Evidence of Limited Structural Organization in Globin Intron Sequences of Messenger RNA

Wayne Dawson [1] **Kenji Yamamoto** [2]

dawson@ims.u-tokyo.ac.jp backen@lute.is.s.u-tokyo.ac.jp

[1] Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108, Japan

[2] Lab. Clinical Microbiology and Immunology, Bun'in Hospital, Medical School
University of Tokyo, 6-28-3 Mejirodai, Bunkyo-ku, Tokyo 112, Japan

Abstract

The structural significance of introns in α and β globin genes were evaluated in terms of the folding (or stacking) free energy of the nascent RNA sequences as obtained from secondary structure calculations. From calculations of the folding free energy of exons and introns, it was found that introns tend to have a more folded (globular) structure compared to the exon counterparts segments. The results suggest that significant biological information is present in all intronic sequences including spliceosome type introns.

1 Introduction

Since the discovery of introns, considerable progress has been made at understanding the general mechanisms that drive the intron splicing apparatus in eukaryotic messenger RNA (mRNA). Notwithstanding, in many respects, introns seem to carry no functional role in the pre-RNA (heterogeneous nuclear RNA (**hnRNA**)) sequence other than to interrupt the protein coding sequence of mRNA. Yet despite nearly a billion years of evolution [10, 33], this functionless intron has remained in the eukaryotic genes. Moreover, the number of introns has often increased in the pre-RNA sequence particularly for the higher forms of life [33]. To the contrary then, there is reason to think that the intron *is* important to the genetic apparatus.

In some cases, the 3D structure of the intron in nascent mRNA exhibits something akin to an enzymatic role in which the intron splices itself out of the gene [15, 25]. However, we also have introns that depend on a complex splicing apparatus to successfully extricate themselves from the pre-RNA sequence [22, 26]. Moreover, in some cases, much of the intron can be deleted or exchanged with another intron sequence, yet (unless there are cryptic sites involved) there is *no* change in the splice points [35]. Hence, not only do we lack understanding about the purpose of the intron, we still don't really know how much information (if any) is contained on a given intron sequence.

The majority of the introns depend on the same consensus sequence "GU $\cdots$ AG" in which "GU" is located at the $5'$ end of the intron and "AG" is located at the $3'$ end [31]. This dinucleotide pattern is the basis for the so-called "GU $\cdots$ AG rule". Generally, the exons do not appear to influence any steps in this splicing process neither chemically nor enzymatically although more recent findings are suggesting some influence [12]. Presently, four major types of introns are known.

The spliceosome type introns are the most common in which a host of small nuclear RNA (**snRNA**) sequences [21, 22, 23, 26] (and usually a number of auxiliary enzymes called SR proteins) [22] are required to cleave the introns from the hnRNA sequences. The precise cleaving mechanism of hnRNA is still under investigation but generally appears to involve a complex interplay between the snRNA, the SR proteins, and several short consensus sequences on the intron where the spliceosome apparatus (snRNA) hybridizes and eventually splices the intron out. We refer to this large splicing apparatus as the small nuclear ribonucleoprotein complex (**snRNP**).

Another type of intron has similar properties to the snRNA mentioned above, but the splicing sequence follows an "AT $\cdots$ AC rule". The proportion of these introns is much smaller than the former, however, they compose roughly 1/10000 of the population of snRNA and they often require their own special snRNP (U12) to carry out splicing [31].

The other types of introns so far isolated are called self splicing introns: type I and type II. Group I exons are found in the lower eukaryotic cells, ciliates like *Tetrahymena thermophile*, and slime molds like *Physarum polycephalum* [3]. Group II introns are found in organelles and bacteria. For these intron sequences, the secondary structure and tertiary structure determine the catalytic behavior of the introns [14, 15]. These type of structures are sometimes referred to as ribozymes.

The most complete way to examine the current problem of intron interactions within the hnRNA sequences would be a quantum chemistry calculation at the *ab initio* level. However, such calculations on these complex biological system are completely inaccessible to the current genre of computers. The next level is semiempirical calculations in a subset such as the 3D structure. However, at present, a robust calculation scheme of the general 3D structure remains rather elusive. MD simulation of the relaxation of large biological structures (minimum of 1 - 1000 μs for a 400 nucleotide sequence) are extremely expensive when one considers the number of atoms involved, the computation time for each iteration, and the increment (1 fs) for each iteration [39]. The demands on computer memory can also be daunting for a simple calculation even at the semiempirical level. Finally, these expensive calculations require a well developed strategy to interpret the complex interactions found in molecular dynamics simulations. It is the purpose of this work to help clear a path for this strategy.

Hence, to gain a foothold on future calculations of this far more complex 3D structural analysis, we have begun by resorting to a much simpler strategy of estimating the free energy obtained from the secondary structure of these sequences. Secondary structure calculation have been found to assist researchers in development the full 3D structure [40]. Generally, the 3D structure can be obtained by folding the secondary structure into the more complex tertiary structure [13, 40].

Estimation of the free energy of these sequences is useful in establishing how compact or globular these structures are. A very compact (highly folded) structure also has a large negative free energy and a loosely bonded structure (marginally folded) should exhibit a small negative free energy. Free energy calculations are very inexpensive to carry out in a secondary structure calculation.

We chose the globin family because it is common throughout the Chordata phylum and spans all the vertebrates. The GenBank database contains about 40 representative complete *coding sequences* (**CDS**), which is fairly large for a typical database of CDS. With such a *comparatively* large data set, the results of this investigation are expected to *approximate* the general properties of introns and exons in the globin gene family. Furthermore, to the best of our knowledge, globin hnRNA only exhibits constitutive splicing which eliminates any unusual splicing related effects. Finally, no recognizable Alu, L1, or MER1/2 sequences are found in any of the introns in these sequences; although Alu and L1 repeats are contained within sections of the entire globin gene [9]. All of the genes in the globin family appear to have three exons and two introns. The exon sequences have a number of common conserved elements, but the introns are often vastly different even between related families of species.

In this work, we report that, in the globin family, there is *some* evidence for internal structure in these snRNP based introns [6, 38]. We also suggest that this research may also serve to compliment search techniques for $5'$/exon/intron/$3'$ boundaries in unknown sequences.

2 Methods

Unless otherwise noted, all calculations were carried out using the SGI Origin 2000 computer at the Human Genome Center, Institute of Medical Science University of Tokyo. The software used in this work included a program written by the authors [6, 38], and the Wisconsin GCG Package which includes M. Zuker's mfold and foldrna programs [40].

(I) Gather data:
(1) Obtain hnRNA CDS[a] of gene.
(2) divide CDS into introns, exons, $5'$, and $3'$ subsequences.
(3) For *each* subsequence: calculate the free energy (ΔG) for *N-shuffled* sequences[b] .

[a] CDS = complete coding sequence
[b] $N \equiv$ sample set of shuffled sequences.

$\longrightarrow$

(II) Evaluate the distribution:
(1) find[a] $max(\Delta G) = \Delta G_j$, $min(\Delta G) = \Delta G_i$,
(2) evaluate[b] $\Theta = abs(\Delta G_j - \Delta G_i)$

[a] $(i,j) \in \Delta G_N \exists (\Delta G_i, \Delta G_j) \| [\lim_{N\rightarrow\infty} min(\Delta G_i) \rightarrow min(\Delta G)$, $and \lim_{N\rightarrow\infty} max(\Delta G_j) \rightarrow max(\Delta G)]$
[b] For a "spontaneous" (favorable) reaction the free energy must yield a negative value.

$\longrightarrow$

$\hookrightarrow$

(III) Construct a histogram:
(1) divide Θ into m sections.
(2) calculate $G(l) = max(\Delta G) - \Delta G_l$, (where $l \in N$).
(3) locate $G(l) \parallel \frac{i\Theta}{m} \leq G(l) < \frac{(i+1)\Theta}{m}$ (where $i < m$, and $l \in N$)
(4) construct an m-bin histogram based on assignment in (3).

$\longrightarrow$

(IV) Analyze the m-bin histogram:
(1) use the m-bin histogram indices as a sequence independent normalization[a] .
(2) Find the index of maximum density for the *shuffled* sequences.
(3) Find the index of original *unshuffled* sequence.
(4) Compare the indices.

[a] For all histograms compared in this way, m *must* be the same value.

Figure 1: Flow chart of the approach used to evaluate the exon and intron structure in hnRNA.

In this work, the word "structure" will be used to refer to both the *physical structure* (as in the geometrical arrangement of atoms or molecules in a sequence of hnRNA) and *chemical structure* (as in the chemical bonding and functional groups of a particular arrangement of atoms or molecules). When referring to secondary structure, or to bulk structural features (such as hair pins) we will specify these higher levels of abstraction as *secondary structure*. The word "function" will be used to refer to the biological function of a large protein or nucleotide sequence.

2.1 Computational methods

A flow chart of the approach used in this work to examine the exon and intron structure is diagrammed in Fig. 1.

First, we obtained complete CDS of the hnRNA. In the current work we found 10 α-globin species with 24 complete CDS, and 12 β-globin species with 20 complete CDS.

Next, we sliced the hnRNA into its intron, exon, $5'$ and $3'$ sections. A set of N-shuffled sequences for each section was created and the free energy (FE) was evaluated from the respective secondary structure.

From the results of the calculated FE values in Fig. 1(I), and the distribution of FE obtained in Fig. 1(II), a histogram of m indices is constructed (Fig. 1(III)).

Finally, maintaining the same binning between histograms, we then compared the FE of the original sequence with the mean free energy of the shuffled sequences and assigned the appropriate index. This methodology is carried out for all introns and exons in a given hnRNA.

A typical histogram of the FE is shown in Fig. 2. The horizontal axis represents the bin number (index) which also carries the function of normalization when the number of bins is kept constant for all histograms (Fig. 1(IV) and 2). In Fig. 2, the numbered columns along the horizontal axis represent the normalized FE (or a m-bin normalization of the free energy) in which the bin index numbers extend from the 1^{st}-bin to the m^{th}-bin. In the current convention, the 1^{st}-bin represents the largest (*i.e.*, the most negative) free energy, and 11^{th}-bin represents the smallest free energy (*i.e.*, the least negative) obtained in the shuffled distribution. The mean FE is located between the 6^{th}-bin and 7^{th}-bin. The FE of the original introns and exons are then plotted with respect to the distribution found in Fig. 2 and are shown in the exon-intron plot of Fig. 3.

2.2 Analysis

The very first problem one encounters in examining the energy distribution of all possible shuffled sequences generated from a given coding sequences of length n, is that evaluating *all* possible conformations is intractable. For example, suppose we have a sequence of 20 nucleotides (**nt**) and equal percentages of AUGC contained in the sequence. The number of possible sequences will be equal to a multinomial distribution

$$n!/(n_a!n_u!n_g!n_c!) \tag{1}$$

where n is the sequence length, $n = n_a + n_u + n_g + n_c$, n_a is the number of A in the sequence, n_u the number of U, etc. For a sequence of length 20 and $n_a = n_u = n_g = n_c = 5$, we obtain $20!/(5!)^4 = 1.17 \times 10^{10}$. The number of secondary structure conformations would be at least half this number, due to symmetry considerations and degeneracy effects in the secondary structure conformations, but would be of similar order. Moreover, we would still need to evaluate all sequences to determine this degeneracy and symmetry precisely. A typical exon or intron of mRNA contains sequences in excess of 100 nucleotides, which means that we would need to evaluate more than 10^{57} sequences to determine all possible conformations and completely describe the distribution.

Since such a complete evaluation is obviously outside the power of even the most advanced computers of this age, we have reasoned that shuffling a given sequence of length n into a set of N possible shuffled sequences will provide an *approximation* of this distribution. As reasoned in Fig. 1(II), as the number of random sequences N increases, the observed FE distribution will gradually approach the true FE distribution [7]. In the calculations done in this work, we have used values of N such that $N \geq 400$.

To estimate the error in this small number of calculations, we evaluated the fluctuation of the mean FE value for a given sequence (length: 100 nt) [4]. Each mean FE was determined by shuffling the given sequence N times. From the distribution of 200 separate mean FE calculations, we found that the ratio of the standard deviation (σ) to the range of the FE (Θ) comes out roughly to $\sigma/\Theta < 0.01$ (see Fig. 1). This means that the separation between bins in Fig. 2 is equal to roughly 10σ (where the range is divided into 11 bins: *i.e.*, $m = 11$). Other authors have found that 100 sequences is sufficient to make these estimations at a confidence level of 95% [20]. Hence, a threshold of 400 shuffled sequences should be sufficient to obtain the mean value of the distribution based on our calculation, and others [8, 18, 20].

We also examined the tendency of these sequences as a function of nt length. For sequences with fixed AUGC percentages, the property is linear as a function of nt length. For sequences with different AUGC concentration, the slopes are different but the linear dependence is the same [4].

The distribution seen in Fig. 2 is typical and appears to be fairly independent of the nt-length or percentages of AUGC components, although there *could be some* difference in the degree of dispersion (variance). In the sequences evaluated so far, the variance appears to be stable. If there is any significant fluctuation in the variance, it is most likely to appear at the limits of these distributions. For sequences with moderately equivalent percentages of AUGC (between 20 and 40%), the FE distribution obtained from secondary structure calculation appears to be a stable distribution [4]. Hence, we expect that this finite sampling method reflects a sufficient representation of the actual distribution of these sequences. In particular, this should be the case for the sequences analyzed here, since the distribution of AUGC in these globin genes fits within the 20 $\sim$ 40 percent range. Moreover, all these distributions

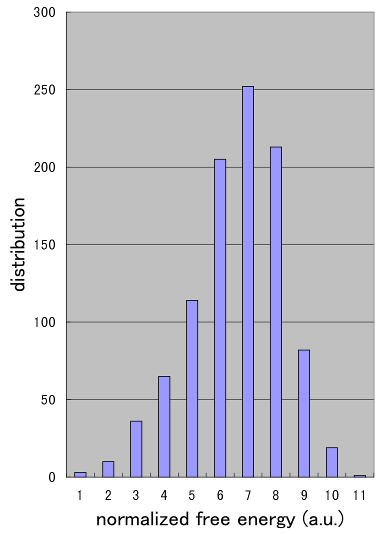

Figure 2: Distribution of free energy in shuffled sequences of a globin exon.

strongly resemble the familiar normal distribution. There is some indication that the distribution is log normal, but the shift is very close to a typical Gaussian distribution with a small contribution of skew symmetry (~ -0.5). Hence, we have adopted Gaussian statistics in our analysis. However, because of the small component of skew in the FE distributions, we will refer to the plot in Fig. 2 as a Gaussian-like density distribution (GLDD).

The next issue that one encounters in evaluating the free energy of a secondary structure calculation, is that secondary structure calculations remain a very imprecise technique for estimating the minimum free energy of a nucleotide sequence. At present, the estimated certainty is claimed to be about 70% accurate [13]. It should be noted, however, that the introduction of secondary structure analysis has greatly assisted researchers in other techniques such as searching for conserved sequences to deduce the structural relationship of hnRNA and mRNA. Hence, with some judicious considerations, we can expect to extract some *general* understanding. In particular, if numerous FE calculations of families of intron sequences consistently yield a similar result, this trend is reasonable to accept as genuine, but a lone calculation that has an exceptionally different value is probably questionable. As a result of these underlying issues, we have focused mainly on the general trend or behavior of these intron and exon sequences.

Thus, we have made the following assumptions in this work. First, we have used a normal distribution to evaluate the mean FE of the GLDD. Second, we have assumed that the optimal secondary structure of the shuffled sequences can make a sufficiently qualitative representation of the true stacking energies to make these estimates useful. Finally, we have assumed that the tertiary structure will not significantly change this result and yield a far different mean FE.

3 Results

The results in this report should only be considered suggestive; however, using only the secondary structure as a discriminating tool, we have found what appears to be a significant difference in the free energy of secondary structures of introns and exons found in the globin family.

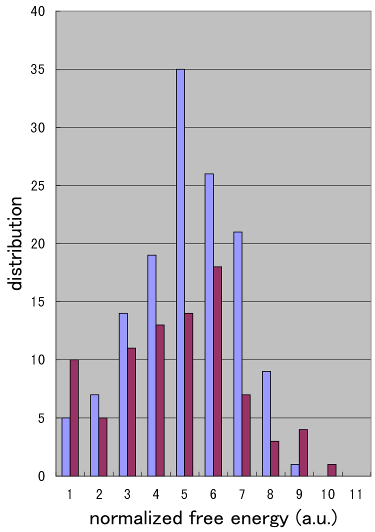

Figure 3: Distribution of free energy in intron (dark gray) and exon (light gray) sequences of globin gene from various warm blooded species. The sequence data were obtained from GenBank.

The distribution of exons and introns in the α/β-globin sequences are plotted in Fig. 3. The *mean* FE is roughly 6 in both Fig. 2 and 3. Due to the small negative skew (~ -0.5) of the FE distribution, the actual mean is slightly larger than 6 (mean: 6.6). However, since the error in the FE estimation using secondary structure stacking energies is likely to be a much greater source of error, we will quote 6 as the mean FE value. Sequences with a FE lower than 6 possess a more stable structure (more folded or globular). Sequences with a FE greater than 6 have a less stable structure and consequently, are less folded (or globular) than the mean sequences.

The wild type exons are distributed from bin 1 to bin 9 in Fig. 3(light gray) with a broad maximum at roughly bin 5 or 6 (near the mean FE). The dispersion of the exon FE values tends to resemble a normal distribution which broadly spans most of the FE range. The exon distribution shows *some* bias toward a larger negative FE, but less that the introns.

The distribution of intron FE values Fig. 3(dark gray) does not appear to take on a normal distribution about the mean FE. One group of introns appear to cluster around bin 5 (1 to 2 bins more negative than the mean FE) and another set of introns appears to have a very negative FE (centering at bin 1 in Fig. 3). On the other hand, *very few* of the introns appear to have a FE that is significantly greater than the mean FE of the shuffled sequences. Hence, we can likely assume that the majority of the introns in α and β-globin sequences are globular in nature, and that as a whole, this is the *tendency* of intron sequences.

The exon sequences show considerable conservation even when all the sequences are viewed together for the α and β globin genes. On the other hand, there is very little conservation between distant introns (Table 1). It should be expected that there will be strong conservation of the exon sequences since they are used in the protein coding sequences. Since the preserved homology for the introns is almost non-existent, this means that the biasing effect seen in the introns is probably real, and not an artifact of sequence conservation.

In the present analysis, we did not attempt to classify the nature of the resulting secondary structure. However, *in some cases* it was found that the initial fragments of the intron sequences have a shape in which the GU and AG ends are partially free but roughly a $5 \sim 10$ nt sequence is present

gene family	$ex(1)$	$ex(2)$	$ex(3)$	$in(1)$	$in(2)$
α-globin	0.75	0.75	0.75	0.07	0.10
β-globin	0.50	0.60	0.65	0.07	0.12

Table 1: Average similarity score for exon and intron sequences of α/β-globin used in the calculations.

which joins the $5'$ and $3'$ ends in a similar fashion. However, there are a number of other patterns that commonly occur. This may be part of the origin of the longer (but not well defined) "consensus sequence" that surrounds the doner and acceptor sites. This consensus sequence has been found both experimently [2] and by conditional probability analysis [30]. This may also indicate some part of the actual structure of the intron. However, the tertiary structure (of which little is known) is also likely to be significant for short intron sequences. These issues of structure and function will be considered in a future monograph.

4 Discussion

We preface this section by reiterating that, since the results are based on secondary structure calculations of the stacking free energies, only the general trend of the wild type intron and exon distributions will be considered. Whereas the results reported here are not nearly as impressive as recent results found for the cytochrome c family of genes [6], similar trends are observed in the globin family as well.

4.1 General considerations

We first address some general matters in relation to interpreting the calculated FE given here.

One point to consider is that it is not clear whether the secondary structure of hnRNA should be expected to segregate into intron and exon regions as we have "forced" the sequences to do in these calculations. However, others have tried fixed windows to scan a gene and have observed similar trends as reported here [8, 18]. There is certainly no precedent *against* using a sliding window and selecting the regions of interest. Rather this may even be important (*vide infra*: section 4.4).

As far as we know, the globin genes studied in this work are not subject to alternative splicing [12, 32]. It would appear that introns spanning regions of alternative spliced genes (*e.g.*, *Mcl1* gene of *D. Melanogaster* [18]) exhibits the *opposite* results found here (*i.e.*, intron 5 is less folded). Our calculations show roughly intron 5 (in ref. [18]) has a FE located at roughly bin 7 or 8. However, alternative splicing mechanisms are still poorly understood and are likely to require various regulatory proteins to govern the selection process [12, 32]. We don't anticipate that the behavior should be the same as genes that obey a constitutive splicing mechanism.

Further work is continuing on the functionality issues. We need to examine whether these results will come out the same when a select sub-set of the intron sequence is shuffled instead of the entire intron. More work is needed in studying the homology for clues of the intron structure. This might help to narrow down where the structural features occur.

4.2 Intron Distribution

4.2.1 Bimodality in the intron sequences

The intron distribution in Fig. 3 suggests that there could be two types of introns in globin. There is no correlation between intron number, but there is *some* tendency for the shorter introns ($n \lesssim 150nt$) to center around bin 5 whereas the longer introns tend to center around bin 1. The origin of this "bimodality" is unclear.

At present we think this effect could be the result of uncertainties due to unaccounted for solvent effects. These short nt segments are likely to critically depend on the stabilization energies introduced by the local concentration of various metallic ions such as Mg^{2+} and Na^{1+}. These are known to play an important role in the structure and possibly the function of type I introns [28, 5]. Although there is no data for snRNP related genes, ion stabilization is expected for most nucleic acid sequences based on their chemistry. Solvent effects are likely to play a significant role in stabilizing (or destabilizing) *short* nucleotide sequences because of the greater degree of exposure of the central core. Longer sequences appear to be highly stabilized by comparison due to their globular character. Hydrophobic/hydrophilic effects are more likely to play a central role in long intron sequences [8].

Ultimately, to interpret the results in Fig. 3, we will need to compute the free energy generated by the 3D structure with the inclusion of solvent interactions and charged ions. However, since the purpose of this current work is to show that there is a detectable effect resulting from a mere secondary structure calculation, we can safely say that there seems to be *some* evidence of structure in the intron sequences. We are currently looking into tertiary structure issues in relation to these shorter sequences.

One remaining issue is what level of confidence to place on any FE calculations based on a secondary structure calculation. Whereas the secondary structure is likely to generate more "scatter" than the actual distribution (or at least a "different kind of scatter"), a deviation of 40σ (between bin 1 and bin 5) seems unlikely even from such an extremely crude calculation. This would suggest that the bimodal character of the intron distribution is expressing some real property of introns.

There is also a matter of concern whether there could be some other artifact causing the higher degree of organization. One such possibility would be a strong homology conservation in the intron sequences since the exon sequences are strongly conserved. However, whereas the exons retain roughly 60% of their homology within their respective α or β families, the introns only retain roughly 10% . Therefore, whatever the origin of the intron band, it is unlikely that the bimodality is the result of some artifact in the conserved homology between species. Furthermore, there is no relation between the intron number and the magnitude of the FE. Another possibility would be various repeating sequences like *Alu*s. The *Alu*s are generally more globular than the random sequence. However, whereas these sequences have been found on some of the genes analyzed here, none of them were located in any part of the hnRNA sequence [9]. Other effects such as accounting for dinucleotide bias does not appear to change this result substantially, but is still under investigation.

4.2.2 Origin of the Lower Free Energy of the Intron

In Fig. 3, it is apparent that the majority of the data from the secondary structure calculation suggest *some* limited degree of structural organization in the introns. This contradicts the general notion that introns only rely on a trivial set of conserved sequences (the GU $\cdots$ AG rule and the "A" branch point; *e.g.*, see [35]). These conclusions were based on the fact that parts of an intron sequence could be spliced off with no apparent change in the cleavage of the introns as long as there were no inserted sequences which could also contain intron character. Such a lack of selectivity would strongly suggest that every property between intron selection to the transesterification could be handled by the snRNP, and that the intron sequence provides no significant information other than its consensus sequence. However, since the 3D structure is basically unknown for most of the hnRNA, it is still possible that there was no catalytic function associated with the deleted sections of the intron sequence in these experiments.

It is generally accepted that the dinucleotide pair GU $\cdots$ AG and the branch point is not sufficient information to fully define a constitutive splicing region. A cursory examination of any hnRNA sequence will show numerous dinucleotide sequences of "GU" and "AG" in any exon or intron region. Currently this plethora of possibilities in the hnRNA sequences are resolved by laborious examination of the mRNA and hnRNA sequences by well trained experts and selection is rarely determined strictly by the consensus sequence alone [2]. However, since the intron is spliced out of the hnRNA with almost perfect precision, this would suggest that the interaction mechanism is more complex than mere sequence information.

We suggest that structural organization is necessary for the intron in order for the snRNP complex to *recognize* the intron. More precisely, there is likely to be both chemical and structural aspects involved in the binding of SR proteins and snRNA to the intron-exon junctions. Such a large and complex apparatus (*i.e.*, recognition system) is likely to be "designed" to attack very specific chemical conformational configurations and functional groups. These chemical and configurational (steric) factors have been found already in type I introns [1, 11, 25, 28, 34] and type II introns [14, 15]. Hence, it is clear that specialized introns *are* structured and *can* form complex enzymatic-like structural architectures.

From a different angle, each part of the snRNP functions much like a subunit. However, because of the bulk size of this complex, most of the subunits must be in close proximity to work effectively since the primary rate limiting steps are probably both diffusion limited and rotationally restricted by the large hydrogen bonded network of water molecules. Likewise, it would not be thermodynamically efficient for these large subunits to seek out these consensus sequences, then to seek out all the other snRNA subunits (meanwhile dragging large sections of the pre-RNA substrate with them through histones, DNA, *etc.*), and only then to finally carry out catalysis. Rather, the intron should already be in the "proper configuration" (*i.e.*, the biologically active structure) whereupon the snRNP complex simply "recognizes" this "configuration", then carries out the two step transesterification process. Such a sequence of events is more thermodynamically feasible, and greatly reduces the rate limiting factors in the physical interaction. Furthermore, since thermodynamically efficient organisms are more likely to possess a selective advantage, reducing or optimizing the number of diffusion limiting steps is likely to have a relatively rapid fixation time in the population of pre-Cambrian proto-eukaryotic organisms [8, 20].

Moreover, the compact structure of hnRNA appears to influence the splicing order carried out by the SR proteins and snRNA yielding a unique electrophoresis pattern. Such splicing order strongly suggest steric effects and that splicing begins at the surface of a hnRNA molecule [19]. Moreover, the splicing does not appear to occur in a linear progression starting from the $5'$ end of hnRNA and ending with the $3'$ end. Rather splicing occurs according to some subset of possible orders [19]. All this suggests that some *shape* in the mRNA structure is required to initiate splicing.

Finally, there is already clear evidence of order in type I and type II introns, suggesting that at least *some* subsections of all intron sequences *might* convey some kind of significant information, which then controls, moderates or regulates the activity or function of the splicing apparatus. Although the majority of the catalytic potential of the spliceosome intron has obviously been lost, if all these introns share some common evolutionary history, then it should be reasonable to expect that some minimal catalytic structure is *necessary* in snRNP introns. Catalytic functions generally exhibit structural constraints as well as chemical functional group requirements. Thus, it should be reasonable to expect that configurational information must be preserved in various parts of the intron in order for the snRNP complex to successfully splice the intron. (See Appendix A for more information about intron types.)

4.3 Exon Distribution

The exons also exhibit some visible evidence of biasing in the overall exon FE distribution in Fig. 3. Although the biasing is less pronounced than the corresponding intron sequences, it is significant. We suggest that this is the consequence of conserved sequences that discourage the formation of excessive CpG islands. It is well known from mutation studies of mRNA of rabbit β-globin that certain types of sequences are conserved because the CpG islands are likely to result in hot spots for mutation [29]. Perhaps, as a result of this, the tertiary structure of the globin exon gene also shows some organization in its sequencing as a means of protecting against undesirable mutations in the protein sequence, and/or as a result or serendipitous positioning of conserved sequences associated with this phylogenetic conservation.

The biasing in the exon distribution in Fig. 3 may also be related to the extent to which the sequences are conserved in the globin family genes. In Table 1, it can be seen that the both α and β globin exons have a very high degree of sequence conservation (over 60%) despite the broad diversity of species analyzed in this work. This biasing is likely to be observed in the FE distribution. Moreover, it may be beneficial for the final mRNA to form a globular structure because the mRNA could be easily be exposed to the possibility of lethal mutations if all sections of its sequence were left unhybridized.

4.4 Relation to Bioinformatics

Bioinformatics has contributed substantially to our understanding of gene splicing. Recent work in conditional probability techniques combined with heuristic methods has helped to extend the simple consensus sequence (GU$\cdots$AG) [30]. Other work using k-gram analysis, in which a sequence of nt is analyzed as though it were a coded instruction set and the average size of a "word" in that sequence is analyzed. Intuitively, the exons produce a strong 3-gram signal, but introns appear to vary and generate 2, 4, 5, and 7-gram signals [16]. The k-gram studies suggest that the information content in the intron sequences consists of dinucleotide, tetranucleotides etc., but the information content in the coding regions consists of multiples of trinucleotide sequences [16]. If one assumes that the snRNP complex strictly utilizes a "text" analyzing mechanism to process the sequence information in the intron, then both these approaches are likely to yield success. Indeed, at least some significant part of the properties of introns is suggestive of this "instructions manual" type of concept.

However, it also appears that a simple text analyzing approach is not sufficient to completely explain how the introns are recognized by the snRNP complex. Many false positives are obtained in this form of analysis. Therefore, based on the *suggestive* nature of the results reported here, it seems reasonable that the current approach could complement research in finding *unknown* intron:exon boundaries. We suspect that the results we have obtained contain an important clue to the splice point detection mechanism, and could contribute to better search methods. We are currently looking for ways to apply this secondary structure aspect to searching unknown sequences.

5 Conclusions

The current analysis indicates that, contrary to established notions of intron behavior, there is *some* evidence for greater structural order and ribozyme-like function in the spliceosome introns. This suggest that all introns carry some degree of catalytic behavior, where spliceosome introns are the weakest of the known intron types. These results should help establish the critical mechanisms to consider when attempting to establish how introns are spliced out of the nascent RNA and the order in which they are excised. In addition, this may be another clue as to the function that introns play in the eukaryotic cell. Some limited structure is also observed for exons in the globin family which is likely a consequence of conserved sequences that protect evolutionary hot spots in these genes. Lastly, these studies of intron and exon structure may provide a complementary approach to detecting intron-exon boundaries in nascent RNA.

Acknowledgments

Research funding was supported through a grant from the Japan Society for the Promotion of Science. Support was provided by University of Tokyo Human Genome Center staff. Thanks are also extended to Prof. Yonezawa, Prof. Takagi, Dr. Morishita, and A. Nakaya, for their helpful discussions. Special thanks to Dr. T. Tsunoda for commenting on the original manuscript.

Appendix A. Type I, type II and spliceosome introns

Let k be the index of a given exon (or intron) such that $ex(k)$ (or $in(k)$) represent the k^{th} exon (or intron) in a hnRNA sequence. If we use a left to right sequence convention, then the doner site is ex(k):in(k) and the acceptor site is in(k):ex($k+1$).

Type I introns have been studied extensively and appear to strongly depend on the 3D-structure. Roughly ten domains are generally found and typically five triple helical junctions are also observed [24, 36]. The 3D-structure and chemical kinetics of the type I intron have been studied extensively and a number of the structures have been worked out [24, 36]. In this case, the majority of the structure and domains could be resolved by examining the consensus sequences alone [1, 3, 5, 11, 24, 25, 28, 34, 36]. However, in the case of type I introns, the length of the consensus sequences are considerable compared to the non-structural segments of these introns. Only a small part of these sequences can be completely deleted without a probable loss of functionality. For a detailed description of their structure and function see [3, 11, 24, 25].

The type II introns somewhat resemble precursors of the spliceosomal RNA apparatus [14, 15]. A typical feature of this hnRNA is the presence of 6 domains in the secondary structure. Experiments on plasmid sequences of domain 5 bear a strong resemblance to the function of U2/U6 in the snRNA [37]. However, the basic structure of type I and type II introns are very different and are not likely to have originated from the same evolutionary pathway [37]. The full tertiary structural details of type II introns are still unclear; however, some aspects of the consensus sequence are found in both type II introns and the hnRNA which utilize the snRNA (particularly the GU $\cdots$ AG splice points and the "A" branch point).

In the spliceosome process the snRNA U1 attaches to the $5'$ end (ex(k):in(k)) of the intron sequence, U2 connects to a branch point on the intron and then connects to U1. The branch point structure and location are partially characterized by a consensus sequence of form YNYYRAY; where Y is a pyrimidine, R is a purine and N can be any base. In this sequence, "A" is the only conserved element. U4 appears to act as an inhibitor for U6 in the final splicing operations. U5 binds to the $3'$ end of the intron (in(k):ex($k+1$)) in the last step of the transesterification process where the $3'$ end of the intron is excised and the $5'$ end of ex($k+1$) is mated with the $3'$ end of ex(k) to form ex(k):ex($k+1$). This step is accomplished after U5 appears to have triggered the release of U4 such that U6 can attach to U2 and also release U1 from the $5'$ position. As U1 is released, U6 replaces U1 at the $5'$ position of the intron. For further details, see [22].

References

[1] Batey, R.T. and Doudna, J.A., The parallel universe of RNA folding, *Nature Structural Biology*, 5:337-340, 1998.

[2] Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., and Chambon, P., Ovabumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries, *Proc. Natl. Acad. Sci*, 75:4853-7, 1978. (This is a general technique that is used today in much of the sequencing that occurs.)

[3] Cech, T.R., Conserved Sequences and Structures of Group I Introns: Building an Active Site for RNA Catalysis - a review, *Gene*, 73:259-271, 1988.

[4] Chou, Y., Dawson, W.K., and Yamamoto, Free energy distribution map of nucleotide sequences of varying composition, manuscript in preparation: to be submitted to *J. Theoretical Biology*; preprint available.

[5] Christian, E.L., and Yarus, M., Metal Coordination sites That Contribute to Structure and Catalysis in the Group I Intron from Tetrahymena. *Biochem.*, 32:4475-4480, 1993.

[6] Dawson, W.K., and Yamamoto, K., Evidence of structural information in cytochrome P450 family intron sequences of messenger RNA, *J. Comput. Biology*, submitted to the RECOMB 99 conference; Dawson, W.K., Yamamoto, K., Yonezawa, A., and Takagi, T., Ex-folded structure and in-folded structure of ribonucleic acid molecules, *Transactions A Bulletin of the American Physical Society*, 43:1386, 1998;

[7] Fontana, W., and Schuster, P., Continuity in Evolution: On the Nature of Transitions, *Science*, 280:1451-1455, 1998; Ibidem, The Possible and the Attainable in RNA Genotype-Phenotype Mapping, *J. Theor. Biol.*, in press.

[8] Forsdyke, D.R., A stem-loop 'kissing' model for the initiation of recombination and the origin of introns, *Mol. Biol. Evol.*, 12:949-958, 1995.

[9] GenBank: for example OCU60902, HUMHBA4, or RABBGLOB.

[10] Gilbert, W., Glynias, M., On the ancient nature of introns, *Gene*, 135:137-144, 1993.

[11] Herschlag, D., and Cech, T.R., Catalysis of RNA Cleavage by the Tetrahymena thermophila Ribozyme, *Biochem.*, 29:10159-10171, 1990.

[12] Inoue, K., Ohno, M., and Shimura, Y., Aspects of splice site selection in constitutive and alternative pre-mRNA splicing, *Gene Expression*, 4:177-182, 1995.

[13] Jaeger, J.A., Turner, D.H., and Zuker, M., Predicting Optimal and Suboptimal Secondary Structure for RNA, *Meth. in Enzymology* 183:281-306,1990; Jaeger, J.A., Turner, D.H., and Zuker, M., Improved predictions of Secondary Structures for RNA, *Proc. Natl. Acad. Sci.*, 86:7706-7710, 1989.

[14] Jarrell, K.A., Dietrich, R.C., and Perlman, P.S., Group II Intron Domain 5 Facilitates a trans-Splicing Reaction, *Mol. Cell. Biol.*, 8:2361-2366, 1988.

[15] Koch, J.L., Boulanger, S.C., Dib-Hajj, S.D., Hebbar, S.K., and Perlman, P.S., Group II Introns Deleted for Multiple Substructures Retain Self-Splicing Activity, *Mol. Cell. Biol.*, 12:1950-1958, 1992.

[16] Konopka, A.K., Sequences and Codes: Fundamentals of Biomolecular Cryptology, *Biocomputing, Informatics and Genome Projects,* (E.W. Smith *eds.*), Academic Press, Inc. N.Y. 1994, p. 119-174.

[17] Le, S.-Y., and Maizel, J.V., Jr., A method for assessing the statistical significance of RNA folding, *J. Theor. Biol.*, 138:495-510, 1989.

[18] Leicht, B.G., Muse, S.V., Hanczyc, M., and Clark, A.,G., Constraints on Intron Evolution in the Gene Encoding the Myosin Alkali light chain in Drosophila, *Genetics*, 139:299-308, 1995.

[19] Lewin, B., Genes VI, Oxford University Press, Inc., N.Y. 1997.

[20] Li, W.-H., and Graur, D., Fundamentals of Molecular Evolution, Sinauer Associates, Inc., Sunderland (USA), 1991.

[21] Madhani, H.D., and Guthrie, C., A Novel Base-Pairing Interaction between U2 and U6 snRNAs Suggests a Mechanism for the Catalytic Activation of the Spliceosome, *Cell*, 71:803-817, 1992.

[22] Madhani, H.D., and Guthrie, C., Dynamic RNA-RNA Interactions in the Spliceosome, *Annu. Rev. Genet.*, 28:1-26, 1994.

[23] McPheeters, D.S., and Abelson, J., Mutational Analysis of the Yeast U2 snRNA Suggests a Structural Similarity to the Catalytic Core of Group I Introns, *Cell*, 71:819-831, 1992.

[24] Michel, F., and Westhof, E., Modelling of the Three-Dimensional Architecture of Group I Catalytic Introns Based on Comparative Sequence Analysis, *J. Mol. Biol.*, 216:585-610, 1990.

[25] Narlikar, G.J., and Herschlag, D., Mechanistic Aspects of Enzymatic Catalysis: Lessons from Comparison of RNA and Protein Enzymes, *Annu. Rev. Biochem.*, 66:19-59, 1997.

[26] Nilsen, T.W., RNA-RNA Interactions in the Spliceosome: Unraveling the Ties That Bind, *Cell* 78:1-4, 1994.

[27] Price, S.R., Evans, P.R., and Nagai, K., Crystal structure of the spliceosomal U2B$''$ -U2A$'$ protein complex bound to a fragment of U2 small nuclear RNA, *Nature*, 394:645-650, 1998.

[28] Pyle A.M., Ribozymes: A Distinct Class of Metalloenzymes, *Science*, 261:709-714, 1993.

[29] Salser, W., Globin mRNA Sequences: Analysis of Base Pairing and Evolutionary Implications, *Cold Spring Harbor Symposium on Quantum Biology*, 42:985-1103, 1977.

[30] Salzberg, S.L., A method for identifying splice sites and translational star sites in eukaryotic mRNA, *CABIOS*, 13:365-376, 1997.

[31] Sharp, P.A., Burge, C.B., Classification of Introns: U2-type and U12-type, *Cell*, 91:875-879, 1997.

[32] Smith, C.W.J., Patton, J.G., and Nadal-Ginard, B., Alternative splicing in the control of gene expression, *Annu. Rev. Genet.*, 23:527-77, 1989.

[33] Stoltzfus, A., Spencer, D.F., Zuker, M., Logsdon, J.M. Jr., Doolittle, W.F., Testing the exon theory of genes: the evidence from protein structure, *Science*, 265:202-207, 1994.

[34] Strobel, S.A., Ortoleva-Donnelly, L., Ryder, S.P., Cate, J.H., and Moncoeur, E., Complementary sets of noncanonical base pairs mediate RNA helix packing in group I intron active sites, *Nature Structural Biology*, 5:60-65, 1998.

[35] Stryer, L., Biochemistry, (*ed* 4), W.H. Freeman and Company, N.Y. 1995. pp. 861.

[36] Wang, J.-F., Downs, W.D., and Cech, T.R., Movement of the guide Sequence During RNA Catalysis by a Group I Ribozyme, *Science*, 260:504-508, 1993.

[37] Weiner, A.M., mRNA Splicing and Autocatalytic Introns: Distant cousins or the Products of Chemical Determinism, *Cell*, 72:161-164, 1993.

[38] Yamamoto, K., and Yoshikura, K., RNA folding and evolution, *Visualizing biological information*, C. A. Pickover *eds.*, World Scientific, London, 158-164, 1995.

[39] Zhong, Q., Jiang, Q., Moore, P.B., Newns, D.M., and Klein, M.L., Molecular dynamics simulation of a synthetic ion channel, *Biophysical Journal*, 74:3, 1998.

[40] Zuker, M., On Finding All Suboptimal Foldings of an RNA Molecule, *Science*, 244:48-52, 1986.