

Bioinformática aplicada à caracterização de íntrons

Bioinformatics applied to the characterization of introns

Darling de Andrade Lourenço

Universidade Federal de Pelotas, Brasil

darlinglourenco@gmail.com

Altino Branco Choupina

Instituto Politécnico de Bragança, Portugal

albracho@ipb.pt

Resumo

Íntrons são sequências intervenientes de nucleotídeos que estão presentes no genoma. São caracterizadas por dividir os genes em éxons, que são as regiões codificantes. Os íntrons não são codificados no processo de tradução do DNA, uma vez que são removidos num processo anterior, denominado *splicing*. Apesar de não serem sequências codificantes, apresentam funções importantes nos organismos e mutações nessas sequências podem gerar danos, como doenças. Nesse sentido, a bioinformática mostra-se como uma ferramenta promissora na caracterização de íntrons: sítios de *splicing*, junções de éxons, sequências intrônicas em DNA genômico, entre outros. O objetivo deste estudo foi apresentar algumas ferramentas de bioinformática, disponíveis na *web*, para a caracterização de íntrons. Com recurso à literatura científica, quatro ferramentas foram escolhidas para a discussão: DNA *functional site miner*, NNSPLICE, GENSCAN e *Human Splicing Finder*. Essas ferramentas utilizam diferentes algoritmos para a predição sítios de *splicing*, junções íntrons-éxons, predição múltiplos genes em sequências parciais e completas e também para a predição de mutações em íntrons e variantes de genes. Concluímos que as ferramentas se complementam entre si e que depende da situação em que se deseja utilizar. Além disso, as predições *in silico* são relevantes e promissoras, mas baseadas em estatísticas, o que não descarta a necessidade de validação *in vivo*.

Palavras-chave: *Biologia molecular; Biologia computacional; Genômica estrutural.*

Abstract

Introns are intervening sequences of nucleotides that are present in the genome. They are characterized by dividing genes into exons, which is the coding region. Introns are not encoded in the translation process since they are removed in a previous process, called *splicing*. Although they are not coding sequences, they have important functions in organisms and mutations in the sequences can generate damages and lead to diseases. Thus, bioinformatics is a promising tool for the characterization of introns, being able to determine *splicing* sites, junctions of exons, intronic sequences in genomic DNA, among others. This study aims to present some bioinformatics web-based tools available for the characterization

of introns. Resorting to the available scientific literature, we choose four bioinformatics tools for discussion: DNA functional site miner, NNSPLICE, GENSCAN, and Human Splicing Finder. These tools are developed from different algorithms for splicing sites prediction, introns-exons junctions, multiple genes prediction in partial or complete DNA sequences and also for prediction of mutation in introns e genes variants. Therefore, we conclude that these bioinformatic tools are complementary and the selection of one is dependent on the desired application. Besides, *in silico* predictions are relevant and promising, but statistic-based, which does not discard the need for *in vivo* validation.

Keywords: *Molecular biology; Computational biology; Structural genomic.*

INTRODUÇÃO

Genes são sequências de ácido desoxirribonucleico (DNA) e constituem o genoma dos organismos. Os genes podem ser transcritos e traduzidos em peptídeos que darão origem a uma proteína ou podem ser sequências que não codificam para peptídeos (NIH, 2019). De acordo com o dogma central da biologia molecular, um gene é transcrito a ácido ribonucleico mensageiro (mRNA) que é então traduzido a peptídeos, que se organizarão e darão origem a um polipeptídio ou proteína (Lewin, 2008).

Os genes dos organismos eucariontes apresentam regiões codificadoras e reguladoras bem mais complexas do que os genes dos organismos procariontes. Como se pode observar na Figura 1, os genes de eucariontes são compostos por uma região reguladora a montante da região codificadora, composta pelo promotor (sítio de ligação da RNA-polimerase); uma região codificadora e uma região terminadora. Como característica singular, os genes eucariotos têm sua região codificadora constituída por sequências intervenientes, os íntrons, que interrompem a região transcrita do gene dividindo-o em vários segmentos, os éxons. As regiões íntrônicas são traduzidas para o pré-mRNA, entretanto, são removidas do mRNA através do processo denominado splicing de RNA (Zaha, Ferreira & Passaglia, 2000). O processo de splicing apenas é possível devido às informações contidas nos íntrons, os denominados sítios de reconhecimento de união, que unem os éxons sem alterações na ordem original (Lewin, 2008).

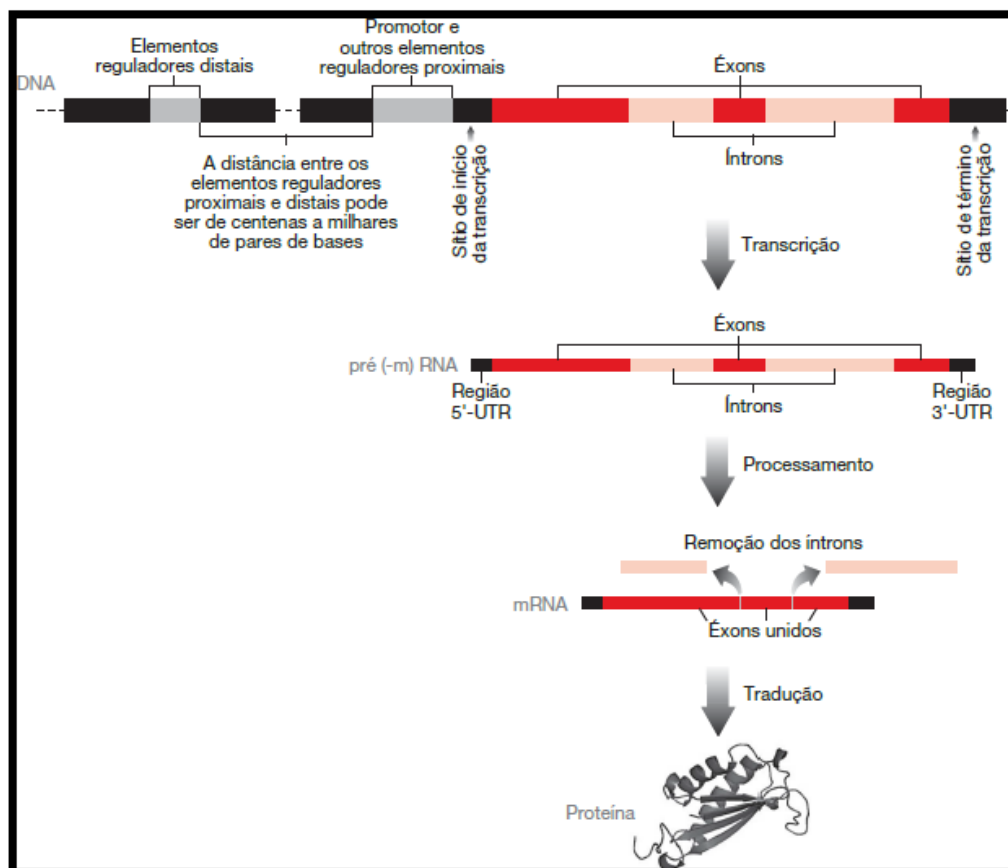


Figura 1 - Estrutura do gene eucarioto.

Existem duas hipóteses que tentam explicar a origem dos íntrons: *introns-early* e *introns-late*. Na primeira, é assumido que os íntrons já existiam nos genes do ancestral comum entre procariontes e eucariontes e que foram perdidos nos procariotos devido à pressão seletiva de replicação rápida do DNA. Já na segunda, assume-se que os íntrons foram inseridos após a divergência entre procariontes e eucariontes durante a evolução (Acharya, 2006; Deusdado, 2008; Lewin, 2008).

O processo de *splicing* é subdividido em 4 categorias, de acordo com suas características estruturais e mecanismos: *splicing* nos spliceossomas, *auto-splicing* (Grupo I e II), *splicing* de RNA transportador (tRNA) e *splicing* alternativo. É importante destacar que as categorias de *splicing* possuem distribuição filogenética distinta (Irimia & Roy, 2014).

O *splicing* nos spliceossomas são encontrados em todos os genomas eucarióticos e é caracterizado pelo seu mecanismo de remoção de íntrons que é catalisado no spliceossoma, uma maquinaria ribonucleoproteica complexa formada por 5 RNAs pequenos nucleolares (snRNA,

small nuclear RNA), U1, U2, U4, U5 e U6 snRNAs e mais de 200 proteínas (Andersson *et al.*, 2007; Lane *et al.*, 2007; Wahl, Will, & Lührmann, 2009).

O *auto-splicing* de grupo I encontra-se em representantes da maioria dos eucariontes, tanto em genes de RNA ribossomal (rRNA) nuclear quanto em genes de plastídeos, mitocôndrias e algumas bactérias e vírus e em protozoários (Haugen, Simon, & Bhattacharya, 2005; Lewin, 2008). O *splicing* de tRNA ocorre no núcleo de todos os eucariontes e no genoma de organismos do domínio Archaeae (Irimia & Roy, 2014; Marck & Grosjean, 2003; Randau & Söll, 2008). O processamento de remoção dos íntrons difere um pouco nessa categoria visto que o mesmo é removido por enzimas ao invés de riboenzimas (Lewin, 2008; Randau & Söll, 2008). O *auto-splicing* do grupo II encontram-se nos genomas bacterianos, de cloroplastos e mitocondriais de eucariontes que divergiram e acredita-se que os mesmos datam de antes da origem dos eucariontes (Candales *et al.*, 2012; Ferat & Michel, 1993; Koonin, 2006; Lambowitz & Zimmerly, 2004). As reações de remoção são catalisadas pelo próprio RNA intrônico, que possui atividade riboenzimática (Chan *et al.*, 2018).

O *splicing* alternativo é um processo que ocorre nos eucariontes complexos e é o mesmo que permite que esses organismos possuam uma enorme variedade de proteínas que não se encontra correspondida no genoma, ou seja, os eucariontes possuem um número pequeno de genes codificantes quando comparado com o número de diferentes proteínas que existem. Isso porque durante o *splicing* alternativo os éxons são recombinados de diversas formas, gerando uma variabilidade de mRNA que codificarão para isoformas de proteínas a partir de um único gene (Lewin, 2008). Além disso, o *splicing* alternativo também pode gerar variações de transcritos não funcionais que atuam no *down-regulation* de expressão gênica (Bingham, Chou, Mims & Zachar, 1988).

Apesar de durante muito tempo se ter acreditado que os íntrons eram apenas sequências intercalantes, que além de não originarem um RNA funcional também não possuíam outras funções, hoje sabe-se que os íntrons desempenham funções importantes para a manutenção da homeostase nos organismos. Algumas das funções desempenhadas pelos íntrons são a expressão de genes para RNA não-codificantes (ncRNA) como microRNA (mRNA) e RNA nucleolar pequeno (snoRNA); desempenhar o papel de reguladores de transcrição tanto por ser um fator limitante no *splicing* cotranscricional quanto por agir indiretamente, através do ancoramento de elementos regulatórios de DNA (Irimia & Blencowe, 2012; Maeso *et al.*, 2012; Patel, McCarthy & Steitz, 2002).

Além disso, mutações nas sequências dos íntrons também geram distúrbios nos organismos, como doenças. Fibrose cística, β -talassemia, esclerose lateral amiotrófica 1 e distrofia muscular de Duchene são doenças humanas que estão associadas com mutações intrônicas profundas, que resultam na alteração do processo de *splicing* (Vaz-Drago, Custódio & Carmo-Fonseca, 2017). Entretanto, entre as doenças, as mutações intrônicas são muito mais estudadas no cancro, estando associada com cancro de bexiga, faringe e pulmão (Ward & Cooper, 2010).

Devido ao papel desempenhado pelos íntrons nos organismos, o seu estudo e caracterização tornam-se de suma importância para diversas finalidades, como genômica estrutural e farmacogenômica. Nesse sentido, as ferramentas de bioinformática são muito promissoras para auxiliar nessa caracterização de sequências intrônicas e suas alterações (Floreza, 2006). Sendo assim, o objetivo deste trabalho foi realizar um levantamento de ferramentas bioinformáticas que possam ser utilizadas na caracterização de íntrons, como plataformas de *web* e *softwares*, que possam ser utilizados para determinar as características dos íntrons.

DESENVOLVIMENTO

Para o desenvolvimento deste trabalho foram realizadas pesquisas na literatura científica sobre íntrons, utilizando artigos disponíveis nas bases de dados PubMed-NCBI (National Center for Biotechnology and Information, www.ncbi.nlm.nih.gov/pubmed) e ScienceDirect (<http://www.sciencedirect.com/>).

Para as informações na área de bioinformática, foram realizadas pesquisas na plataforma Google (www.google.com.br) com as palavras chave: *bioinformatics*, *introns*, *prediction*, *splicing*. Algumas das plataformas *web* existentes que foram utilizadas estão descritas na Tabela 1. A seleção dessas plataformas fez-se com recurso a literatura científica disponível (Desmet, Hamroun, Collod-Bérout, Claustres & Beroud, 2010; Valle-Aviles, Valentin-Berrios, Gonzalez-Mendez & Rodriguez-Del Valle, 2007).

RESULTADOS E DISCUSSÃO

Algumas das plataformas *web* encontrados estão listadas na Tabela 1 abaixo.

Tabela 1 – Plataformas web de caracterização de íntrons.

Plataforma	Método	Endereço
DNA <i>functional site miner</i> (DNAFSMiner)	Máquina de vetor de suporte	http://dnafsmineer.bic.nus.edu.sg/
NNSPLICE	Redes neurais	https://www.fruitfly.org/seq_tools/splice.html
GENSCAN	Decomposição Máxima de Dependência	http://hollywood.mit.edu/GENSCAN.html
<i>Human Splicing Finder</i> (HSF)	Matriz posição-peso	http://www.umd.be/HSF/

O DNA *functional site miner* (DNAFSMiner) é um *software* para a *web* que reconhece sítios funcionais em sequências de DNA (Liu, Han, Li & Wong, 2005). Esse *software* está dividido em duas ferramentas: a *TIS Miner* que é utilizada para a predição de sítios de iniciação da tradução em sequências de DNA, mRNA e cDNA e a *Poly(A) Signal Miner* que é utilizada para predizer os sinais de poliadenilação (sítios poliA) em sequências de DNA humano. Essas ferramentas são úteis na identificação e confirmação indireta de sítios específicos em sequências genômicas analisadas (Olivier *et al.*, 2008; Sekyere, Dunn & Richardson, 2005).

O NNSPLICE é uma ferramenta de predição de sítios de *splicing* desenvolvida por Reese (1997). Essa ferramenta fornece um método de reconhecimento de sítio de *splicing* através da análise da estrutura dos sítios doadores e sítios receptores utilizando redes neurais que reconhecem cada um dos sítios (Reese, 1997). O NNSPLICE têm sido usado em investigações relacionadas a mutações nos íntrons de oncogenes em alguns tipos de cancro, como o melanoma e, em doenças autossômicas recessivas, como a microesferofacia (Alías *et al.*, 2018; Rodrigues *et al.*, 2018).

O GENSCAN é uma ferramenta *web* que identifica estruturas completas de íntrons e éxons em sequências de DNA genômico (Burge & Karlin, 1997). O programa é modelado de acordo com os modelos ocultos de Markov e, além de predizer as estruturas de íntrons e éxons, também prediz genes múltiplos em uma sequência parcial ou completa (Burge & Karlin, 1997). O GENSCAN têm sido utilizado para diversas finalidades, entre elas, a caracterização e anotação de genes em genomas rascunhos (C. Li *et al.*, 2018), investigações que visam novas formas de controle biológico de pragas aplicadas, por exemplo, a Dengue, utilizando mecanismos de biologia molecular (Serrato-Salas *et al.*, 2018). Além disso, também tem sido utilizado para a

análises de sequências codificadoras e relações evolutivas entre essas sequências em diferentes espécies (J. Li, Li, & Lu, 2018).

O *Human Splicing Finder* (HSF) é uma ferramenta capaz de prever mutações em sinais de *splicing* e também de identificar motivos de *splicing* apenas em sequências humanas, o que o difere das outras ferramentas (Desmet *et al.*, 2009). O HSF tem sido usado em investigações a cerca de sequências humanas como por exemplo, na identificação de alterações nos sítios de *splicing* causadas por variantes deletérias em oncogenes envolvidos no cancro de próstata (Paulo *et al.*, 2018) e em sequências de canais de sódio relacionados a síndrome de morte infantil (Männikkö *et al.*, 2018). Além disso, também é utilizado na identificação e caracterização de variantes intrônicas em genes relacionados ao cancro de mama (Caputo *et al.*, 2018).

CONCLUSÃO

No presente trabalho apresentaram-se algumas ferramentas baseadas na *web* para a caracterização de sequências intrônicas em sequências de DNA. As ferramentas apresentadas realizam a predição de diversas características presentes em íntrons e éxons, como a predição de sítios de *splicing* em sequências parciais ou completas, análise de sítios doadores e receptores de sequências e análises de mutações e variantes em sequências intrônicas.

É importante ressaltar que cada ferramenta utiliza diferentes métodos de predição e que, apesar dos resultados serem passíveis de comparação, cada ferramenta complementa a outra. A escolha da ferramenta também depende da situação em que se deseja emprega-la. Além disso, as predições *in silico* são relevantes e promissoras, mas baseadas em estatísticas, o que não descarta a necessidade de validação *in vivo*.

Referências

- Acharya, S. (2006). *Some aspects of physicochemical properties of DNA and RNA*. Uppsala Universitet.
- Alías, L., Crespi, J., González-Quereda, L., Téllez, J., Martínez, E., Bernal, S., & Gallano, M. P. (2018). Next-generation sequencing reveals a new mutation in the LTBP2 gene associated with microspherophakia in a Spanish family. *BMC Medical Genetics*, 19(1), 1-8.
- Andersson, J. O., Sjögren, Å. M., Horner, D. S., Murphy, C. A., Dyal, P. L., Svård, S. G., & Roger, A. J. (2007). A genomic survey of the fish parasite *Spironucleus salmonicida* indicates genomic plasticity among diplomonads and significant lateral gene transfer in eukaryote genome evolution. *BMC Genomics*, 8(1), 1-8.
- Bingham, P. M., Chou, T. B., Mims, I., & Zachar, Z. (1988). On/off regulation of gene expression at the level of splicing. *Trends in Genetics : TIG*, 4(5), 134-138.

- Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268(1), 78–94.
- Candales, M. A., Duong, A., Hood, K. S., Li, T., Neufeld, R. A. E., Sun, R., & Zimmerly, S. (2012). Database for bacterial group II introns. *Nucleic Acids Research*, 40, D187–90.
- Caputo, S. M., Léone, M., Damiola, F., Ehlen, A., Carreira, A., Gaidrat, P., & Rouleau, E. (2018). Full in-frame exon 3 skipping of BRCA2 confers high risk of breast and/or ovarian cancer. *Oncotarget*, 9(25), 17334–17348.
- Chan, R. T., Peters, J. K., Robart, A. R., Wiryaman, T., Rajashankar, K. R., & Toor, N. (2018). Structural basis for the second step of group II intron splicing. *Nature Communications*, 9(1), 1–10.
- Desmet, F. O., Hamroun, D., Collod-Bèroud, G., Claustres, M., & Bèroud, C. (2010). Bioinformatics identification of splice site signals and prediction of mutation effects. *Research Advances In Nucleic Acids Research*, 1–14.
- Desmet, F. O., Hamroun, D., Lalande, M., Collod-Bèroud, G., Claustres, M., & Bèroud, C. (2009). Human Splicing Finder: An online bioinformatics tool to predict splicing signals. *Nucleic Acids Research*, 37(9), 2–14.
- Deusdado, S. (2008). *Análise e Compressão de Sequências Genômicas*. Universidade do Minho.
- Ferat, J.-L., & Michel, F. (1993). Group II self-splicing introns in bacteria. *Nature*, 364(6435), 358–361.
- Florea, L. (2006). Bioinformatics of alternative splicing and its regulation. *Briefings in Bioinformatics*, 7(1), 55–69.
- Haugen, P., Simon, D. M., & Bhattacharya, D. (2005). The natural history of group I introns. *Trends in Genetics*, 21(2), 111–119.
- Irimia, M., & Blencowe, B. J. (2012). Alternative splicing: decoding an expansive regulatory layer. *Current Opinion in Cell Biology*, 24(3), 323–332.
- Irimia, M., & Roy, S. W. (2014). Origin of spliceosomal introns and alternative splicing. *Cold Spring Harbor Perspectives in Biology*, 6(6), 1–22.
- Koonin, E. V. (2006). The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biology Direct*, 1(1), 1–22.
- Lambowitz, A. M., & Zimmerly, S. (2004). Mobile Group II Introns. *Annual Review of Genetics*, 38(1), 1–35.
- Lane, C. E., van den Heuvel, K., Kozera, C., Curtis, B. A., Parsons, B. J., Bowman, S., & Archibald, J. M. (2007). Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. *Proceedings of the National Academy of Sciences*, 104(50), 19908–19913.
- Lewin, B. (2008). *Genes IX*. Burlington: Jones & Bartlett Learning.
- Li, C., Liu, X., Liu, B., Ma, B., Liu, F., Liu, G., & Wang, C. (2018). Draft genome of the Peruvian scallop *Argopecten purpuratus*. *GigaScience*, 7(4), 1–6.
- Li, J., Li, C., & Lu, S. (2018). Systematic analysis of DEMETER-like DNA glycosylase genes shows lineage-specific Smi-miR7972 involved in SmDML1 regulation in *Salvia miltiorrhiza*. *Scientific Reports*, 8(1), 1–13.
- Liu, H., Han, H., Li, J., & Wong, L. (2005). DNAFSMiner: A web-based software toolbox to recognize two types of functional sites in DNA sequences. *Bioinformatics*, 21(5), 671–673.
- Maeso, I., Irimia, M., Tena, J. J., Gonzalez-Perez, E., Tran, D., Ravi, V., ... Garcia-Fernandez, J. (2012). An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement. *Genome Research*, 22(4), 642–655.
- Männikkö, R., Wong, L., Tester, D. J., Thor, M. G., Sud, R., Kullmann, D. M., ... Matthews, E. (2018). Dysfunction of NaV1.4, a skeletal muscle voltage-gated sodium channel, in sudden infant death syndrome: a case-control study. *The Lancet*, 391(10129), 1483–1492.
- Marck, C., & Grosjean, H. (2003). Identification of BHB splicing motifs in intron-containing tRNAs

- from 18 archaea: evolutionary implications. *RNA (New York, N.Y.)*, 9(12), 1516–1531.
- NIH. (2019). What is a gene? - Genetics Home Reference - NIH. Retrieved February 26, 2019, from <https://ghr.nlm.nih.gov/primer/basics/gene>
- Olivier, V., Blanchard, P., Chaouch, S., Lallemand, P., Schurr, F., Celle, O., & Ribière, M. (2008). Molecular characterisation and phylogenetic analysis of Chronic bee paralysis virus, a honey bee virus. *Virus Research*, 132(1–2), 59–68.
- Patel, A. A., McCarthy, M., & Steitz, J. A. (2002). The splicing of U12-type introns can be a rate-limiting step in gene expression. *The EMBO Journal*, 21(14), 3804–3815.
- Paulo, P., Maia, S., Pinto, C., Pinto, P., Monteiro, A., Peixoto, A., & Teixeira, M. R. (2018). Targeted next generation sequencing identifies functionally deleterious germline mutations in novel genes in early-onset/familial prostate cancer. *PLOS Genetics*, 14(4), 1-18.
- Randau, L., & Söll, D. (2008). Transfer RNA genes in pieces. *EMBO Reports*, 9(7), 623–628.
- Reese, M. G. (1997). Improved splice site detection in Genie. *Journal of Computational Biology*, 4(3), 311–323.
- Rodrigues, M., Mobuchon, L., Houy, A., Fiévet, A., Gardrat, S., Barnhill, R. L., & Stern, M. H. (2018). Outlier response to anti-PD1 in uveal melanoma reveals germline MBD4 mutations in hypermutated tumors. *Nature Communications*, 9(1), 1-6.
- Sekyere, E. O., Dunn, L. L., & Richardson, D. R. (2005). Examination of the distribution of the transferrin homologue, melanotransferrin (tumour antigen p97), in mouse and human. *Biochimica et Biophysica Acta - General Subjects*, 1722(2), 131–142.
- Serrato-Salas, J., Hernández-Martínez, S., Martínez-Barnette, J., Condé, R., Alvarado-Delgado, A., Zumaya-Estrada, F., & Lanz-Mendoza, H. (2018). De Novo DNA synthesis in *Aedes aegypti* midgut cells as a complementary strategy to limit dengue viral replication. *Frontiers in Microbiology*, 9, 1-12.
- Valle-Aviles, L., Valentin-Berrios, S., Gonzalez-Mendez, R. R., & Rodriguez-Del Valle, N. (2007). Functional, genetic and bioinformatic characterization of a calcium/calmodulin kinase gene in *Sporothrix schenckii*. *BMC Microbiology*, 7, 1-12.
- Vaz-Drago, R., Custódio, N., & Carmo-Fonseca, M. (2017). Deep intronic mutations and human disease. *Human Genetics*, 136(9), 1093–1111.
- Wahl, M. C., Will, C. L., & Lührmann, R. (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell*, 136(4), 701–718.
- Ward, A. J., & Cooper, T. A. (2010). The pathobiology of splicing. *The Journal of Pathology*, 220(2), 152–163.
- Zaha, A., Ferreira, H., & Passaglia, L. M. P. (2000). *Biologia Molecular Básica*. Porto Alegre: Artmed.