A Quantitative Approach to Investigating the Hypothesis of Prokaryotic Intron Loss

Robert Sinclair

Mathematical Biology Unit

Okinawa Institute of Science and Technology

Okinawa 904-0412

Japan

sinclair@oist.jp

Keywords: Intron evolution, intron loss, nucleomorph, Bacteria, Archaea, Thermotogae, Legionellales

Abstract: Using a novel method, we show that ordered triplets of motifs usually associated with spliceosomal intron recognition are underrepresented in the protein coding sequence of complete Thermotogae, archaeal and bacterial genomes. The underrepresentation observed does not extend to the noncoding strand, suggesting that the cause of the asymmetry is related to mRNA rather than DNA. Our data do not suggest that the underrepresentation is due to gene transfer from eukaryotes. We speculate that one possible explanation for these observations is that the protein coding sequence of Thermotogae, Archaea and Bacteria was at some time in the past subjected to selection against certain motifs appearing in an order which might initiate splicing in environments harboring a functional spliceosome. This is consistent with, but certainly does not prove, a hypothetical scenario in which at least some prokaryote lineages once possessed a functional spliceosome. Thus, we present a new quantitative method, observations obtained using the method, and a speculative discussion of a possible explanation of the observations.

Introduction

The origin of spliceosomal introns has been a matter of debate for some time [1-17]. A review of this debate is beyond the scope of this paper, as would be any attempt at justifying our pragmatic decision to use the term "prokaryote" here [18]. Like many others, we focus on introns of the type removed by the major spliceosome [19], but also consider minor spliceosomal [17] and some self-splicing introns [9]. Our contribution is twofold. First, we introduce a novel, quantitative method of analysis, which is designed to detect traces of current or prior spliceosomal activity in complete genomes or chromosomes. We demonstrate the potential of the method by first applying it to two cryptophyte nucleomorph genomes [20], one of which has spliceosomal introns, the other of which has lost both spliceosomal introns and the spliceosome itself [21]. Second, we apply our method to complete Thermotogae [22,23], archaeal [24] and bacterial genomes, showing that these do indeed show signs consistent with the hypothesis of prokaryotic intron loss, for spliceosomal introns of both major and minor type [19,25]. We do not exclude all alternative hypotheses, but do show that horizontal transfer of genes from eukaryotes to prokaryotes, know to be rare [26], is unlikely to be the explanation for our observations.

On the most abstract level, one can view our method as a black box that can, to a certain limited extent, recognize spliceosomal introns with only the Thermotogae, archaeal or bacterial genomes as biological input. See Tables 3, 4 and 5. The fact, that this is possible at all, suggests that Thermotogae, archaeal and bacterial genomes encode information sufficient to enable such a recognition. Thus, even when our method is divorced from the intron loss scenario which motivated its development, it still provides indirect support for the hypothesis of exposure to an active spliceosome in the coding sequence of complete Thermotogae,

archaeal and bacterial genomes. In a broad sense, our data suggest that Thermotogae, Archaea and Bacteria "know something" about introns, and we wonder whether this apparent knowledge could in fact be a memory [27].

Let us now describe the line of thought which led to our method. The central principle behind our work is this: Intronless genes must, by definition, avoid being spliced, whenever an active spliceosome is present.

Given that the 5', branch point and 3' splice site motifs potentially define an intron sequence, one way in which an intronless gene might avoid the activity of a spliceosome is to refrain from presenting these motifs in the order associated with splicing. Thus, one expects that these motifs would be underrepresented in what we will call the canonical order, with respect to the same motifs in the reverse order. Since the spliceosome acts only on mRNA, one would only expect to see such an underrepresentation on the coding strand.

The idea of using the order of sequence units is of course not new. Gene order [28] continues to be used as a phylogenetic marker for the simple reason that it is expected to be relatively stable over evolutionarily relevant periods. Since we make no attempt to reconstruct any phylogeny, the known practical difficulty of estimating evolutionary distance using gene order alone [29] does not impact our specific use of motif order as a relatively stable marker. The observations, that splice site motifs appear to be most strongly conserved in eukaryotes with nearly-complete intron loss [30,31], and that short introns appear to be spliced via an "intron definition" mechanism [32], suggest that splice site motifs may be a suitable set of features to use in the investigation of the prokaryotes' hypothetical loss of introns.

It should also be noted that the idea of RNA selection pressure, which we invoke when we state that we expect no underrepresentation on the non-coding strand, has also been used before in discussions concerning the evolution of splicing fidelity and alternative splicing in eukaryotes [33,34].

In the literature, one finds discussions of the hypothesis that prokaryotes may once have possessed active spliceosomes but then lost their introns [2,14-16], via reverse transcription [35] or deletion [36], followed by the eventual loss of the genes required for construction of the spliceosome, since it would no longer have had any purpose. Such scenarios imply that essentially all of the genes in the genome in question would have been intronless at a time when a spliceosome was still active. Therefore, one might expect that there was a time when prokaryotic genes were under selection to avoid presenting splice site motifs in the canonical order. With the hypothesized loss of the spliceosome, this selective pressure would have ceased, and the underrepresentation, a direct result of avoidance, would have decayed with time. In this paper, we in fact demonstrate an underrepresentation in completely sequenced genomes of Thermotogae, Archaea and Bacteria. Due to the statistical nature of our method of analysis, we do not expect individual genes to be informative and therefore focus on complete genomes.

It is only natural to ask whether various selective forces, such as codon usage bias or selection for a certain level of GC content, will lead to the loss of memory, at the DNA sequence level, of ancient events like the putative existence of a spliceosome. In order to test the robustness of our results, we have used random codon reassignments, not respecting original genomic GC content, in an attempt to provoke such a loss of memory. The positive

result of this experiment, for the full bacterial data set, is presented in Figure 1, suggesting that our analysis is robust enough to capture an ancient signal.

The question of which motifs would be appropriate to use in a search for evidence of prokaryotic genes' exposure to an active spliceosome is not a simple one, particularly for major spliceosomal introns. Many of the motifs recognized in intron-poor eukaryotes may be derived or secondary [37]. Minor spliceosomal introns do at least appear to possess well-conserved 5' splice site and branch point sequences wherever they are found in eukaryotic genomes [17]. We take the point of view that both the available prokaryotic genomic data and motifs known from eukaryotes should be taken into account. Thus, we have taken a diverse subset of the known eukaryotic motifs as a starting point, and have let the results of analysis of prokaryotic genomes guide us. From a philosophical point of view, we are therefore in danger of using circular argumentation. Any photograph of a new species potentially suffers from this problem if the photographer adjusted the focus of the camera in taking it, so one has to expect philosophical problems of this sort when more direct evidence is unavailable, as in our case. Our response is to provide means of falsification and to formulate testable predictions of our approach wherever possible. We do this in the Results and Discussion Section below.

Meaning of the Tables

It will be useful to give a brief, informal description of the method, to allow the reader to understand the tables without reading the detail of the Materials and Methods Section. The basis of the analysis is counting matches to patterns composed of three motifs (one can imagine GT, A and AG here, for the sake of discussion) with variable spacings between them, both in the order they are given and also in the reverse order (which would be AG, A

and GT). This counting is performed on the coding sequence of each protein coding gene. If there are more matches to the given order, the gene is considered to show a bias towards the given order. If there are more matches to the reverse order, the gene is considered to show a bias towards the reverse order. A chromosome in which more genes show a bias towards the given order than the reverse order is considered to show a bias towards the given order. A chromosome in which more genes show a bias towards the reverse order than the given order is considered to show avoidance of the given order. The extent of avoidance is given a numerical value by taking ratios of these numbers of chromosomes. The significance of avoidance is quantified in terms of a one-sided P-value, computed with respect to shufflings of nucleotide content within each gene. In the tables, the numbers of chromosomes showing bias towards or avoidance of the given order of motifs are presented as well as the one-sided P-value. A star, indicating significance, appears whenever this P-value is less than 2.5%.

In many cases, an analysis has also been performed using the inverse complements (which would be AC, T and CT) of the given motifs. These are marked as "inv.cpl." in the tables. Matching these constitutes an analysis of the other, non-coding strand. What is important is that we expect to see a lack of significance in these cases.

As described above, this analysis gives equal weight to all genes. The intention has been to be conservative. We do not know for certain if longer genes are more informative than short ones for our purposes. Since it is however natural to ask whether this weighting introduces an unintended bias, we have performed some calculations, using the full bacterial data set, in which genes were weighted according to their coding sequence length. These are described in the Results and Discussion Section. They suggest that our results are not dependent upon the use of equal weighting for all genes.

We have also given equal weight to all chromosomes. Since prokaryotic genomes encoded on multiple chromosomes are relatively rare, we have not performed specific tests for any bias. We do believe that this issue is best understood in the broader context of the biased phylogenetic distribution of sequenced prokaryotic genomes [38].

Genomes

The nucleomorph genomes of the cryptophytes *Guillardia theta* [39] and *Hemiselmis andersenii* [21] provide us with a pair of eukaryotic genomes which, although not closely related [21,40], are the current best data set for studying complete intron loss. The *Guillardia theta* nucleomorph possesses a few short (42 to 52nt) AT-rich spliceosomal introns with 5' and 3' splice site consensus sequences GTAAGTAT and AG respectively. A branch-point consensus sequence has not been identified, although it is reasonable to assume an adenosine in the intron sequence serves as a branch point [41], and we note that CTAA is the core of a common branch point motif in intron-poor eukaryotes [31]. In order that the motifs not be over-specified, which would result in poor statistics, we chose the motifs GTNNGT, TAA, and TAG or CAG as our 5', branch point and 3' splice site motifs, respectively. For *Guillardia theta*, we searched for underrepresentation corresponding to introns with lengths from 42 to 52nt. The nucleomorph of *Hemiselmis andersenii* possesses neither spliceosomal introns nor a spliceosome. We allowed for the possibility it may have had even shorter introns than the *Guillardia theta* nucleomorph before it lost them, noting that chlorarachniophyte nucleomorphs do have very short introns [42], and so searched for underrepresentation corresponding to intron lengths from 37 to 52nt.

It is thought that Thermotogae have been involved in significant horizontal gene transfer [23,43]. The fact that the overwhelming majority of these transfers have been identified as being within prokaryotes does allow us to treat Thermotogae as prokaryotes, and this is consistent with our aim of quantitatively investigating the hypothesis of prokaryotic intron loss. As a first test of our method, the Thermotogae have the advantage of being in this sense generic and also of being a phylum with a conveniently small number of fully sequenced genomes (eleven) to analyse. In the case of Thermotogae, we first used the motif triplets GTNNGT / TAA / TAG and GTNNGT / TAA / CAG to investigate possible avoidance of eukaryotic major spliceosomal splice site motifs. GTNNGT is our attempt at a balance between too much specificity, which would reduce the statistical strength of the analysis due to too infrequent matchings, and too little specificity, which would include nucleotide patterns a spliceosome may not recognize (see [30], Figure 1 of [44] and Figure 2 of [32] for a variety of examples of known 5' splice site motifs). The choice of TAA for the branch point motif is motivated by the fact that most branch point sequences in intron-poor species studied to date contain TAA rather than the more general TRA (see Table 1 of [31] and also [30]). To investigate possible avoidance of eukaryotic minor spliceosomal intron splice site motifs, we first used GTNNCC / TAA / CAG as well as ATNNCC / TTAA / CAC, making use of the greater conservation of the motifs within eukaryotes [17]. Since many unicellular eukaryotes harbour short to ultrasmall introns [30,42,45,46], we specified intron lengths from 17 to 52nt.

In order to see whether the results reported for Thermotogae could in fact be representative of other prokaryotic lineages, we also applied our method of analysis to the domains Archaea and Bacteria, each as a whole, using the same seed motifs and intron length range as for the Thermotogae. The principal restriction here is in the number of chromosomes (90 for Archaea and 1177 for Bacteria), which increases the computer time required for the analysis.

We have also applied our method of analysis to complete bacterial genomes of the intracellular pathogens of eukaryotes *Legionella pneumophila* [47], *Legionella longbeachae* [48] and *Coxiella burnetii* [49], the last an obligate intracellular acidophile. All of these are reported to have a number of genes similar to eukaryotic genes [50], otherwise rare in Bacteria [26]. 11 genomes (5 *Legionella pneumophila* strains, 1 *Legionella longbeachae* strain and 5 *Coxiella burnetii* strains) are the only complete genomes in the γ-proteobacterial order Legionellales in RefSeq [51] Release 44. We used all of them.

Results and Discussion

The *Guillardia theta* nucleomorph genome showed evidence of underrepresentation for the motifs GTNNGT, TAA and TAG (see Table 1), consistent with avoidance of an active spliceosome using these motifs, which we know to be active.

In contrast to this, there is no evidence of underrepresentation for the motifs GTNNGT, TAA and CAG. This would imply that the nucleomorph spliceosome does not splice introns with these motifs – a testable prediction. The list of introns included in Table 2 of the Supplementary Information of the *Guillardia theta* nucleomorph genome paper [39] does however include an intron with just these motifs in the gene identified as Yrpl24. Chromosome 2 of the *Guillardia theta* nucleomorph contains this intron sequence in its entirety, within the pseudogene for gene rpl24 with the locus tag GTHECHR2056. The same chromosome also encodes another gene rpl24 with locus tag GTHECHR2057, to which the automatic annotation pipeline has assigned the protein id XP_001713328.1. A cDNA sequence (GenBank: EG716478.1) extracted from *Guillardia theta* cells (from total RNA [52]) contains this intron sequence, except for the two initial nucleotides GT of the 5' splice site,

still joined to the remainder of the rpl24 pseudogene (i.e. not spliced) and also most of the tRNA-Arg gene with locus tag GTHECHR2t102. We took what would have been the coding sequence of the pseudogene, removed the putative intron, and searched for an EST bridging the putative intron site, but without success. We interpret all this to mean that the only intron listed in the Supplementary Information of the *Guillardia theta* nucleomorph genome paper with a 3' splice site motif of CAG is not an intron that is spliced by the *Guillardia theta* nucleomorph spliceosome. Thus, we see no available evidence to contradict the claim that the *Guillardia theta* nucleomorph spliceosome does not splice short introns with the motifs GTNNGT, TAA and CAG.

We see no significant signal on the noncoding strand. The lack of significance in this case, most easily seen by noting (in Table 1) that two of the three chromosomes showed neither under- nor overrepresentation of the set of motifs, indicates a lack of underrepresentation of intron-like sequences on the non-coding strands of *Guillardia theta* nucleomorph genes, consistent with our tentative interpretation of the observed underrepresentation being a result of hypothetical selection against splicing only, and therefore only applicable to mRNA.

The *Hemiselmis andersenii* nucleomorph data mirrors that for *Guillardia theta*, and this is important because it strongly suggests that genomes which have lost their spliceosome can carry a trace of its prior activity. As for *Guillardia theta*, we find no significant underrepresentation for the motifs GTNNGT, TAA and CAG. For the motifs GTNNGT, TAA and TAG, we do see significant underrepresentation, once again only on the coding strand. See Table 2. The fact that the observed underrepresentation is weaker for *Hemiselmis andersenii* than for *Guillardia theta* is consistent with a scenario involving a slowly decaying signal following spliceosome loss.

In the case of Thermotogae, we see no significant underrepresentation using the motifs GTNNGT, TAA and TAG. This is reminiscent of the observation that the eukaryote *Trichomonas vaginalis* does not splice introns with a TAG 3' splice site motif, instead sharing a long and required ACTAACACACAG 3' splice site motif with at least one *Giardia intestinalis* intron [53].

Using the motifs GTNNGT, TAA and CAG with Thermotogae genomes, we do see significant underrepresentation on the coding strand but not on the noncoding strand (Table 3). We investigated whether the signal would survive if we specified the branch point sequence more completely, and find that the motifs GTNNGT, CTAA and CAG are also significantly underrepresented on the coding strand, in spite of the expected decrease in the raw numbers of matches (which would typically lead to poorer statistics). We also investigated the effect of changing the nucleotide upstream of the branch point adenosine, since this can vary in eukaryotes [31], but only find significance using the TAA motif. We also (Table 3) observed significant underrepresentation for splicing motifs usually associated with the minor spliceosome: GTNNCC, TAA and CAG [25], and so investigated further, to see whether non-canonical AT-AC termini were also avoided. We did not detect any significant avoidance for such motif triplets, suggesting that a putative Thermotogae spliceosome, if it ever existed, may have been a major spliceosome of a permissive type [54]. These results are useful in that they indicate that our method could in principle detect splicing signals that differ from the ones we see today in intron-poor eukaryotes.

Corresponding results for Archaea are presented in Table 4. One difference is that we now also observe avoidance of the motifs GTNNGT, TGA and CAG, which is more consistent

with the variation observed in intron-poor eukaryotes [31], in particular the common branch point consensus sequence TRA.

In the case of Bacteria, our results are presented in Table 5 and Figure 1. The large number (1177) of chromosomes restricted our analysis, but strong trends can still be discerned. In addition to the types of avoidance seen in Thermotogae and Archaea, we now also see avoidance of non-canonical AT-AC termini of a clear minor spliceosomal type, corresponding to the motifs ATNNCC, TTAA and YAC [17]. This is potentially interesting, since it has been supposed that major spliceosomal introns are ancestral to minor spliceosomal introns [19].

In the bacterial data set, we also detect avoidance of the group II intron-like splice site motifs GTGNG, CTA [55] and AT, which is interesting from an evolutionary point of view [9], even though these introns are not spliceosomal. Given that we are examining only short intron-like sequences, this signal could conceivably be due to a mechanism of ORF-less group II intron repression in bacteria [56,57], but studying this question would be beyond the scope of this paper.

To determine whether our assignment of equal weight to each gene introduces an unwelcome bias into our analysis, we have recomputed the analyses for the motif triples GTNNGT / TAA / CAG and GTNNGT / TAA / TAG with the full bacterial data set, but this time weighting genes according to their coding sequence length in nucleotides. Apparent significant avoidance of these two triples is central to the conclusions of this paper. Our raw results for the former triple are 520 chromosomes showing avoidance of the given order and 657 chromosomes showing a bias for the given order. 2000 shufflings resulted in a one-sided P-value estimate of 0.0005, which is consistent with the earlier estimate (Table 5), based

upon equal gene weighting, of <0.001. In the case of the latter triple, we find the chromosome numbers 478 and 690, respectively, and estimate the one-sided P-value to be <0.01. This is also consistent with the estimate based upon equal gene weighting. We conclude that our results are robust with respect to the choice of weighting.

Figure 1 is important because it demonstrates that the use of motif triplet ordering does provide what one would call a stable phylogenetic marker. First, when using the full bacterial data set and shuffling 20,000 times, the distribution of the ratio (of the number of chromosomes showing avoidance to the number of chromosomes showing either avoidance or bias for the given order of the motifs GTNNGT, TAA and CAG) was centred at 50%. This is the blue peak in Figure 1. One would hope to see this for a large and highly diverse genomic data set. We then took the entire bacterial data set again, randomly reassigning all codons (excepting the start and stop codons) in all genes on all chromosomes 4000 times, performing our analysis anew each time. The distribution of these randomly reassigned genomic data sets is the green peak in Figure 1, which contains the raw genomic result (black vertical line), and remains separate from the blue peak. Our conclusion is that random reassignment of codons, which changes GC content as a side effect, does not destroy the signal we observe.

Horizontal transfer of intronless protein coding genes from eukaryotes to prokaryotes [26,58] could in principle explain our observations. In order to understand whether this explanation is likely to be true, we also analysed the eleven available completed bacterial genomes of intracellular pathogens of eukaryotes in the proteobacterial order Legionellales. *Legionella pneumophilia* is known to harbour a relatively large number of eukaryote-like proteins. A recent, careful study [50] identified 14 eukaryotic-like proteins as having been definitely

acquired from eukaryotes, making up less than 1% of the genome, out of a total of 102 eukaryotic-like proteins, and suggested that gene acquisition from eukaryotes is an ongoing process. If our observations were due to horizontal transfer from eukaryotes to prokaryotes, then these intracellular pathogens of eukaryotes would be expected to show a particularly clear signal of avoidance. Our data, presented in Table 6, however, do not show significant avoidance at all. Thus, the hypothesis, that the avoidance we observe in Archaea and Bacteria, each treated as a whole, may be explained by horizontal transfer of genes from eukaryotes to prokaryotes, would appear to be unlikely.

We have observed that Thermotogae exhibit significant avoidance of intron-like sequences on the coding strand of protein coding genes, but Legionellales do not. It is tempting to interpret this as indicating that Thermotogae would have lost introns more recently than Legionellales, but we do not draw any such conclusion here since thermophiles do appear to have especially low mutation rates in comparison with mesophiles [59], and may thus retain an ancient signals more robustly, whereas intracellular pathogens tend to be subject to unusually intense genome reduction [60], and this may contribute to the erasure of ancient signals.

It is on the basis of all these data that we suggest we may have found evidence in favour (not proof) of the hypothesis of prior exposure to an active spliceosome in Thermotogae, archaeal and bacterial genomes. Note that we have used all available archaeal and bacterial genomes in our analyses of Archaea and Bacteria, respectively, making no attempt to correct for sampling bias [38]. For this reason, we do not claim to have provided evidence here for every archaeon or every bacterium on the level of individual isolates or even phyla (except for the Thermotogae). Rather, we have observed a general, but significant, trend for each domain separately. Although one cannot assume that putative prokaryotic splicing signals would

necessarily be similar to the ones we see today in intron-poor eukaryotes, our data do suggest a degree of similarity.

Our proposition, that we may have evidence consistent with (but not proving) prokaryotic intron loss, could be falsified in at least one way: Any prokaryotic mRNA binding factor using the same, or very similar, motifs to a eukaryotic spliceosome might explain our results without needing to invoke the prior existence of active spliceosomes in prokaryotes. We have conducted a literature search, but are not aware of any viable candidate.

The basic ideas upon which our work is based, (i) that intronless genes must avoid attracting the attention of the spliceosome, and (ii) that this should be reflected in an underrepresentation of intron-like sequences in intronless coding genes, can be tested. On the basis of our analysis of the *Guillardia theta* nucleomorph genome, we predict that the *Guillardia theta* nucleomorph spliceosome is unable to splice introns with a CAG 3' splice site sequence.

Materials and Methods

Complete genomic sequences were downloaded from RefSeq [51] Release 44.

Given three ordered motifs, limits on the lengths of matched sequences, and a set of complete genomes or chromosomes, we perform our analysis in the following way. For practical reasons, we pad shorter motifs with the symbol "N" [61] until all motifs have the same length. Thus, the triplet GTNNGT / TAA / TAG becomes GTNNGT / TAANNN / TAGNNN. For each genome or chromosome, we perform the following operations on every contiguous (intronless) protein coding sequence. In each coding sequence, excepting the start and stop codons, we count the number of matches of the motifs in the given order, and also in the

reverse order. We require at least a one nucleotide gap between motifs, and a total length, including the padding, between the specified limits. For example, the sequence ACTAGTACAGTAACGGTGTAAGTAGATAACTTTTTAGGACT contains exactly one match for each ordering. Each coding sequence is then assigned the value +1, -1 or 0, depending upon whether there were more matches to the given order, more matches to the reverse order, or equal numbers of matches to both orders, respectively. For each genome or chromosome, we sum these values for all contiguous protein coding sequences, giving all genes analyzed equal weight. The genome or chromosome is then assigned a value of +1, -1 or 0, depending upon whether there were more contiguous protein coding sequences with the value +1, more contiguous protein coding sequences with the value -1, or equal numbers of contiguous protein coding sequences with the values +1 and -1, respectively. The type of underrepresentation we are interested in expresses itself in there being an underrepresentation of genomes or chromosomes with the value +1 as compared to those with the value -1. Let G (for "given" order) denote the number of genomes or chromosomes with value +1, and R (for "reverse" order) the number of genomes or chromosomes with the value -1. In Tables 1 to 6, G is provided in the second column, while R is provided in the third column.

To evaluate the significance of any underrepresentation, we compute a one-sided P-value in the following manner, the purpose of which is to compensate for gradients in codon use along protein coding sequences [62]. We perform the same analysis as described above for 20,000 (unless otherwise specified) independently generated shuffled copies of the given genomes or chromosomes, shuffling within each contiguous protein coding sequence as follows: We divide up the subsequence between the start and stop codons into windows of 9nt width, and randomly permute the nucleotides in each window, not allowing in-frame stop codons to be created in the process. In each case, let G' and R' denote the counts corresponding to G and

R as described above, but for the shuffled sequence data. We approximate the one-sided P-value by the ratio LE/(LE+GT), where LE is the number of times, out of the 20,000 shufflings, G'/(G'+R') is less than or equal to G/(G+R), and GT is the number of times it is greater than G/(G+R). Cases in which G+R=0 and/or G'+R'=0 did not occur. Note that small values of this one-sided P-value do mean that there is significant underrepresentation of the motifs in the given order, but large values do not automatically mean that there is significant overrepresentation. We have used P<2.5% as our threshold for significance throughout.

We look for underrepresentation on the noncoding strand by using motifs which are the inverse complements of those used for the coding strand, requiring care with the concepts of given and reverse order. No significant under- or overrepresentations were actually observed on the noncoding strand.

Acknowledgements

I would like to thank Dr. Takayuki Naito of the Molecular Neuroscience Unit at OIST, Japan, and Prof. Byrappa Venkatesh of the Molecular Genetics Lab, Institute of Molecular and Cell Biology, Singapore, both of whom have commented on results and interpretations at various stages in this long-term project, giving valuable constructive criticism and advice. This manuscript is an expanded version of a poster of the same name presented at the Annual Meeting of the Society for Molecular Biology and Evolution in Lyon, France, in July 2010, in which the method was described and applied to the cryptophyte nucleomorph, Thermotogae, Clostridium and Crenarchaeota genomes (from RefSeq 41).

Figure 1: Robustness of phylogenetic signal.

The black vertical line represents the percentage of 1177 bacterial chromosomes which show avoidance of the motifs GTNNGT, TAA and CAG, in the given order, with respect to the reverse ordering. The green peak was obtained by randomly reassigning all codons in all genes 4000 times, modeling the effects of diverse selective forces on codon bias and GC content. The blue peak was obtained by randomly shuffling nucleotides within 9-base-pair windows in all genes 20,000 times, and therefore represents a null model of non-informative gene sequences. What can be seen is that the codon reassignment peak not only contains the actual data, but remains separate from the null model. We conclude that the phylogenetic signal is likely to be robust with respect to the effects of selective forces on codon bias and GC content.

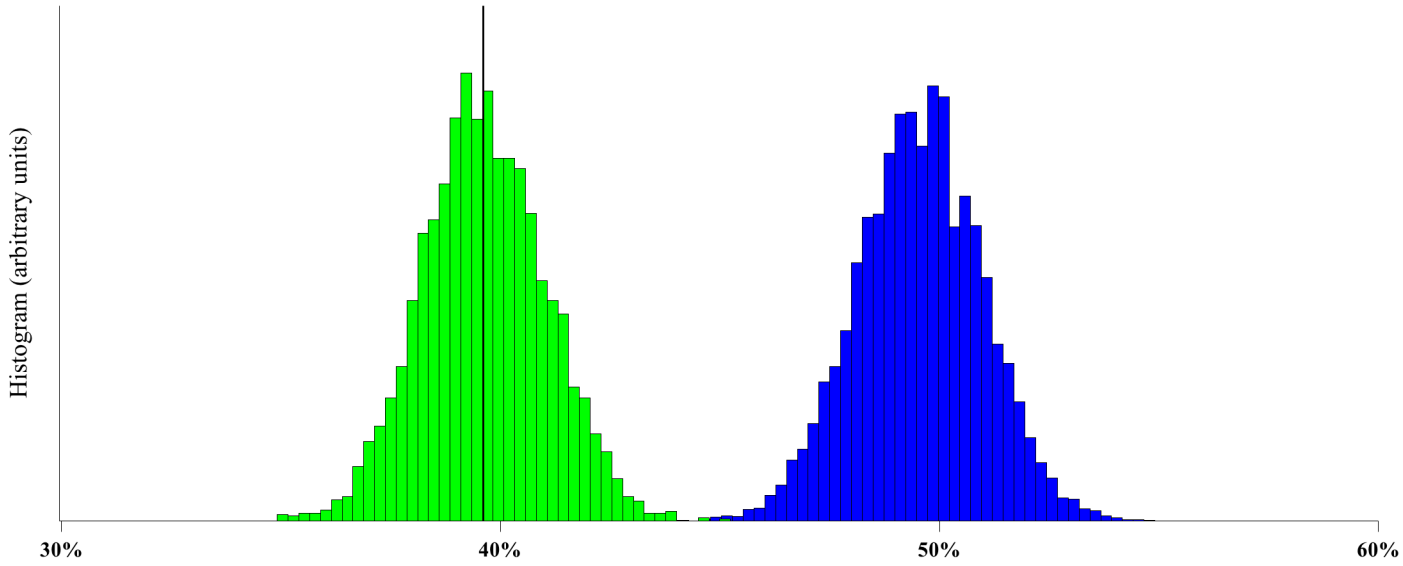
Histogram (arbitrary units)
30%
40%
50%
60%

Table 1: Analysis of *Guillardia theta* nucleomorph chromosomes.

Motifs Given Order	Chromosomes Given Order	Chromosomes Reverse Order	One-Sided P-Value
GTNN**GT** / **TAA** / **TAG** **AC**NN**AC** / **TTA** / **CTA** (inv.cpl.)	1 1	2 0	0.003 * 1
GTNN**GT** / **TAA** / **CAG**	3	0	1
GTNN**CT** / **AGA** / **GAT**	2	1	0.177

This nucleomorph possesses an active spliceosome. Splice site motifs GTNNGT, TAA and TAG are avoided in coding sequences on the coding strand. On the non-coding strand we see neither significant under- nor overrepresentation, as evidenced by the fact that two out of three chromosomes contained equally many genes with bias for or against the given order of inverse complemented motifs. Splice site motifs GTNNGT, TAA and CAG are not avoided in coding sequences. The high one-sided P-value is due to 74% of all shuffled genomic datasets sharing the counts of 3 and 0 chromosomes showing a bias towards the given or reverse order, respectively. Thus, we observe no under- or overrepresentation for the motifs GTNNGT, TAA and CAG. The motifs GTNNCT / AGA / GAT are intended to be a control.

Table 2: Analysis of *Hemiselmis andersenii* nucleomorph chromosomes.

Motifs Given Order	Chromosomes Given Order	Chromosomes Reverse Order	One-Sided P-Value
GTNN**GT / TAA / TAG** **AC**NN**AC / TTA / CTA** (inv. cpl.)	1 2	1 1	0.013 * 0.569
GTNN**GT / TAA / CAG**	2	1	0.246
GTNN**CT / AGA / GAT**	2	1	0.247

This is a eukaryotic genome possessing neither spliceosomal introns nor a spliceosome. We observe avoidance of the splice site motifs GTNNGT, TAA and TAG on the coding strand only, consistent with a memory of intron loss.

Table 3: Analysis of 11 Thermotogae genomes.

Motifs Given order	Genomes Given Order	Genomes Reverse Order	One-sided P-value (n)
GTNN**GT / TAA / CAG**	1	9	0.005 *
ACNN**AC / TTA / CTG** (inv. cpl.)	5	6	0.499
GTNN**GT / TAA / TAG**	7	4	0.941
ACNN**AC / TTA / CTA** (inv. cpl.)	8	3	0.905
GTNN**GT / CTAA / CAG**	2	9	0.006 *
GTNN**GT / TGA / CAG**	7	4	0.689
GTNN**GT / TCA / CAG**	7	4	0.495
GTNN**GT / TTA / CAG**	9	2	0.935
GTNN**GT / TCC / CAG**	5	6	0.045
GNN**TTG / TAA / GCA**	4	6	0.266
GTNN**CC / TAA / CAG**	2	9	0.011 *
ATNN**CC / TAA / CAC**	6	5	0.58 (2000)
ATNN**CC / TAA / TAC**	4	7	0.27 (2000)

The one-sided P-value appears to indicate similarity to intron-like sequences involving the common major spliceosomal splice site motifs GTNNGT, CTAA and CAG, but also suggests permissiveness with respect to the atypical 5' motif GTNNCC. The value of n provided in parentheses is the number of shuffled copies of the genomes used in approximating the P-value, wherever less than 20,000.

Table 4: Analysis of 90 archaeal chromosomes.

Motifs Given order	Genomes Given Order	Genomes Reverse Order	One-sided P-value (n)
GTNN**GT / TAA / CAG**	28	60	0.003 *
ACNN**AC / TTA / CTG** (inv. cpl.)	49	39	0.583
GTNN**GT / TAA / TAG**	33	55	0.043
ACNN**AC / TTA / CTA** (inv. cpl.)	49	40	0.418
GTNN**GT / TGA / CAG**	29	60	0.004 *
GTNN**GT / TCA / CAG**	50	37	0.759
GTNN**GT / TTA / CAG**	42	45	0.552
GTNN**GT / TCC / CAG**	44	45	0.031
GNN**TTG / TAA / GCA**	34	54	0.068
GTNN**CC / TAA / CAG**	34	56	0.003 *
ATNN**CC / TTAA / CAC**	48	38	0.54 (2000)
ATNN**CC / TTAA / TAC**	47	43	0.51 (2000)
GANN**CT / TTA / GCT**	47	43	0.543
AGNN**TC / TAA / AGC** (inv. cpl.)	46	44	0.874

The one-sided P-value appears to indicate similarity to intron-like sequences with the splice site motifs GT, TRA and CAG, reminiscent in particular of eukaryotic major spliceosomal introns. The value of n provided in parentheses is the number of shuffled copies of the genomes used in approximating the P-value, wherever less than 20,000.

Table 5: Analysis of 1177 bacterial chromosomes.

Motifs Given order	Genomes Given Order	Genomes Reverse Order	One-sided P-value (n)
GTNN**GT / TAA / CAG** **AC**NN**AC / TTA / CTG** (inv. cpl.)	448 671	683 486	<0.001 * 0.94 (2000)
GTNN**GT / TAA / TAG** **AC**NN**AC / TTA / CTA** (inv. cpl.)	428 584	676 552	<0.01 (2000) * 0.05 (2000)
GTNN**GT / TGA / CAG**	592	568	0.01 (2000) *
GTNNCC / TAA / CAG	542	608	<0.01 (200) *
ATNN**CC / TTAA / CAC**	472	640	<0.01 (2000) *
ATNN**CC / TTAA / TAC**	492	616	<0.01 (2000) *
GTGNG / CTA / AT	352	818	<0.01 (2000) *
GANN**CT / TTA / GCT** **AG**NN**TC / TAA / AGC** (inv.cpl.)	588 572	571 568	0.06 (2000) 0.87 (2000)

The one-sided P-value appears to indicate similarity to intron-like sequences, with the major splice site motifs GTNNGT, TRA and YAG, the minor splice site motifs ATNNCC, TTAA and YAC, and also motifs reminiscent of group II intron splice sites: GTGNG, CTA and AT. The value of n provided in parentheses is the number of shuffled copies of the genomes used in approximating the P-value, wherever less than 20,000.

Table 6: Analysis of 11 genomes of the proteobacterial order Legionellales.

Motifs Given Order	Chromosomes Given Order	Chromosomes Reverse Order	One-Sided P-Value
GTNN**GT / TAA / CAG**	4	7	0.089
GTNN**GT / TAA / TAG**	7	4	0.816

These are genomes of bacterial intracellular pathogens of eukaryotes, all with a number of eukaryotic-like proteins. The lack of significant avoidance suggests that gene transfer from Eukaryotes to Bacteria cannot explain the observation of avoidance in Bacteria in general, since one would otherwise expect to see significant avoidance in these Bacteria, which live in such intimate association with eukaryotes.

References

1 Doolittle, W. F. & Stoltzfus, A. Molecular evolution. Genes-in-pieces revisited. *Nature* **361**, 403, doi:10.1038/361403a0 (1993).
2 Doolittle, W. F. Genes in Pieces - Were They Ever Together? *Nature* **272**, 581-582 (1978).
3 Martin, W. & Koonin, E. V. Introns and the origin of nucleus-cytosol compartmentalization. *Nature* **440**, 41-45, doi:nature04531 [pii]
10.1038/nature04531 (2006).
4 Roy, S. W. & Gilbert, W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet* **7**, 211-221, doi:nrg1807 [pii]
10.1038/nrg1807 (2006).
5 Gilbert, W. Why genes in pieces? *Nature* **271**, 501 (1978).
6 Lynch, M. & Richardson, A. O. The evolution of spliceosomal introns. *Curr Opin Genet Dev* **12**, 701-710, doi:S0959437X0200360X [pii] (2002).
7 Jeffares, D. C., Mourier, T. & Penny, D. The biology of intron gain and loss. *Trends Genet* **22**, 16-22, doi:S0168-9525(05)00323-9 [pii]
10.1016/j.tig.2005.10.006 (2006).
8 Penny, D., Hoeppner, M. P., Poole, A. M. & Jeffares, D. C. An overview of the introns-first theory. *J Mol Evol* **69**, 527-540, doi:10.1007/s00239-009-9279-5 (2009).
9 Koonin, E. V. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct* **1**, 22, doi:1745-6150-1-22 [pii]
10.1186/1745-6150-1-22 (2006).
10 de Souza, S. J. The emergence of a synthetic theory of intron evolution. *Genetica* **118**, 117-121 (2003).
11 Logsdon, J. M., Jr. The recent origins of spliceosomal introns revisited. *Curr Opin Genet Dev* **8**, 637-648, doi:S0959-437X(98)80031-2 [pii] (1998).
12 Gilbert, W. Genes-in-pieces revisited. *Science* **228**, 823-824 (1985).
13 Koonin, E. V. Intron-dominated genomes of early ancestors of eukaryotes. *J Hered* **100**, 618-623, doi:esp056 [pii]
10.1093/jhered/esp056 (2009).
14 Rodriguez-Trelles, F., Tarrio, R. & Ayala, F. J. Origins and evolution of spliceosomal introns. *Annu Rev Genet* **40**, 47-76, doi:10.1146/annurev.genet.40.110405.090625 (2006).
15 Glansdorff, N., Xu, Y. & Labedan, B. The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biol Direct* **3**, 29, doi:1745-6150-3-29 [pii]
10.1186/1745-6150-3-29 (2008).
16 Darnell, J. E. & Doolittle, W. F. Speculations on the early course of evolution. *Proc Natl Acad Sci U S A* **83**, 1271-1275 (1986).
17 Russell, A. G., Charette, J. M., Spencer, D. F. & Gray, M. W. An early evolutionary origin for the minor spliceosome. *Nature* **443**, 863-866, doi:nature05228 [pii]
10.1038/nature05228 (2006).
18 Pace, N. R. Time for a change. *Nature* **441**, 289, doi:441289a [pii]
10.1038/441289a (2006).
19 Basu, M. K., Rogozin, I. B. & Koonin, E. V. Primordial spliceosomal introns were probably U2-type. *Trends Genet* **24**, 525-528, doi:S0168-9525(08)00230-8 [pii]

10.1016/j.tig.2008.09.002 (2008).
20 Moore, C. E. & Archibald, J. M. Nucleomorph genomes. *Annu Rev Genet* **43**, 251-264, doi:10.1146/annurev-genet-102108-134809 (2009).
21 Lane, C. E. *et al.* Nucleomorph genome of Hemiselmis andersenii reveals complete intron loss and compaction as a driver of protein structure and function. *Proc Natl Acad Sci U S A* **104**, 19908-19913, doi:0707419104 [pii]
10.1073/pnas.0707419104 (2007).
22 Nelson, K. E. *et al.* Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature* **399**, 323-329, doi:10.1038/20601 (1999).
23 Zhaxybayeva, O. *et al.* On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. *Proc Natl Acad Sci U S A* **106**, 5865-5870, doi:0901260106 [pii]
10.1073/pnas.0901260106 (2009).
24 Garrett, R. A. & Klenk, H.-P. *Archaea : evolution, physiology, and molecular biology*. (Blackwell Pub., 2007).
25 Patel, A. A. & Steitz, J. A. Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol* **4**, 960-970, doi:10.1038/nrm1259
nrm1259 [pii] (2003).
26 Keeling, P. J. & Palmer, J. D. Horizontal gene transfer in eukaryotic evolution. *Nature reviews. Genetics* **9**, 605-618, doi:10.1038/nrg2386 (2008).
27 Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *Journal of theoretical biology* **8**, 357-366 (1965).
28 Sankoff, D. *et al.* Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proc Natl Acad Sci U S A* **89**, 6575-6579 (1992).
29 Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* **6**, 361-375, doi:nrg1603 [pii]
10.1038/nrg1603 (2005).
30 Lee, R. C., Gill, E. E., Roy, S. W. & Fast, N. M. Constrained intron structures in a microsporidian. *Mol Biol Evol* **27**, 1979-1982, doi:msq087 [pii]
10.1093/molbev/msq087 (2010).
31 Irimia, M. & Roy, S. W. Evolutionary convergence on highly-conserved 3' intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. *PLoS Genet* **4**, e1000148, doi:10.1371/journal.pgen.1000148 (2008).
32 Lim, L. P. & Burge, C. B. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* **98**, 11193-11198, doi:10.1073/pnas.201407298
98/20/11193 [pii] (2001).
33 Xing, Y. & Lee, C. Alternative splicing and RNA selection pressure--evolutionary consequences for eukaryotic genomes. *Nat Rev Genet* **7**, 499-509, doi:nrg1896 [pii]
10.1038/nrg1896 (2006).
34 Crotti, L. B. & Horowitz, D. S. Exon sequences at the splice junctions affect splicing fidelity and alternative splicing. *Proc Natl Acad Sci U S A* **106**, 18954-18959, doi:0907948106 [pii]
10.1073/pnas.0907948106 (2009).
35 Mourier, T. & Jeffares, D. C. Eukaryotic intron loss. *Science* **300**, 1393, doi:10.1126/science.1080559
300/5624/1393 [pii] (2003).

36 Loh, Y. H., Brenner, S. & Venkatesh, B. Investigation of loss and gain of introns in the compact genomes of pufferfishes (Fugu and Tetraodon). *Mol Biol Evol* **25**, 526-535, doi:msm278 [pii]
10.1093/molbev/msm278 (2008).
37 Slamovits, C. H. & Keeling, P. J. A high density of ancient spliceosomal introns in oxymonad excavates. *BMC Evol Biol* **6**, 34, doi:1471-2148-6-34 [pii]
10.1186/1471-2148-6-34 (2006).
38 Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**, 1056-1060, doi:nature08656 [pii]
10.1038/nature08656 (2009).
39 Douglas, S. *et al.* The highly reduced genome of an enslaved algal nucleus. *Nature* **410**, 1091-1096, doi:10.1038/35074092
35074092 [pii] (2001).
40 Deane, J. A., Strachan, I. M., Saunders, G. W., Hill, D. R. A. & McFadden, G. I. Cryptomonad evolution: Nuclear 18S rDNA phylogeny versus cell morphology and pigmentation. *Journal of Phycology* **38**, 1236-1244 (2002).
41 Zelin, E., Wang, Y. & Silverman, S. K. Adenosine is inherently favored as the branch-site RNA nucleotide in a structural context that resembles natural RNA splicing. *Biochemistry* **45**, 2767-2771, doi:10.1021/bi052499l (2006).
42 Slamovits, C. H. & Keeling, P. J. Evolution of ultrasmall spliceosomal introns in highly reduced nuclear genomes. *Mol Biol Evol* **26**, 1699-1705, doi:msp081 [pii]
10.1093/molbev/msp081 (2009).
43 Logsdon, J. M. & Faguy, D. M. Thermotoga heats up lateral gene transfer. *Current biology : CB* **9**, R747-751 (1999).
44 Schwartz, S. H. *et al.* Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res* **18**, 88-103, doi:gr.6818908 [pii]
10.1101/gr.6818908 (2008).
45 Russell, A. G., Shutt, T. E., Watkins, R. F. & Gray, M. W. An ancient spliceosomal intron in the ribosomal protein L7a gene (Rpl7a) of Giardia lamblia. *BMC Evol Biol* **5**, 45, doi:1471-2148-5-45 [pii]
10.1186/1471-2148-5-45 (2005).
46 Russell, C. B., Fraga, D. & Hinrichsen, R. D. Extremely short 20-33 nucleotide introns are the standard length in Paramecium tetraurelia. *Nucleic Acids Res* **22**, 1221-1225 (1994).
47 Chien, M. *et al.* The genomic sequence of the accidental pathogen Legionella pneumophila. *Science* **305**, 1966-1968, doi:10.1126/science.1099776 (2004).
48 Cazalet, C. *et al.* Analysis of the Legionella longbeachae genome and transcriptome uncovers unique strategies to cause Legionnaires' disease. *PLoS genetics* **6**, e1000851, doi:10.1371/journal.pgen.1000851 (2010).
49 Seshadri, R. *et al.* Complete genome sequence of the Q-fever pathogen Coxiella burnetii. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 5455-5460, doi:10.1073/pnas.0931379100 (2003).
50 Lurie-Weinberger, M. N. *et al.* The origins of eukaryotic-like proteins in Legionella pneumophila. *Int J Med Microbiol* **300**, 470-481, doi:10.1016/j.ijmm.2010.04.016 (2010).
51 Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-65, doi:gkl842 [pii]

10.1093/nar/gkl842 (2007).
52 Patron, N. J., Inagaki, Y. & Keeling, P. J. Multiple gene phylogenies support the monophyly of cryptomonad and haptophyte host lineages. *Curr Biol* **17**, 887-891, doi:S0960-9822(07)01255-9 [pii]
10.1016/j.cub.2007.03.069 (2007).
53 Vanacova, S., Yan, W., Carlton, J. M. & Johnson, P. J. Spliceosomal introns in the deep-branching eukaryote Trichomonas vaginalis. *Proc Natl Acad Sci U S A* **102**, 4430-4435, doi:0407500102 [pii]
10.1073/pnas.0407500102 (2005).
54 Denoeud, F. *et al.* Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* **330**, 1381-1385, doi:science.1194167 [pii]
10.1126/science.1194167 (2010).
55 Chu, V. T., Adamidi, C., Liu, Q., Perlman, P. S. & Pyle, A. M. Control of branch-site choice by a group II intron. *EMBO J* **20**, 6866-6876, doi:10.1093/emboj/20.23.6866 (2001).
56 Tourasse, N. J. & Kolsto, A. B. Survey of group I and group II introns in 29 sequenced genomes of the Bacillus cereus group: insights into their spread and evolution. *Nucleic Acids Res* **36**, 4529-4548, doi:gkn372 [pii]
10.1093/nar/gkn372 (2008).
57 Edgell, D. R., Belfort, M. & Shub, D. A. Barriers to intron promiscuity in bacteria. *J Bacteriol* **182**, 5281-5289 (2000).
58 Dunning Hotopp, J. C. Horizontal gene transfer between bacteria and animals. *Trends in genetics : TIG* (2011).
59 Drake, J. W. Avoiding dangerous missense: thermophiles display especially low mutation rates. *PLoS genetics* **5**, e1000520, doi:10.1371/journal.pgen.1000520 (2009).
60 Casadevall, A. Evolution of intracellular pathogens. *Annu Rev Microbiol* **62**, 19-33, doi:10.1146/annurev.micro.61.080706.093305 (2008).
61 Cornish-Bowden, A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res* **13**, 3021-3030 (1985).
62 Qin, H., Wu, W. B., Comeron, J. M., Kreitman, M. & Li, W. H. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics* **168**, 2245-2260, doi:168/4/2245 [pii]
10.1534/genetics.104.030866 (2004).