Chapter 14

Applying Word-Based Algorithms: The IMEter

Ian F. Korf and Alan B. Rose

Abstract

Important patterns can be found in strings of characters such as nucleotides in a DNA sequence by examining the frequency of occurrence of specific character combinations or words. The abundance of words can reveal the presence of underlying trends governing the order of characters, even if the biological reasons for those trends remain mysterious. As an example of one way in which word frequencies have provided insight, we describe the IMEter, a word-based algorithm for analyzing introns and their effect on gene expression. The IMEter demonstrates that introns located near the beginning of genes are compositionally distinct from later introns and that these differences are closely related to the ability of some introns to increase gene expression. This word-based approach has proven more successful than deletion analysis at identifying the sequences responsible for elevating expression because they are dispersed throughout stimulatory introns.

Key words: Markov model, word based, nucleotide frequency, odds ratios, intron, gene expression, motif, intron-mediated enhancement, IMEter.

1. Sequences as Markov Models

Whether the medium is conversation, chalkboards, journal articles, or computers, biological sequences are often represented as text. While we know that DNA, RNA, and protein molecules are dynamic entities with physical and chemical properties that depend on their shape and environment, representing biological sequences as one-dimensional *strings* is convenient and allows one to take advantage of analysis techniques pioneered in Cryptography and Natural Language Processing.

One of the simplest analyses one can perform with any text is to count the frequencies of individual symbols. **Table 14.1** shows the frequency of each letter in Darwin's Origin of Species. Not

Dmitry A. Belostotsky (ed.), *Plant Systems Biology, vol. 553*
© Humana Press, a part of Springer Science+Business Media, LLC 2009
DOI 10.1007/978-1-60327-563-7_14 Springerprotocols.com

Table 14.1
Letter frequencies in Origin of Species

Symbol	%	Symbol	%
A	7.98	N	7.17
B	1.69	O	7.21
C	3.50	P	1.89
D	3.70	Q	0.09
E	13.18	R	6.27
F	2.78	S	6.88
G	1.82	T	9.00
H	4.99	U	2.56
I	7.43	V	1.19
J	0.07	W	1.60
K	0.37	X	0.24
L	4.19	Y	1.64
M	2.51	Z	0.05

surprisingly, the most common letter is “e”, which accounts for approximately 13% of all letters. Similarly, one can count the letters from the *Arabidopsis thaliana* genome and observe that the genome is approximately 36% GC (**Table 14.2**). Although these analyses are very simple, they provide useful information. For example, the letter frequencies in Darwin’s book closely resemble the letter frequencies in English in general, and one could use such information to deduce that Origin of Species was published in

Table 14.2
Genomic nucleotide frequencies

	A. thaliana					*D. radiodurans*
			Intron			
Symbol	Genome	Exon	All	Proximal	Distal	Genome
A	32.00	29.85	26.73	25.7	27.3	16.54
C	18.02	20.14	15.46	14.8	16.2	33.51
G	18.01	20.16	17.16	17.0	15.8	33.45
T	31.97	29.84	40.64	42.5	40.7	16.49

English without ever reading the book. Similarly, one can examine the nucleotide compositions of various genomes and recognize that each organism has a characteristic (though not necessarily unique) composition, and one might use these frequencies to look for horizontal gene transfer or contamination events.

While simple letter counting can be useful, it throws away the context of each letter. Let us first consider the immediate context of a letter, which is defined by adjacent letters. It may seem strange, but in text and sequence analysis, only the preceding letter is used as the context. The reason for this is that we model sequences as the products of Markov processes. A useful way to think of a Markov model (or chain) is as a machine that randomly generates plausible observations. Let us consider a Markov model for the daily weather with three states called *Sunny*, *Cloudy*, and *Rainy*. We must define transition probabilities between the various states that determine how often and in what order the observations are generated. For example, we might define the probability of transitioning between Sunny and Cloudy to be greater than Sunny and Rainy because clouds usually precede a rain storm. **Figure 14.1** shows such a model. If we want the weather model to mimic actual weather patterns, the transition probabilities should match local weather observations. Given the model in **Fig. 14.1**, we can start in a state, such as Sunny and then create a new day of weather by "rolling dice" to determine if tomorrow will be Sunny, Cloudy, or Rainy. It should be obvious that whatever tomorrow's weather is depends only on today, not last year. This property where the future is independent of the distant past is known as the *Markov property*.

Getting back to biology, let us now consider that genomes are products of Markov processes. The simplest Markov model one can make is that each nucleotide is generated without respect to context. That is, we can simply define probabilities for A, C, G, and T and draw these at random to generate a sequence. Let us say we do this for the *A. thaliana* genome and the first three nucleotides

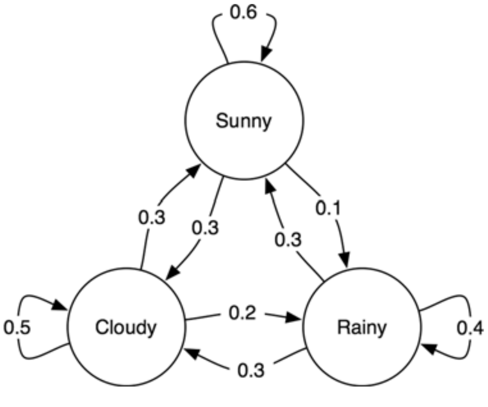

Fig. 14.1. Markov model for daily weather.

generated are GCT. The probability with which this particular sequence was generated by the model is approximately 0.010 (0.18 × 0.18 × 0.32). To take nearby context into account, we must record the conditional probability of each letter given the preceding letter(s). That is, we need to know the probability of generating a T given that we have already seen a T (this is similar to the weather example above). If we are concerned with only a single preceding letter, this is called a first-order Markov chain. If we are concerned with two letters of context, for example to model the probability of generating a G given that we have already seen an A followed by a T, this is a second-order Markov chain. In general terms, biological sequences are often modeled by an *n*th-order Markov chain where *n* is some non-negative integer. In practice, the value of *n* is commonly in the range of 1–5.

The letter frequencies from **Tables 14.1** and **14.2** can now be understood as 0th-order Markov models. They predict that the probability of observing CG is simply the probability of C multiplied by the probability of G. While you might believe this to be true of the *A. thaliana* genome, you certainly would not believe this of English since you have never seen any words where C precedes G (because there are none). The immediate context of a letter is clearly very important in language. Is this also true of biological sequences? **Table 14.3** shows a first-order Markov model for the *A. thaliana* genome. There are some interesting properties. For example, the probabilities of the homodimers (AA, CC, GG, TT) are greater than expected. Also, the probabilities are not symmetric. For example, the probability of C followed by G is not the same as G followed by C. For this reason, when modeling biological sequences, it is generally a good idea to go beyond 0th-order Markov chains. Exactly how far beyond zero depends on several factors. One important consideration is the size of the training set. A 15th-order Markov chain requires approximately 1 billion (4^{15}) contexts. If such a model was applied to the human genome (approximately 3 billion bp), each observation would receive less than one count on average, which does not lead to a very useful model.

Table 14.3
First-order Markov chain from A. thaliana

	Preceding symbol			
Symbol	**A**	**C**	**G**	**T**
A	36.18	35.24	35.58	23.98
C	16.36	18.81	16.65	20.02
G	18.57	12.99	18.74	19.86
T	28.89	32.97	29.03	36.14

In addition to the immediate context of a symbol, there are also larger contexts. Chromosomes do not have uniform compositions throughout their lengths, and a single Markov model does not capture regional variations very well. **Table 14.2** shows nucleotide frequencies for *A. thaliana* exons and introns. The differences are obvious even at the 0th-order level. The compositional difference between coding and non-coding sequences has been used by many researchers to identify protein-coding genes in genomic DNA. Early efforts (1) used second-order Markov models to report regions that were likely to be coding but did not attempt to define exon–intron boundaries. Today, gene-finding programs are much more sophisticated (2) and take into account such biological features as splice sites, promoters, and poly-A sites. Still, at the root of these gene prediction algorithms, the sequence is generally represented as an *n*th-order Markov model.

Previously we noted that GCT would be generated by an *A. thaliana* 0th-order Markov model with a probability of approximately 0.010. If the model had been derived from *Deinococcus radiodurans* (*see* **Table 14.2**), the sequence would be generated with a probability of approximately 0.019 (0.335 × 0.335 × 0.165). If we are given just the sequence GCT and asked to guess its origin, the odds are almost twice as great that it came from *D. radiodurans* than from *A. thaliana*. Similarly, given models for coding and non-coding sequence, it is very simple to examine an unknown sequence (or segment of a sequence) to determine which model is more likely to have generated the sequence. In essence, this is how gene-finding algorithms work, though they use higher order Markov chains.

Probabilities and odds ratios are typically calculated as logarithms in bioinformatics applications. One reason for this is that genome data is large, and if you repeatedly multiply numbers, you may overflow or underflow the numeric representation in the computer (e.g., if you keep squaring a number on a calculator, you eventually reach a limit as the value approaches infinity or zero). The base of the logarithm is generally 2 or *e*, and the corresponding units are *bits* or *nats.* In general, log-odds values are called *scores.* For example, in sequence alignment, the score for any two amino acids is the log-odds ratio of the observed and expected pairings, where the observed pairing is calculated from multiple alignments of related proteins and the expected pairing is the random expectation from individual amino acid frequencies. A high score indicates the amino acids are found more often than by chance. For example, the score for valine and leucine is positive because these are chemically similar amino acids that can often substitute for one another without compromising protein function. In gene finding, the score of a predicted exon reflects the log-odds ratio of having been generated by a coding vs. non-coding model. A positive score indicates the region is probably coding while a negative score indicates it is probably non-coding.

2. Words as Functional Units

Both in language and in sequence analysis, the concept of a *word* is very useful. In text, it is not letters but words that convey information. Similarly, in biology, specific strings such as the amino acids RGD (extra-cellular matrix attachment) or the nucleotides GAATTC (*Eco*RI restriction site) have known functional roles. In both language and sequence, the meaning of a word often depends on its context.

Identifying words in text is trivial if the language uses spaces as delimiters, but in languages without delimiters, such as ancient Roman, one must use context to parse letters into words. The words and meanings in biology are much more difficult to parse than natural languages for several reasons: (a) there are only four letters in DNA; (b) there are no delimiters or punctuation; (c) there is no dictionary of legal words; (d) we do not know the rules of the language; and (e) words can often be misspelled and retain their biological meaning.

In sequence analysis we use a simplified definition of *word* that does not require knowing its meaning (function). A word is simply a sequence of length *k* where *k* is some positive integer. Words are also called *k*-tuples or *k*-mers or even oligos. To identify all the words in a sequence, one simply moves a window of length *k* one letter at a time along the entire sequence (without making truncated words at either end). Words created this way are therefore very similar to *n*th-order Markov models.

Like letter frequencies, word frequencies can be informative in text and sequence. **Table 14.4** shows the top 20 words from the Origin of Species. As you might expect, *the* tops the list and *species* is quite common. If we construct a quasi-genomic version of the book by removing all the spaces and punctuation, the text becomes one long chromosome-like string that is difficult to interpret.

malsonecatforinstancetakingtocatchratsanothermiceonecata
ccordingtomrstjohnbringinghomewingedgameanotherhareso
rrabbitsandanotherhuntingonmarshygroundandalmostnigh
tlycatchingwoodcocksorsnipesthetendencytocatchratsrathertha
nmiceisknowntobeinheritednowifanyslightinnatechangeofha
bitorofstructurebenefitedanindividualwolfitwouldhavethebe
stchanceofsurvivingandofleavingoffspringsomeofitsyoungwo
uldprobablyinheritthesamehabitsorstructureandbytherepeti
tionofthisprocessanewvarietymightbeformedwhichwouldeith
ersupplantorcoexistwiththeparentformofwolforagainthewol
vesinhabitingamountai

Since we no longer know where the word boundaries are, we can use the sequence analysis concept of a word to identify all possible words of some size *k* and determine their frequencies.

Table 14.4
Word frequency analysis of Origin of Species

Rank	Word	3-mer	4-mer	5-mer	6-mer	7-mer	8-mer
1	the	the	tion	ofthe	specie	species	differen
2	of	and	nthe	ation	pecies	thesame	havebeen
3	and	ion	ofth	speci	softhe	thatthe	especies
4	in	ing	fthe	pecie	hesame	natural	selectio
5	to	tio	thes	ecies	thesam	differe	election
6	a	nth	ther	inthe	thatth	ifferen	varietie
7	that	ent	that	which	differ	havebee	arieties
8	as	tha	othe	tions	hatthe	avebeen	naturals
9	have	oft	atio	other	ations	especie	lselecti
10	be	her	spec	ction	ofthes	lection	alselect
11	is	sof	peci	onthe	natura	selecti	characte
12	on	fth	have	natur	atural	electio	haracter
13	species	hes	inth	softh	ection	varieti	aturalse
14	by	for	cies	atthe	tionof	arietie	turalsel
15	which	ati	ecie	andth	genera	rieties	uralsele
16	or	int	hich	thesa	especi	fromthe	ralselec
17	are	ere	whic	tothe	eofthe	animals	distinct
18	it	hat	sand	esame	iffere	aturals	ifferent
19	for	oth	tthe	hesam	havebe	lselect	conditio
20	with	ies	with	their	fferen	ication	ondition

Table 14.4 shows that even without knowing the true word boundaries, it is possible to identify common words and phrases. Now imagine if the text was written in a different language and was interspersed with lots of seemingly random letters. This is the problem we have to deal with in analyzing biological sequences.

Even though genomes are complex entities and our knowledge of genome biology is still in its infancy, it is possible to make significant advances using methods as simple as word frequency analysis. As an example, we describe our research on how introns affect gene expression. While our work utilized the *A. thaliana* genome (3), and this work would have been more difficult without the entire sequence, the sine qua non was not the genome, but rather knowing what to model. In other words, the key was understanding the biology.

3. The Biology of Intron-Mediated Enhancement

Shortly after introns were discovered, it was noted that several genes were expressed very poorly when their introns were removed. Conversely, inserting introns into reporter genes that lacked them, including bacterial genes such as lacZ or GUS, often increased the expression of those genes in transgenic organisms. Most of the introns characterized could only affect expression from within transcribed sequences and in their natural orientation, indicating that they operated by a still undefined mechanism that is distinct from transcriptional enhancer elements. This intron-mediated enhancement (IME) (4) has been observed in a broad diversity of organisms including mammals, fungi, nematodes, insects, and plants, suggesting that it is an ancient and fundamental feature of eukaryotic gene regulation.

A puzzle arose from attempts to identify the sequences within introns that are responsible for elevating expression. Some efficiently spliced introns clearly stimulate expression much more than others do, suggesting the presence of enhancing sequences within some introns. However, motifs that are conserved between stimulatory introns could not be found by conventional homology searches due to the large heterogeneity in intron sequence and size coupled with the relatively small number of introns known to stimulate expression. Attempts to locate enhancing regions by deletion analysis also have failed because no unique sequences are individually required for the intron to increase mRNA accumulation (5). That is, introns that contain large internal deletions but are still spliced usually stimulate mRNA accumulation as much as does the full-length intron, even when the combined deletions span the entire intron. How can differences between introns be sequence based if no unique sequences are involved? One possible explanation is that the enhancing sequences are redundant and distributed throughout introns. This idea was confirmed using hybrid introns constructed from parts of enhancing and non-enhancing introns (6). The dispersed nature of the expression-affecting sequences contrasts with more familiar regulatory elements such as enhancers, promoters, RNA secondary structures such as stem loops, binding sites for proteins or small RNAs, or aptamers, whose functions depend on discrete and localized individual sequences.

Three lines of evidence suggest that introns near the start of their genes are more likely than other introns to stimulate expression. First, virtually all of the introns known to boost expression are first introns, although not all first introns have this ability. Second, introns that elevate expression when located in the 5′-UTR lose this effect when they are moved to the 3′-UTR. Third, an enhancing intron that is placed progressively farther downstream in a gene starts to lose its enhancing ability at about 500 bp from the promoter and has no effect ~1 kb or more from the 5′ end (7).

Defining the sequences required for IME is desirable because it would provide a toehold for the biochemical isolation of trans-acting factors that bind to those sequences, which could be an important path to understanding the novel mechanism through which introns affect expression. In addition, identifying the enhancing sequences would provide a means to predict whether or not an untested intron is likely to elevate expression. Eventually, with the knowledge of what sequences enhance expression, it may be possible to design synthetic introns that are more powerful than any naturally occurring one, which would be very useful for transgenic applications seeking to maximize gene expression.

4. The IMEter

The key to modeling IME is the hypothesis that IME signals are enriched in introns located near the promoter (proximal) compared to those farther down the transcript (distal). Starting with simple letter counting, we can look at the sequence composition of proximal and distal introns. **Table 14.2** shows the single letter frequencies where proximal introns are defined as those that begin within 500 bp of the promoter and distal introns are defined as those that begin more than 500 bp from the promoter. The fact that the compositions are not identical suggests that there may be important differences between proximal and distal introns. Even if the compositions were identical, however, there may be higher order differences not visible at the 0th order.

Our program for predicting IME, the IMEter (6), is very similar to algorithms for predicting coding regions. Instead of comparing arbitrary genomic regions to the compositions of coding and non-coding sequences, we compare arbitrary introns to the compositions of proximal and distal introns. A positive score indicates an intron is more similar to proximal introns than distal introns and is therefore more likely to contain elements responsible for IME. Since we do not consider the splice donor and acceptor sequences to be generated from the same model as the body of the intron, the IMEter omits these regions. Before training the IMEter, we must choose several parameters: (a) the word size, (b) the cutoff for proximal introns, (c) the cutoff for distal introns, (d) the length of the splice donor site to omit, and (e) the length of the splice acceptor site to omit. For simplicity, the cutoff for both proximal and distal can be a single value, such as the 500 bp we used for **Table 14.2**.

The IMEter scoring function can be described by the following equation:

$$S = \sum_{i=1+D}^{i \leq L-K-A} \log\left(\frac{P_{w_i}}{Q_{w_i}}\right)$$

where S is the IMEter score, L is the length of the intron, K is the word size, A is the length of the splice acceptor consensus, D is the length of the splice donor consensus, w_{i} is a word of length K at position i, and P and Q are frequency distributions for words of length K in proximal and distal introns.

Training the IMEter requires a set of genes where the position of the 5′ end of the transcript and the positions of introns are known. Fortunately, the *Arabidopsis* genome annotation contains thousands of experimentally identified examples due to the efforts of the full-length cDNA sequencing project (8). While we utilized thousands of genes, we find that it is also possible to train the IMEter on a few hundred conserved genes. As part of our training procedure, we removed highly paralogous genes (to limit overtraining on large gene families) and those genes with suspect features (e.g., very short or GC-rich introns) that may indicate genome annotation errors.

To test whether IMEter scores have biologically meaningful values, we trained the IMEter with an educated guess at the parameters (word size 5, proximal/distal cutoff at 400 bp, 5 bp donor site, 10 bp acceptor site) and examined the scores of introns whose effect on gene expression was already known. The only data set appropriate for this analysis comes from experiments in *Arabidopsis* where the enhancing ability of different introns has been tested with the same reporter gene in single-copy lines. Even though the quantitative data set was small, representing just six introns, it was the largest known for any organism. Furthermore, the data are very reproducible, as indicated by the small amount of variation in expression, presumably because only single-copy transgenic plants were analyzed. When IMEter scores were compared to the expression values, a very strong linear correlation was found between an intron's IMEter score and the degree to which that intron stimulates mRNA accumulation (**Fig. 14.2** filled circles). The tight correlation suggested that IMEter scores might be able to predict the enhancing ability of previously uncharacterized introns. To test this, six additional introns were chosen, and all enhanced expression to the degree expected from their IMEter scores (**Fig. 14.2** open

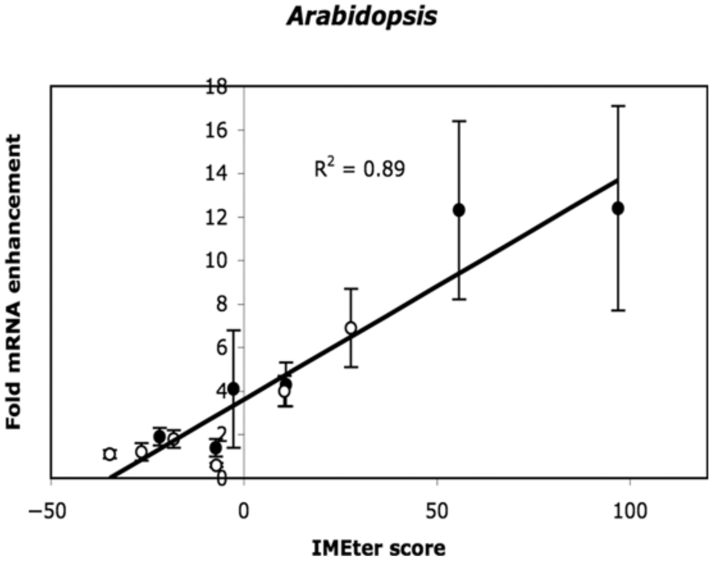

Fig. 14.2. IMEter score is correlated with enhancement.
Copyright The American Society of Plant Biologists and reproduced with permission.

circles). Further evidence supporting the connection between IMEter scores and enhancing ability comes from the 21 other *Arabidopsis* introns reported in the literature to boost expression of different genes. All but one of these introns have a positive IMEter score, and 18 have scores in the top 5% of all *Arabidopsis* introns.

The IMEter can be optimized by changing various parameters. **Figure 14.3** shows how variations in word size and the proximal/distal cutoff affect performance as measured by the R^2 value. A word size of 1, corresponding to a 0th-order Markov model, is not very useful. Larger word sizes perform much better, but as the word size gets over 8, the R^2 value drops off. This is especially apparent when the proximal/distal cutoff is low. This is probably due to the smaller amount of sequence available and the larger number of words. In *Arabidopsis*, a variety of parameter combinations perform approximately equivalently. This may not be true in other genomes or other sequence analysis scenarios, so it is a good idea to survey the parameter space as we have done.

The observation that promoter-proximal and distal introns gave different k-mer profiles indicated that introns are structurally unequal depending on the location of those introns in their genes. To explore genome-wide differences in intron composition, the entire set of *Arabidopsis* introns was randomly divided into two equal groups. The introns in one group were used to train the IMEter, which was then used to analyze the introns in the other. The distribution frequency of IMEter scores forms a bell-shaped curve centered near zero. When only the first introns from genes are considered, the distribution shifts to the right (mean score = 10.6),

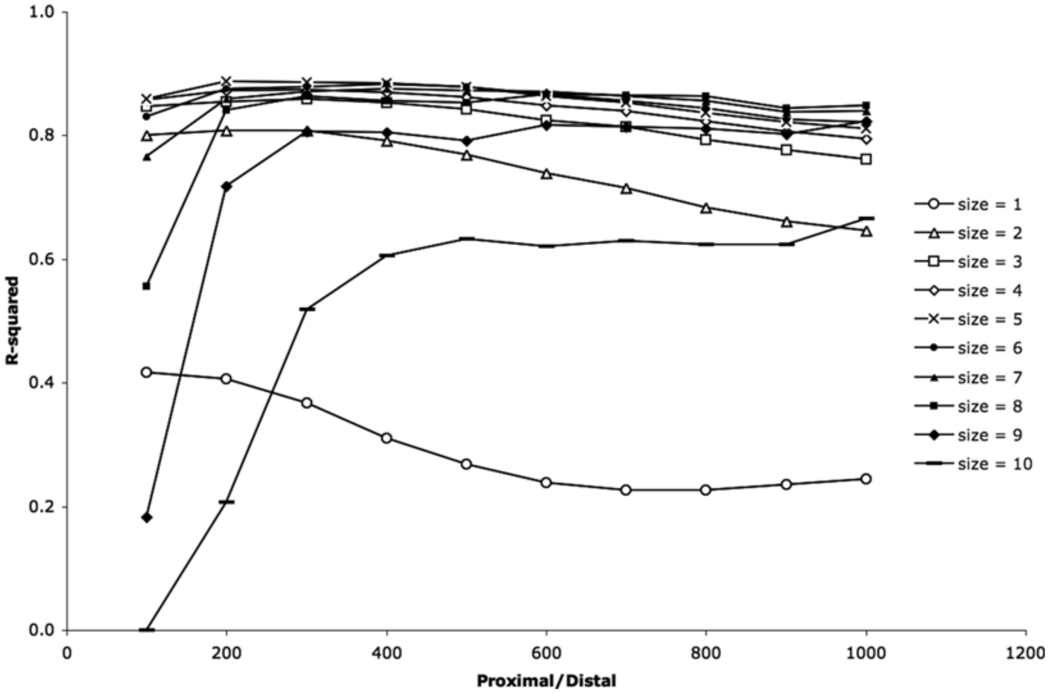

Fig. 14.3. Optimizing IMEter parameters.
Copyright The American Society of Plant Biologists and reproduced with permission.

and virtually all of the highest-scoring introns in the genome are first introns. The relationship between IMEter scores and location can be seen more clearly by plotting the scores of introns against their distance from the start of transcription (**Fig. 14.4**). Average

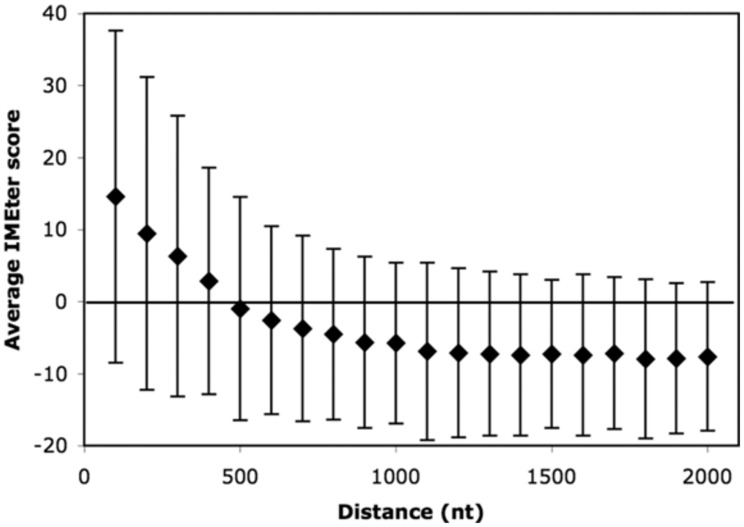

Fig. 14.4. IMEter score of introns as a function of distance from their promoter.
Copyright The American Society of Plant Biologists and reproduced with permission.

IMEter scores are highest in introns near the start and decline with distance, and very few introns more than 1000 nt from the start have a positive score. This pattern is in striking agreement with the ability of an intron to stimulate mRNA accumulation, which also declines with distance from the promoter until it is lost entirely between 550 and 1100 nt from the start of transcription.

4.1. Identifying Enhancing Sequences

One drawback of analyzing word frequencies is that the biological signals that give rise to high scores are not immediately apparent. To identify candidate sequences involved in IME, we employed a motif-finding algorithm, NestedMica (9), to find sequence patterns that are over-represented in the 100 introns with the highest IMEter scores. Several motifs were found, and these were ranked by how well they correlated with the set of introns with known effects on expression. This analysis was therefore very similar to that shown in **Fig. 14.2**, except that a combined motif score was used in place of the IMEter score. The motif that was most correlated with observed enhancement is shown in **Fig. 14.5**, and we call this the IME motif.

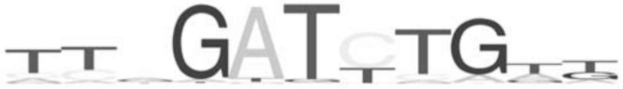

Fig. 14.5. IME motif.
Copyright The American Society of Plant Biologists and reproduced with permission.

Rather than evaluating an entire intron, one can also look for regions of high IMEter score in genomic context by calculating IMEter score in a fixed, sliding window. **Figure 14.6** shows that high scores are most abundant in the intron and occur in the same regions as high IME motif density. While there is a great deal of variation from gene to gene, the general pattern is for IME signals to be concentrated in proximal introns and virtually absent from other regions of the genome.

4.2. IME Signals in Other Species

The IMEter can be applied as described to any organism where there are known exon–intron structures for a few hundred genes or more. Unfortunately, there are no organisms aside from *Arabidopsis* where the IMEter can be quantitatively evaluated because rigorous intron-swapping experiments have not been performed elsewhere. It remains to be established whether or not promoter proximity is relevant to IME in all organisms. Furthermore, the IMEter may be ineffective in species in which introns are very large in size (as in mammals) or small in number (as in *Saccharomyces cerevisiae*). Given the aforementioned caveats, we have examined

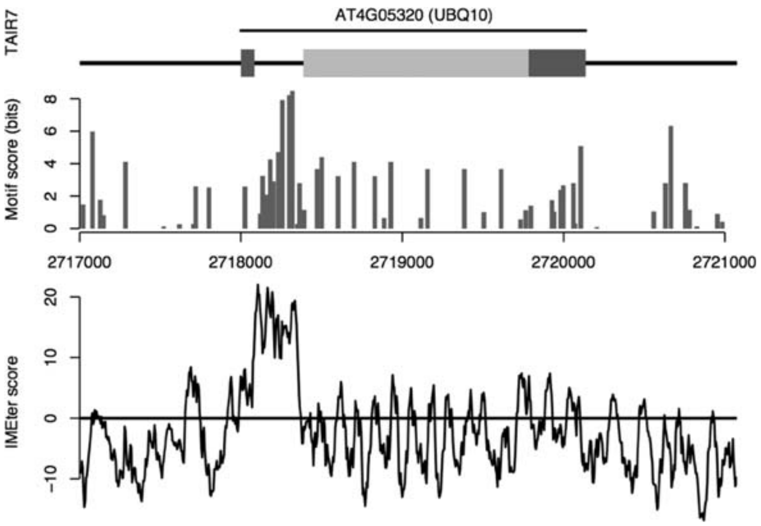

Fig. 14.6. IMEter score and IME motif density in the UBQ10 region. The genomic region of the UBQ10 gene is shown. The exon–intron structure is shown at the top. The dark regions are untranslated, and the light regions are coding. The middle panel shows the score of the IME motif. Higher bars indicate a better match to the consensus. The lower panel shows IMEter score in a 50 bp window.
Copyright The American Society of Plant Biologists and reproduced with permission.

IME signals in rice and find that a rice IMEter behaves similarly to the *Arabidopsis* IMEter (data not shown, see (6)). We also find that there is a good correlation (R^2 0.74) between rice-trained IMEter scores and *Arabidopsis* expression values, which indicates that the IME machinery may be very similar in these organisms.

5. Summary

The IMEter illustrates some of the strengths and weaknesses of word-based algorithms. On the positive side, the IMEter revealed previously unsuspected differences in the composition of introns, a large collection of very diverse elements. No prior assumptions about the length or positions of the relevant sequences were required. However, word frequency analysis is not the appropriate method for all sequence elements. IME signals are both dispersed and redundant, so there was a good fit between the biological signals and the statistical model. If we had been looking for an isolated signal where position was an important factor, for example the TATA box, one would not expect word frequency analysis to be very useful.

Perhaps the most serious weakness of word-based analyses is the difficulty in identifying the functional elements that are being recognized from among the entire dictionary of words. A number of statistical measures can be employed, but ultimately the biological significance of any candidates must be evaluated experimentally. Despite the inherent limitations in word-based analyses, they can be very useful tools for the systems biologist because they provide a means to detect previously unrecognized patterns in complex sets of data, thereby revealing new connections. While it is expected that more sophisticated statistical models (e.g., hidden Markov models) and experimental molecular biology (e.g., gene expression studies, proteomics) are required to identify the biological entities involved, word-based analyses can provide a critical first step for the journey ahead.

References

1. Staden, R. and McLachlan, A.D. (1982) Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res.* **10**, 141–156.
2. Brent, M.R. (2008) Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat. Rev. Genet.* **9**, 62–73.
3. Arabidopsis thaliana Consortium (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature.* **408**, 796–815.
4. Mascarenhas, D., Mettler, I.J., Pierce, D.A., and Lowe, H.W. (1990) Intron-mediated enhancement of heterologous gene expression in maize. *Plant Mol. Biol.* **15**, 913–920.
5. Rose, A.B. and Beliakoff, J.A. (2000) Intron-mediated enhancement of gene expression independent of unique intron sequences and splicing. *Plant Physiol.* **122**, 535–542.
6. Rose, A.B., Elfersi, T., Parra, G., and Korf, I. (2008) Promoter-proximal introns in Arabidopsis thaliana are enriched in dispersed signals that elevate gene expression. *Plant Cell.* **20**, 543–551.
7. Rose, A.B. (2004) The effect of intron location on intron-mediated enhancement of gene expression in Arabidopsis. *Plant J.* **40**, 744–751.
8. Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., Muramatsu, M., Hayashizaki, Y., Kawai, J., Carninci, P., Itoh, M., Ishii, Y., Arakawa, T., Shibata, K., Shinagawa, A., and Shinozaki, K. (2002) Functional annotation of a full-length Arabidopsis cDNA collection. *Science.* **296**, 141–145.
9. Down, T.A. and T.J. Hubbard (2005) NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res.* **33**, 1445–1453.