method for the identification of STRs.

In terms of its practical implementation, note that our approach has a linear running time \( O(n) \), where \( n \) is the length of \( Y^M \). The empirical mutual information of course needs to be evaluated for every possible alignment of \( Y^M \) and \( X_1^m \), with each such calculation done in \( O(n) \) steps, where \( n \) is the length of \( X_1^m \). But \( n \) is typically no longer than a few hundred bases, and, at least to first order, it can be considered constant. Also, repeating this process for all possible repeat periods does not affect the complexity of our method by much, since the number of such periods is quite small and can also be considered to be constant.

**ACKNOWLEDGMENT**

The work of W. Szpankowski was supported in part by the NSF Grants CCF-0513636 and DMS-0503742, NIH Grant R01 GM068959-01, and AFOSR Grant 073071.

**REFERENCES**


