Juliette Zerick (jnzerick@<natural habitat>)
Department of Mathematics, U.C. Davis
PHY256B ~ Spring 2010

Exons and You

[0]

Abstract

The problem of locating coding regions in DNA sequences—*de novo* exon prediction—is an old open problem in theoretical biology. Finding a solution is still attractive, however, since accurate predictors could lead to rapid discovery of new genes. Many attempts at solving the problem have been made, but contradictory results and inaccurate classifiers have been the result. I conjecture that given real-world constraints, *de novo* exon prediction is impossible: no sequence-based model can fully capture the dynamic of the transcription process. As a simple experiment, I created models of exonic and intronic data; assuming the models were correct, the results would suggest that exons and introns contain equally noisy, random data. Such a conclusion we know is invalid; biological systems appear haphazard while we do not yet understand their complexity. Thus the appearance of "noise" more likely indicates faulty assumptions imposed by the model, or variation that the model cannot capture.

Introduction

Motivation

The genomes of many species have been sequenced, and as a result, biologists are drowning in data: the raw genomic data requires annotation. Annotating a genome (associating certain properties with regions of the DNA sequence) is a difficult process. Transcription factors might be located by identifying motifs (frequently-occurring patterns in the sequence); if we can sequence a known protein, the source coding regions in the DNA might be found; and so on.

The more general problem of locating expressed coding regions, or exons, unsurprisingly remains open. In prokaryotes, the DNA sequence only contains exons. In fact, prokaryote DNA often has overlapping exons; the "code reuse" presumably helps with keeping the "code base" small enough to fit in the cell. In eukaryotes, the process of transcription is not performed in a straight shot. In the DNA sequence, exons may be fragmented, interspersed with introns (non-coding regions). During transcription, introns are spliced out of the RNA.

If we can locate exons in the sequence, and then confirm the presence of a transcribed and translated protein in the cell, we might be able to find new genes. Such a method could lead to incredible new advances in the biological sciences.

Punchline

Finding a solution for the exon prediction problem is attractive for many reasons. First, gene prediction could be sped up, or even automated. But second, the problem is highly complex, which makes it even more interesting.

Many solutions have been attempted, but none fully capture the dynamic in the transcription process. I conjecture that by ignoring the nuance of the problem, exon prediction is impossible. I created simple models of good exonic and intronic data; assuming the models were correct, the results would suggest that exons and introns contain equally noisy, random data. As we know this is certainly not the case, the appearance of "noise" more likely indicates faulty assumptions imposed by the model, or variation that the model cannot capture.

Background

The Central Dogma of Molecular Biology (illustrated in Figure 1), in a nutshell, states that from the DNA, RNA is transcribed, and RNA is translated into protein. This process of course is extremely complicated, involving many specialized molecules. RNA polymerase must open the DNA, and then create the complementary strip of RNA. The RNA then must be transported through the nucleus membrane, and into the cell. From there, ribosomes latch onto the RNA, and translate it into protein with amino acids in the cell.

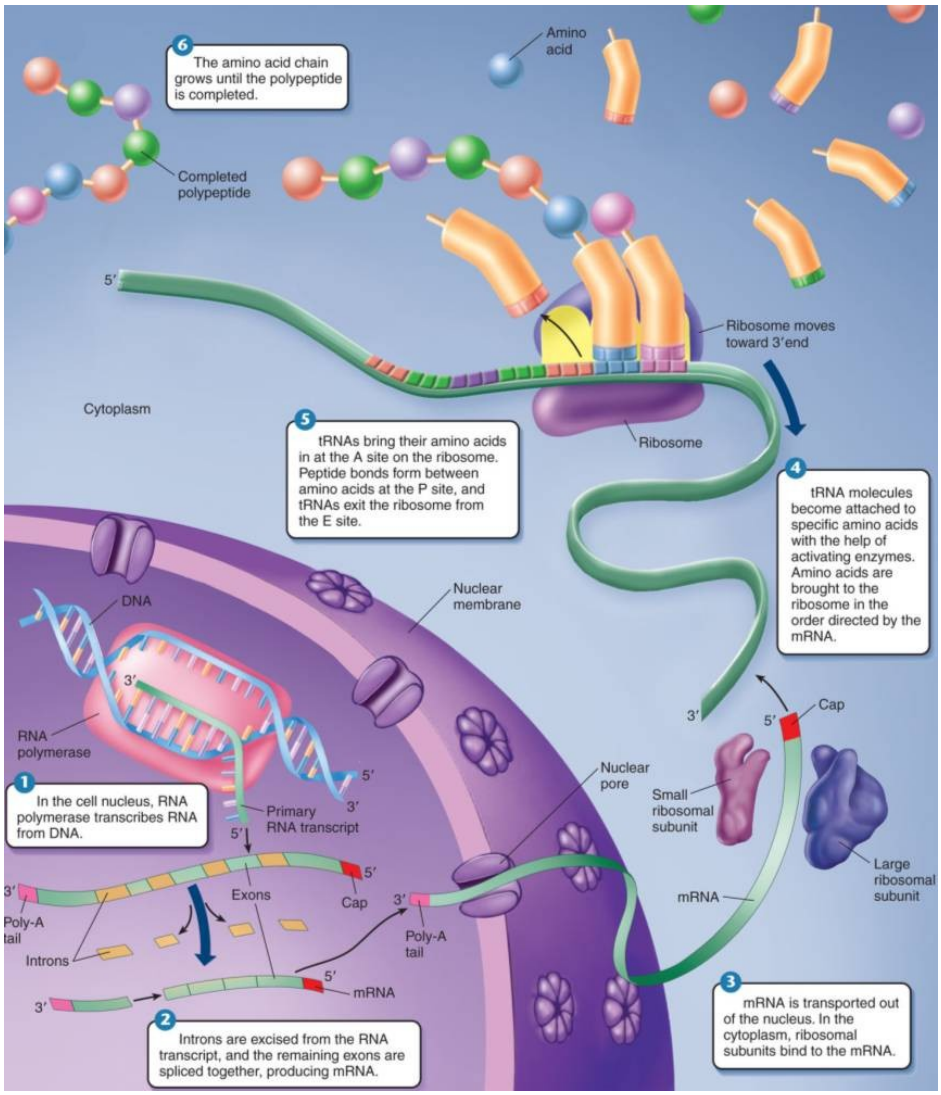

Figure 1: Above is an illustration of the Central Dogma of Molecular Biology. RNA is transcribed from DNA, and proteins are translated from the RNA [1].

From DNA sequence to protein, there are many subprocesses of great interest. Coding theorists consider translation as a decoding process; mutations are effectively decoding errors. Physicists model the action of RNA polymerase on strands of DNA; ripping apart the strand is a rather violent process, and deformations induced by stress can affect where the DNA might open up. The problem of *de novo* exon prediction (prediction based on the DNA sequence alone) is of course focused on the transcription process.

Many attempts have been made to solve this problem. Starting in the mid-1980s, information theorists tried to compare the "information content" of introns and exons, with conflicting conclusions. In [2], it was found that the coding and non-coding regions contained different amounts of information; in [3], the authors the opposite.

In the early 1990s, signal processing theorists interpreted the DNA sequence as a signal. [4] was the major result: it was found that there are long-term correlations found in very long sequences of DNA.

From the 1990s onward, computational biologists have tried using hidden Markov models (HMM) to

predict exon locations, using roughly the same technique: a HMM is constructed from a training set, and its accuracy validated based on its performance on another dataset. While the developers of AUGUSTUS [5], GeneParser [6], and others claimed reasonable accuracy, a review of popular models [7] found that the rate of successfully predicting the locations of exons was on the order of a (fair) coin toss.

Dynamical System

I am interested in the problem of exon prediction. By the Central Dogma of Molecular Biology, it is assumed that from DNA, RNA is transcribed. In eukaryotes, the RNA is encoded from exons in the DNA, while introns are spliced out.

Aside from the Central Dogma, there are no agreed-upon abstract models of the transcription process. Since we must first assume a model of the process, the attempts to find exons have varied greatly (along with their rates of accuracy). As a result, the "equations of motion" that describe the process are still theoretical, and the subject of much debate. I assume that the DNA sequence, along with environmental factors in the cell, determines the location of coding regions, but exact interations are unknown.

Conjecture

I conjecture that any computationally realistic model of the transcription process cannot accurately predict exons *de novo*. In theory, the DNA encodes every protein available in the cell, so there should be sufficient information in the raw sequence to locate exons. In practice, we do not have the computational resources to create a perfect HMM—not to mention the fact that completely annotated data would be necessary for its construction, thus rendering the exon prediction problem solved prior to the building of the model.

I expect that a feasible solution would have to incorporate information external to the sequence, such as environmental conditions in the nucleus, the physical stresses on strands, etc.

Methods

Knowing how many failing attempts existed, I hesitated to construct any model of the transcription process. The model reflects the assumptions made of the nature of the data; at worst, it *imposes* structure on the process that does not exist. I would like to learn as much as I can about the transcription process and proper modeling techniques, but since I had good data, I figured I might as well see what would happen.

My data was a subset of regions annotated in the ENCODE Pilot Project [8]. Over a multi-year period, biologists annotated particular regions of the human genome as best as possible. The end result was extremely good data available for theorists.

My dataset consisted of a subset of [9]. A rendering of the dataset and its annotations is shown in Figure 2.

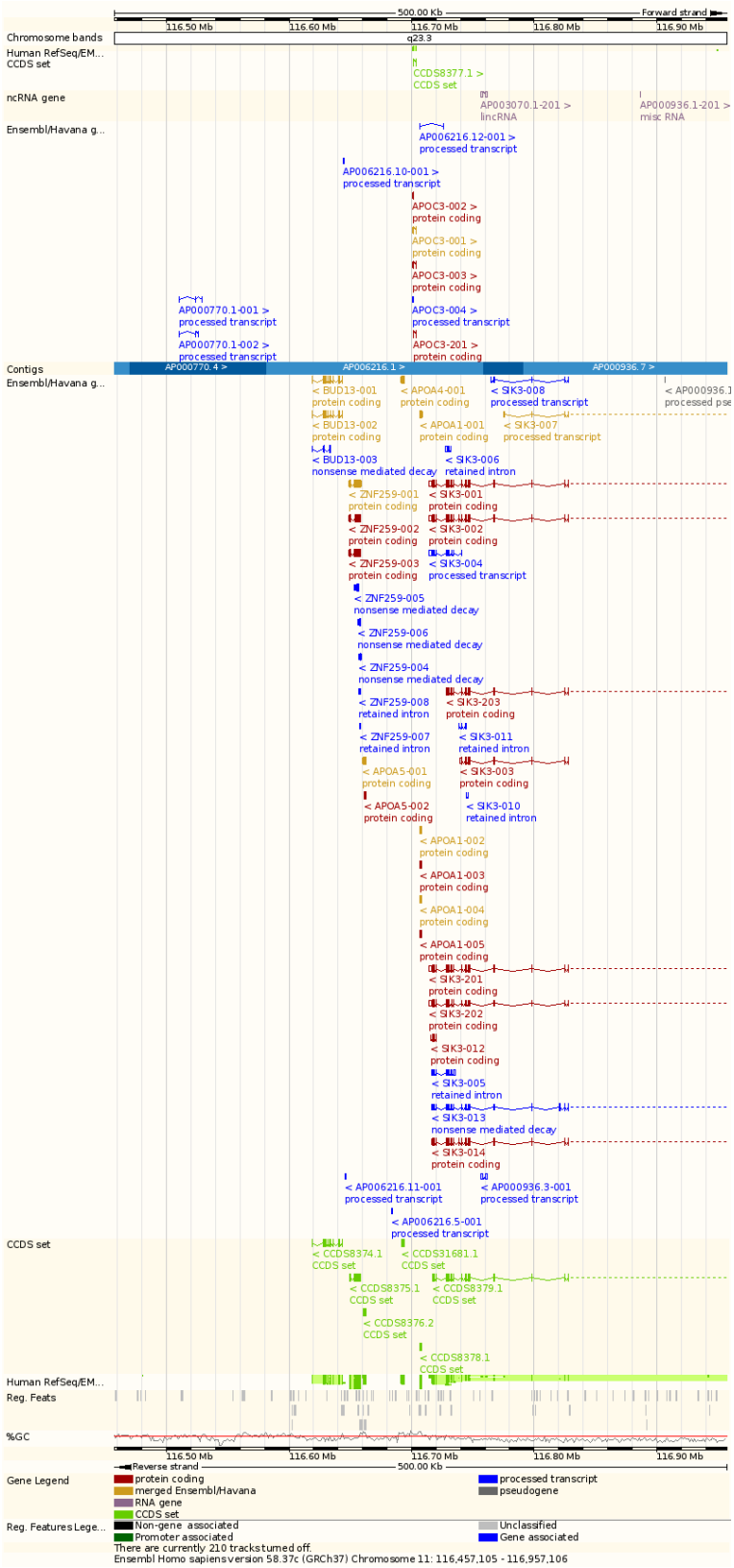

Figure 2: Via the Ensembl Genome Browser, the (many) annotations of the test dataset are depicted as subintervals of a subsequence of Human Chromosome 11.

I had recently examined the algorithm used in NestedMICA [11], a motif-finding program. Rather than searching for possible matches to length-n motifs (which is NP-complete), the authors created a model of the sequence that did not contain motifs (the "uninteresting" parts). Thus if a motif was encountered in a sequence, it would be classified as not being "uninteresting," and thus be identified as an "interesting" motif. I was impressed with the simplicity of such a clever approach, and implemented a similar technique when constructing the model. I first parsed the data, splicing out exons and introns, yielding two datasets: exonic and intronic.

For each dataset, I considered the raw sequence as a base-4 number, mapping nucleotides to binary numbers. I then constructed a depth-5 parse tree, ran the sequence through, and inferred an epsilon-machine.

The scripts used to automate retrieval of data were a mixture of Perl and Python. Perl is more frequently used in bioinformatics, but I have found Python the superior tool for scientific data analysis.

Results

The results were unsurprising. Both datasets generated epsilon-machines similar to a noisy fair coin process (see Figure 3).

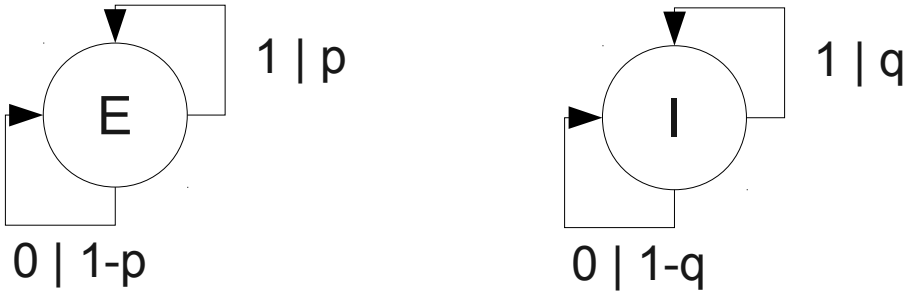

Figure 3: On the left is the recovered epsilon machine of the exonic data, where $p \approx \frac{1}{2}$ (alternatively, $p = \frac{1}{2} + n_p$, *where* $n_p > 0$). On the right is a similar epsilon machine generated by the intronic data with, likewise, $p \approx \frac{1}{2}$.

For complexity measures, $h_\mu = 1$, **E** = 0, and $C_\mu = 0$. The crypticity of the process (given this model) is estimated at $\chi = 0$.

Discussion

Some variation in the models was to be expected, but the "noise" is more likely indicative of underlying structure that is ignored. Since the construction of a HMM in some sense forces the data to hew to a simpler model, small but significant variations will appear as noise, but we know better than to assume that transcription behaves so predictably (or erratically, depending on the viewpoint).

Conclusion

The transcription process is extremely complicated, and still not fully understood. As a result, the exon prediction problem requires far more care than building classification schemes based on erroneous assumptions regarding the behavior of the process.

Even from well-annotated data, a simple HMM cannot encode the nuance of the differences between exonic and intronic sequences. What appears to be noise in a fair coin-like process is more likely significant variation that *cannot* be captured by the model. Thus a better model is required before a classification scheme is even considered.

Exon prediction is an old problem, and finding a solution—even rudimentary—would greatly help advance the field of biology. Its difficulty, and the need for a cautious approach, has been noted for over twenty years [12, 13]. But its complexity is the source of its attraction, and I hope to study the problem further.

References

[0] beaker.jpg . Available online: <http://lockheart.co.uk/bb/beaker.jpg>.

[1] 13_10.jpg . Biology 1100: Survey of Biology, College of DuPage. Available online: <http://bio1100.nicerweb.com/Locked/media/ch13/13_10.jpg >.

[2] Robert Farber, Alan Lapedes (1992). **Determination of eukaryotic protein coding regions using neural networks and information theory**. *Journal of Molecular Biology*, pp. 471-479 (226).

[3] R. Michael Stephens and Thomas Dana Schneider. **Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites**. *Journal of Molecular Biology*, pp. 1124-1136 (228).

[4] Wentian Li, Thomas G. Marr, Kunihiko Kaneko (1994). **Understanding long-range correlations in DNA sequences**. *Physica D*, pp. 392-416 (75).

[5] Mario Stanke and Stephen Waack (2003). **Gene prediction with a hidden Markov model and a new intron submodel**. *Bioinformatics*, pp. 215-225 (19).

[6] Eric E. Snyder and Gary D. Stormo (1995). **Identification of protein coding regions in genomic DNA**. *Journal of Molecular Biology*, pp. 1-18 (248).

[7] Keith Knapp and Yi-Ping Phoebe Chen (2006). **An evaluation of contemporary hidden Markov model genefinders with a predicted exon taxonomy**. *Nucleic Acids Research*, pp. 317-324 (35).

[8] National Human Genome Research Institute. **ENCODE Pilot Project**. ENCODE Project: ENcyclopedia Of DNA Elements. Information available online: <http://www.genome.gov/10005107>.

[9] UCSC Genome Browser, February 2009 build: Chromosome 11: 1-135,006,516. ENCODE Project. Available online: <http://genome.ucsc.edu/cgi-bin/hgTracks?hgsid=162157541&position=chr11>.

[10] Ensembl Genome Browser: Chromosome 11: 116,457,105-116,957,106. Available online: <http://uswest.ensembl.org/Homo_sapiens/Location/View?db=core;r=11:116457105-116957106>.

[11] Mutlu Doğruel, Thomas A. Down, and Tim J.P. Hubbard (2008). **NestedMICA as an ab initio protein motif discovery tool**. *Bioinformatics*, (9:19).

[12] J.M. Barry (1986). **Informational DNA: a useful concept?**. *Trends in Biochemical Sciences*, pp.

317-318.

[13] Peter R. Scibbald, Satindranath Banerjee, and Jack Maze (1989). **Calculating higher-order DNA sequence information measures**. *Journal of Theoretical Biology*, pp. 475-483 (136).