

Fold Pressure and Origin of Introns

Tanmoy Paul*

Department of Zoology, Vivekananda College, Kolkata, India

Abstract

There are two theories of origin of introns - Introns Early and Introns Late. Recently, it has been shown that different pressures in DNA for maintenance of different functions leads to origin of introns by Introns Early theory. In this review it will be seen that fold pressure for DNA repair and protein pressure for useful protein formation cause origin of intron- exon organization that are found in the modern genes. The details of snake venom phospholipase A2, MHC gene clusters and retroviral genomes will show us the reciprocal relationship between fold and protein pressure, difference between nonsynonymous substitutions faced by introns and exons and ultimately how genomic conflicts is resolved by localization of different pressures to different location of genome which results in intron (noncoding part) and exon (coding part) organization of the genome.

Introduction

Eukaryotic genes contains multiple regions called introns that are removed post transcriptionally during splicing. After splicing exons of the pre mRNA are joined to form mature mRNA that is translated to form proteins. There are two theories of origin of introns: Intron Early and Intron Late [1] Intron early model describes ancient origin of introns and their subsequent different loss and maintenance in different lineages. Whereas, intron late model deals with the idea that earliest genes did not contain introns, which is subsequently added to some genes (Krebs et al, 2011). Modern views, supports the intron early model.

RNA being single stranded, forms secondary stem-loop structure. RNA stem-loop structures interacts among themselves by kissing interaction [2]. Such kissing interaction is believed to have played roles in ancient recombination. DNA has been shown to form similar secondary stem-loop structures like RNA from their classical duplex form. There is a genome wide pressure for forming single stranded DNA stem-loop structure extruded from classical duplex DNA -this is called fold pressure.

It has been observed that, there are many pressures on genome for stability and function. GC/AT pressure, fold pressure and protein pressure are important in respect to intron evolution. GC/AT pressure explains genome wide pressure for distinctive balance between proportions of AT and GC base pairs. This indicates stability, as DNA with more GC base pairs are more stable than DNA with more AT base pair percentage. Protein pressures indicates the pressures for encoding proteins with specific functions.

As the genome has to accommodate between different pressures situation arises when genomic conflicts can occur. Is there any situation when conflicts between these pressures can lead to changes in genome architecture? Is this related to origin of introns? Modern papers casts light resolving an ancient mystery-origin of introns.

Homologous Recombination: The Kissing Model

Homologous pairing is the most enigmatic stage of homologous recombination, to be stated more dramatically, how can two homologous needles find each other in the genomic haystack? (Kupiec, 2008). There are two models of homologous recombination: The cut first model; that DNA duplex is nicked at start of recombination and The pair first model; that complementarity between homologous recombination is established before cut. But both models deals with the problem of finding homology.

Tomizawa [3] proposed that, the complementary single stranded RNA interact by kissing interaction between the loops of RNA stem-loops. The interaction Tomizawa proposed is weak, reversible and transient. It was thought that, similar kissing interaction may be involved in homology search during homologous recombination, when duplex DNA extruded to form cruciform structure [4-6]. It has been seen that, in yeast the frequency of recombinational repair is 100%, so it is possible that information about the positions of homologous sequences are stored in the spatial organization of genome [7].

DNA molecule remains in an equilibrium between its classical helical form and extruded form generated by extrusion. This extruded structure consists of stems and loops. The process by which extrusion occurs is completely reversible and generally a segment of DNA vibrates between classical and cruciform (extruded) structure [8], (Figure 1).

But sometimes there are proteins that binds single stranded DNA, in that case extrusion between two forms classical and extruded cannot occur at same frequency [8]. Frequency of vibration is lower for GC rich DNA due to more stability of GC base pairing whereas AT rich sequence can give us high vibration frequency as there are only two hydrogen bonds between A and T [8].

Watson crick base pairing dominates in kissing loops. Whereas, similar (G+C)% ensures opening of classical helices into stem-loop structures synchronously by means of negative supercoiling. Whereas even slightly different (G+C)% values can change the timing of extrusion event so duplex cannot be extruded symmetrically [8].

Rise of Fold Pressure: Recombination and Replication Fidelity

As characteristic stem-loop structure helps in recombination, so in an early RNA world the ability to replace a damaged segment with an

*Corresponding author: Lecturer of Zoology, Department of Zoology, Vivekananda College, Kolkata 700063, West Bengal, India, Tel: +919477455669; Email: tanpaul@rediffmail.com

Received March 26, 2012; Published July 27, 2012

Citation: Paul T (2012) Fold Pressure and Origin of Introns. 1: 210. doi:[10.4172/scientificreports.210](https://doi.org/10.4172/scientificreports.210)

Copyright: © 2012 Paul T. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

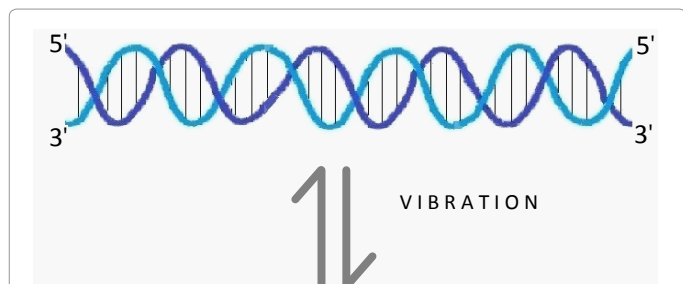


Figure 1: Extrusion of DNA cruciform structure from normal duplex DNA. This two forms remains in equilibrium. The two strands of normal duplex usually extruded simultaneously due to similar base composition, other sequences with similar base composition and base order will have similar vibration frequency. Frequency of extrusion is always greater in AT rich sequence than GC rich sequence as the later has more stability due to three hydrogen bonding.

undamaged one which could continue rapid and accurate replication could have been advantageous. Replicators that modifies their sequence to increase the probability of stem-loop formation may have a survival advantage [9]. Selection favors the sequences that can form distinctive stem-loop structures. Formation of stem-loops is necessary for recombinational repair that increases replication fidelity. Thus, in ancient world fold potential must be advantageous, because it gives its bearer the ability to copy genetic information perfectly.

Rise of Protein Pressure: Neofunctionalization and Positive Selection

Origin of novel function of a protein is called neofunctionalization, it is associated with positive selection of a gene which results in adaptive evolution [10,11]. According to birth and death model of protein evolution, rapid functional and structural diversification of a protein by sequence evolution occurs by gene duplication. The duplicated copies faces two fates: They can be inactivated by conversion to pseudogene, whereas others acquired novel functions, a phenomenon called neofunctionalization [12-14]. Genes that experiences positive selection, often have products that act as unique bioweapons in the evolutionary arms race [11]. In which two protagonist species are involved, and change in one casts an immediate selective pressure on another to resist that change. Seen in predator-prey, parasite-host interaction, in which satisfaction of protein pressure is very necessary for not to lose ground in evolutionary arms race.

Genomic Conflict: Fold Pressure and Protein Pressure

Fold potential are necessary for proper folding into secondary structures for the kissing interaction during recombination. On the other hand protein innovation is also needed for evolutionary arms race. Possibilities of genomic conflict, between protein pressure and fold pressure, arise because a sequence may not be able to locally optimize both base order dependent stem-loop potential or fold potential and base order dependent protein encoding potential [15], (Figure 2).

Thus, from the diagram it can be seen that even simple nucleotide sequence has a problem to compensate between fold and protein pressure [8].

Distribution of positive FORS-D value throughout the genome indicates the evolutionary pressure to promote stem-loop potential whereas, negative FORS-D in some region shows us that the conflict exists between general fold potential and local protein pressure for gene experiencing positive selection [16,11].

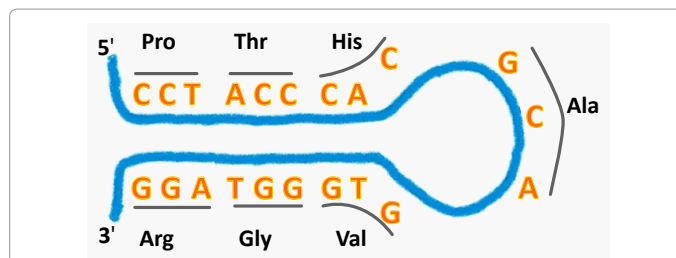


Figure 2: Conflict between genome wide fold pressure and local protein pressure. If Proline is necessary for a protein then it must incorporate Arginine, otherwise the nucleic acid will not be folded properly to stem-loop structure. Therefore, there is potential conflict between need of a nucleic acid for stem-loop formation and its need to encode a particular protein resulting in genomic conflict.

Conflicts resulted in intron evolution Positive FORSD values are widely dispersed throughout genome and greatly exceeds negative FORS-D values [16]. Incapability of a sequence to optimize fold and protein pressures simultaneously resulted in potential conflict between both pressures. The conflict can be bypassed in three ways 1. use of synonymous codons to balance both pressures. 2. broadening of codon choice by use of chemically and functionally similar amino acids 3. Sequence encoding protein can be placed over wide region allowing intervening sequence to optimize fold pressure. When first two options are not sufficient to bypass the conflict only third alternative remains. So introns might correspond to part of genome where constraints of first two ways are most severe. Conflict hypothesis explains the pressures that shape the intron-exon organization found in modern gene [16].

Distribution of Substitution Densities (Fors-D) In Modern Genes

As described above there are three ways to resist the genomic conflict. Three modern genes shows us the fact that when conflict is resolved by taking first two options the third is not needed, as evident from the analysis of retroviral quasispecies. But when the first two ways are not sufficient evolution takes the third way which leads to the origin of introns, as seen in snake venom phospholipase A2 and MHC gene clusters.

Retroviral Quasispecies

There are three ways to bypass potential genomic conflict between fold pressure and protein pressure. In retroviral genomes conflicts is resolved by first two categories: use of synonymous codons and functionally similar amino acids to balance between the above pressures.

Computer analysis of poliovirus and retroviral RNAs reveals the reciprocal relationship between stem-loop potential (fold potential) and sequence variability (protein pressure) [16]. This phenomenon is seen in poliovirus [17] Visua virus [18], human retrovirus HIV1 [19,20]. In HIV1 there are regions of high sequence variability associated with specific functions, characterised by negative FORS-D value, these regions face positive selection pressure by host defenses [16].

But there is also the need to conserve functional genomes among members of viral quasispecies, by recombination. Thus rapidly evolving genomes of retrovirus needs fold potential as specific function is needed. It is seen that, by use of synonymous codon and functionally similar amino acids in HIV1 the two pressures (that is fold and protein pressures) is balanced. Thus base substitutions are rarely accepted in regions rich in stem-loops.

FORS-D analysis of HIV1 subtype B shows that, the sequence encoding part of CD4 recognition sequence (gp 120) have high FORS-D value. That means this sequence is conserved for stem loop potential as expected from a negatively selected sequence [16]. The same phenomenon is also shown by Le and coworkers that on around 7900 nucleotide in env gene of retroviral genome- the R region is most conserved as it has most positive FORS-D value [19-21]. On the contrary there also have regions with negative FORS-D value at 3' end of env gene(nt 7901-8700) satisfying protein pressure.

It is shown that generally, in compact HIV1 genome functional constraints are actually limited stem-loop potential.

Snake Venom Phospholipase A2

Phospholipase A2 is an enzyme (Molecular Weight 85000) [22] found in mammalian tissues, insects and snake venom. Phospholipase A2 (PLA2) can be intracellular or extracellular. Intracellular PLA2 are involved in normal cellular activities like metabolism, cell signalling etc. Whereas, extracellular PLA2 is present in snake venom. PLA2 toxicity, for instance myotoxicity, neurotoxicity etc, is mediated by the Pharmacological Sites present on the PLA2 protein surface that recognise ligand on target cell surface. These Pharmacological Sites are modified structurally and experience positive selection which will lead to toxin diversity [11].

Many species of snakes have multiple PLA2 genes which usually consist of four exons [23]. The lengths of Viperidae PLA2 genes including proximal parts of promoter and flanking regions are 2.0-2.7kb long [24,25].

PLA2 multigene families contain a variety of PLA2 paralogs, that is copies of genes created by gene duplication about 40-50 million years ago with diverse pharmacological activities acquired during course of evolution. It was thought that snake venom PLA2 multigene families arise by gene duplication followed by divergence by substitution from a single ancestral PLA2 gene associated with the generalized function for digestion [11].

Neofunctionalization and arms race Few changes on the surface amino acids is enough for novel protein functions to arise, but there are constraints too. Structural and functional constraints on other parts of molecule limits divergence. It was proposed that, there is a perfect complementarity between pharmacological sites and target sites on prey cell surface. Duplication and divergence mainly by substitution altering the patterns of pharmacological sites alter binding specificities, a few may create new interaction sites leading to emergence of novel function (neofunctionalization) [26]. Pharmacological sites, the part of the enzyme facing arms race is the site of intense positive selection for neofunctionalization. So it is the region where high substitution rate would be expected [26].

Neofunctionalization of one side will create a strong pressure on another side. So was the case for snake and its prey. Thus neofunctionalization of snake PLA2 creates pressure for target cell receptor modification resulting in arms race [27,28].

Distribution of substitution density and positive selection The study of substitution density and indel density shows their peaks in regions corresponding to exons, whereas they are low in introns [16]. It is also found that, there are about 98% and 89% homologies present between 5'UTR and 3'UTR regions of many PLA2 genes, respectively, on the other hand exonic homologies are rather low about 67% [29]. High rate of nonsynonymous substitution is evident in exons of V. ammodytes

PLA2 genes, particularly on first and second position of codons which means a high rate of meaningful mutation indicating positive selection [10].

The expectation for high Ka/Ks ratio in protein coding exons of PLA2 gene are quite relevant when a free ratio model estimates separate Ka/Ks ratio for all lineages in a tree. The result indicates 20%-30% of PLA2 genes face positive selection [11].

The distribution of stable and unstable triplets are consistent with above data. There are unstable triplets present more frequently in exons whereas codons in introns are more stable (indicating fold pressures). It is observed that 5/8 stable triplets are found in introns whereas 5/8 unstable triplets are present in exons, consolidating the indications of positive selection for arms race [11].

Distribution of FORS-D value in PLA2 gene According to Donald Forsdyke [9] there is a reciprocal relationship between FORS-D values and substitution densities in PLA2 gene. He shows that, FORSD values are low in coding parts of PLA2 gene (exon) whereas FORSD values are high in introns and 5' and 3' noncoding region.

Major Histocompatibility Complex Gene Cluster

The major histocompatibility complex molecule (MHC) are involved in T cell mediated adaptive immune response. MHC genes occur in highly polymorphic MHC gene clusters. After endogenous and exogenous antigen processing, proteins are presented with MHC molecule, on antigen-presenting cell (APC) surface to T cell receptor on T cell [30].

MHC and arms race Intracellular and extracellular pathogens mutate their proteins to evade from host immune system and so host MHC also have to respond by appropriate mutation. This situation creates evolutionary arms race between host MHC and microbial antigens [31].

MHC and positive selection it is the peptide recognition domains of MHC proteins that matters in evolutionary arms race with microbial antigen. Thus genes corresponding to peptide binding region face positive selection whereas other conserved regions are negatively selected. A high nonsynonymous/synonymous (Ka/Ks) mutation ratio indicates a region under positive selection. It has been shown that peptide binding region of MHC class I and MHC class II molecule shows a high Ka/Ks value indicating positive selection occurring at this region [32].

MHC class I gene, HLA-A1 has four exons. Exon 1 encodes signal peptide, exon 2-4 encodes three extracellular protein domains (a1,a2,a3), exon 5 encodes transmembrane domain and exon 6 and 7 encode intracellular domains. When FORS-D values are calculated, extremely negative FORS-D values are seen in first four exons whereas, positive FORS-D values are more frequent in introns and flanking regions. The second and third exon encoding highly polymorphic a1 and a2 domains shows maximum nonsynonymous substitution when compared with other class I genes [30,32]. Decreased FORS-D values show inability to avoid the genomic conflict with synonymous codon and replacement with similar amino acid (seen in case of retroviral genome). It is thought that decreased FORS-D values may have resulted from positive selection in second and third exons. Also evident is the fact that low FORS-D values are associated with regions, involved in binding of peptide and/or T cell receptor, facing positive selection [33].

Similar pattern becomes evident when considering MHC class II Ab chain encoding gene [34]. Here, the second exon encodes

polymorphic sequences, involved in interaction with peptide and T cell receptor, thus this exon is positively selected showing lower FORS-D value. Thus polymorphic exons face intense positive selection, reason they have lower average FORS-D value compared to nonpolymorphic exons [35].

Origin of Introns: Introns Early

There are two hypothesis for the origin of introns: Introns Early and Introns Late [36]. According to Introns Early theory introns are present in ancestors of prokaryotes and eukaryotes between genes [37-39]. Introns then lost from all prokaryotic lineages whereas they are maintained in eukaryotes by splicing [36].

According to Donald Forsdyke [16] introns originated by Introns Early hypothesis. Because, it has been observed that introns interrupt genetic information generally not just protein coding genes. Introns can be found within 5' and 3' noncoding regions [40] genes encoding mRNAs but have no protein [41,42]. Consistent with the idea that introns interrupts all the genetic information is the findings that approximately one half of known introns don't interrupt genes reading frame [36]. Therefore, there is no significant link between intron location and protein structure [43]. This line of evidence proposes an ancient origin of introns by Intron Early model.

As recombinational repair can serve for accurate replication and thus stability of genetic material, so ability to recombine may have evolved early in evolution and replicators having this property may be the winner in competition. So, fold potential, ability to form stem-loops for recombination may have evolved before ability to encode novel proteins (protein functions) and this line of argument also supports ancient origin of introns.

Conclusion

It is evident that, understanding classic Watson-Crick base pairing of DNA does not know everything about the informational role that DNA played in evolutionary history. Our study shows us that DNA contains many informations that cannot be explained by the classic complementarity between bases. Even spatial organization of secondary structures of DNA can act as potential information. Understanding the forms of informations that DNA carries, is essential to more intricate understanding of genome architecture and its origin.

New approach, views genome as channels containing multiple forms of informations. A new branch of bioinformatics - Evolutionary Bioinformatics formed and devoted to study of different forms of informations that DNA contains. We have seen that genome is overloaded with different information (fold pressure, GC pressure, protein pressure etc.), so genomic conflict is obvious. It is the task of evolutionary bioinformatics to unveil the mystery that, despite the presence of potential conflicts, how transmission of genetic information occurs reliably through generation.

References

- Belshaw R, Bensasson D (2006) The rise and fall of introns. *Heredity* 96: 208-213.
- Tomizawa J (1984) Control of ColE1 plasmid replication: the process of binding of RNA I to the primer transcript. *Cell* 38: 861-870.
- Tomizawa J (1984) Control of ColE1 plasmid replication: the process of binding of RNA I to the primer transcript. *Cell* 38: 861-870.
- Sobell HM (1972) Molecular mechanism for genetic recombination. *Proc Natl Acad Sci USA* 69: 2483-2487.
- Wagner RE, Radman M (1975) A mechanism for initiation of genetic recombination. *Proc Natl Acad Sci USA* 72: 3619-3622.
- Kleckner N, Weiner BM (1993) Potential advantages of unstable interactions for pairing of chromosomes in meiotic, somatic and premeiotic cells. *Cold Spring Harbor Symp Quant Biol* 58: 553-565.
- Barzel A, Kupiec M (2008) Finding a match: how do homologous sequences get together for recombination? *Nat Rev Genet* 9: 27-37.
- Forsdyke DR (2011) *Evolutionary Bioinformatics*. (2nd edn), Springer, New York.
- Forsdyke DR (1995a) A stem-loop kissing model for the initiation of recombination and the origin of introns. *Mol Biol Evol* 12: 949-958.
- Gubensek F, Kordis D Venom phospholipase A2 genes and their molecular evolution, in R. M. Kini (Ed.) *Venom Phospholipase A2 Enzymes: Structure, Function and Mechanism*, Wiley, Chichester. 1997, pp. 73-95.
- Kordis D (2011) Evolution of phospholipase A2 toxins in venomous animals. *Acta Chim Slov* 58: 638-646.
- Lynch VJ (2007) Inventing an arsenal: adaptive evolution and neofunctionalization of snake venom phospholipase A2 genes. *BMC Evol Biol* 7:2
- Kordis D, Gubensek F (2000) Adaptive evolution of animal toxin multigene families. *Gene* 261: 43-52.
- Kordis D, Krizaj I, Gubensek F (2002) Functional diversification of animal toxins by adaptive evolution. In: A. Menez (Ed.) *Perspectives in Molecular Toxicology*, Wiley, Chichester, pp. 401-419.
- Forsdyke DR (1995b) Conservation of stem-loop potential in introns of snake venom phospholipase A2 genes: An application of FORS-D analysis. *Mol Biol Evol* 12: 1157-1165.
- Forsdyke DR (1995) Reciprocal relationship between stem-loop potential and substitution density in retroviral quasispecies under positive Darwinian selection. *J Mol Evol* 41: 1022-1037.
- Currey KM, Peterlin BM, Maizel JV (1986) Secondary structure prediction of poliovirus RNA: correlation of computer-predicted with electron microscopically observed structure. *Virology* 148: 33-46.
- Braun MJ, Clements JE, Gonda MA (1987) The visna virus genome: evidence for a hypervariable site in the env gene and sequence homology among lentivirus envelope proteins. *J Virol* 61: 4046-4054.
- Le SY, Chen JH, Braun MJ, Gonda MA, Maizel JV (1988) Stability of RNA stem-loop structure and distribution of non-random structure in the human immunodeficiency virus (HIV-1). *Nucleic Acids Res* 16: 5153-5168.
- Le SY, Chen J-H, Chatterjee D, Maizel JV (1989) Sequence divergence and open regions of RNA secondary structures in the envelope regions of 17 human immunodeficiency virus isolates. *Nucleic Acids Res* 17: 3275-3288.
- Le SY, Malim MH, Cullen BR, Maizel JV (1990) A highly conserved RNA folding region coincident with the Rev response element of primate immunodeficiency viruses. *Nucleic Acids Res* 18: 1613-1623.
- Arni RK, Ward RJ (1996) Phospholipase A2--a structural review. *Toxicon* 34: 827-41.
- Davidson FF, Dennis EA (1990) Evolutionary relationships and implications for the regulation of phospholipase A2 from snake venom to human secreted forms. *J Mol Evol* 31: 228-238.
- Kordis D, Gubensek F (1996) Ammodytoxin C gene helps to elucidate the irregular structure of Crotalinae group II phospholipase A2 genes. *Eur J Biochem* 240: 83-90.
- Kordis D, Gubensek F (1997) Bov-B long interspersed repeated DNA (LINE) sequences are present in *Vipera ammodytes* phospholipase A2 genes and in genomes of *Viperidae* snakes. *Eur J Biochem* 246: 772-779.
- Lynch VJ (2007) Inventing an arsenal: adaptive evolution and neofunctionalization of snake venom phospholipase A2 genes. *BMC Evol Biol* 7:2
- Kini RM, Evans H J (1989) A model to explain the pharmacological effects of snake venom phospholipases A2. *Toxicon* 27: 613-635.
- Kini RM (2003) Excitement ahead: structure, function and mechanism of snake venom phospholipase A2 enzymes. *Toxicon* 42: 827-840.

29. Ogawa T, Oda N, Nakashima K, Sasaki H, Hattori M, et al. (1992) Unusually high conservation of untranslated sequences in cDNAs for *Trimeresurus flavoviridis* phospholipase A2 isoenzymes. *Proc Natl Acad Sci USA* 89: 8557-8561.
30. Bjorkman PJ, Parham P (1990) Structure, function, and diversity of class I major histocompatibility complex molecules. *Annu Rev Biochem* 59: 253-258.
31. Klein J, O'hUigin C (1994) MHC polymorphism and parasites. *Phil Trans R Soc Lond B* 346: 351-358.
32. Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335: 167-170.
33. Forsdyke DR (1996) Stem-loop potential in MHC genes: a new way of evaluating positive Darwinian selection. *Immunogenetics* 43: 182-189.
34. Hughes AL, Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc Natl Acad Sci USA* 86: 958-962.
35. Hughes AL (1995) Origin and evolution of HLA class I pseudogenes. *Mol Biol Evol* 12: 247-258.
36. Belshaw R, Bensasson D (2006) The rise and fall of introns. *Heredity* 96: 208-213.
37. Darnell Jr JE (1978) Implications of RNA-RNA splicing in evolution of eukaryotic cells. *Science* 202: 1257-1260.
38. Doolittle WF (1978) Genes in pieces: were they ever together? *Nature* 272: 581-582.
39. Gilbert W (1978) Why genes in pieces? *Nature* 271: 501.
40. Hawkins JD (1988) A survey of intron and exon lengths. *Nucleic Acids Res* 16: 9893-9905.
41. Brannan CI, Dees EC, Ingram DS, Tilghman SM (1990) The product of the H19 gene may function as an RNA. *Mol Cell Biol* 10: 28-36.
42. Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, et al. (1992) The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* 71: 515-526.
43. Stoltzfus A, Spencer DF, Zuker M, Logsdon JM, Doolittle WF (1994) Testing the exon theory of genes: the evidence from protein structure. *Science* 265: 202-207.