**UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ**

**DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO**

**PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA**

**RODOLPHO VALENTINI JUNIOR**

# CODANCIRC: FERRAMENTA PARA PREDIÇÃO DE REGIÕES CODIFICADORAS DE PROTEÍNAS EM RNA CIRCULARES EUCARIÓTICOS UTILIZANDO MODELOS OCULTOS GENERALIZADOS DE MARKOV

**DISSERTAÇÃO**

**CORNÉLIO PROCÓPIO**

**2023**

**RODOLPHO VALENTINI JUNIOR**

# CODANCIRC: FERRAMENTA PARA PREDIÇÃO DE REGIÕES CODIFICADORAS DE PROTEÍNAS EM RNA CIRCULARES EUCARIÓTICOS UTILIZANDO MODELOS OCULTOS GENERALIZADOS DE MARKOV

## CodAnCirc: Tool for predicting protein coding regions in eukaryotic circular RNA using Generalized Hidden Markov Models.

Dissertação apresentada como requisito para obtenção do título de apresentado como requisito para obtenção do título de Mestre em Bioinformática do Programa de Pós-graduação em Bioinformática da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. André Yoshiaki Kashiwabara

Coorientador: Prof. Dr. Alexandre Rossi Paschoal

**CORNÉLIO PROCÓPIO**

**2023**

**Ministério da Educação**
**Universidade Tecnológica Federal do Paraná**
**Campus Cornélio Procópio**

RODOLPHO VALENTINI JUNIOR

**CODANCIRC: FERRAMENTA PARA PREDIÇÃO DE REGIÕES CODIFICADORAS DE PROTEÍNAS EM RNA CIRCULARES EUCARIÓTICOS UTILIZANDO MODELOS OCULTOS GENERALIZADOS DE MARKOV**

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Bioinformática.

Data de aprovação: 03 de Julho de 2023

Dr. Andre Yoshiaki Kashiwabara, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Alexandre Rossi Paschoal, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Laurival Antonio Vilas Boas, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Rogerio Fernandes De Souza, Doutorado - Universidade Estadual de Londrina (Uel)

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 03/07/2023.

Dedico este trabalho a minha mãe, Sra. Nilce Petean Valentini, que sempre me estimulou a prosseguir em meus estudos, apesar dos impedimentos da vida. Por ironia, veio a falecer no mesmo dia, poucas horas antes, que soubesse o resultado de meu ingresso no Mestrado em Bioinformática. Hoje, onde quer que esteja, certamente estará feliz com meu desafio, em sua homenagem, de continuar a estudar até o último batimento de meu coração.

## AGRADECIMENTOS

Em primeiro lugar externo meus sinceros agradecimentos aos meus pais que me deram a vida, aos inúmeros professores ao longo de minha existência, aos professores da UTFPR do campus de Londrina, em especial o Prof. Paulo de Tarso que me apoiou e me incentivou a prosseguir meus estudos, e aos professores deste campus, em especial o Prof. André e ao Prof. Paschoal, respectivamente orientador e coorientador, aos demais professores da UTFPR e das demais instituições conveniadas, Prof. Dr. Alan Mitchell Durham( IME/USP), Profª. Dra. Edna Maria Vissoci Reiche (UEL), Prof. Dr. Thiago Estevam Parente Martins(FIOCRUZ) e Prof. Dr. Bruno Thiago de Lima Nichio, e também a todos os colegas de curso, em especial à doutoranda Nayane de Souza(UFPR/UTFPR) e aos doutorandos Denilson Fagundes Barbosa(UFPR/UTFPR) e Murilo Caminotto Barbosa(UFPR/UTFPR) que benevolentemente auxiliaram sobremaneira, com suas intervenções precisas, o desenvolvimento deste trabalho.

"Em tempos de incertezas como os que estamos vivenciando, é tudo muito dinâmico e intenso. Porém, a cada dia, podemos definir se seremos os protagonistas da nossa vitória ou apenas coadjuvantes ou espectadores de toda essa história."
(Daisaku Ikeda)

## RESUMO

Neste trabalho, abordamos a predição de regiões codificadoras em transcritos eucarióticos de RNA circular, aplicando o banco de dados TransCirc e Modelos Ocultos Generalizados de Markov (GHMM). Para esta finalidade, desenvolvemos o CodAnCirc, uma versão adaptada e otimizada para RNA circulares da ferramenta CodAn (Coding sequence Annotator). A partir de uma seleção de 15 mil RNA circulares de um conjunto total de 320 mil, nos concentramos em exemplares que continham uma única região codificadora de proteínas. Utilizando ferramentas computacionais, preparamos a amostra e conduzimos a análise da tradução de RNA circular. Avaliamos o desempenho das previsões de tradução empregando métricas de acurácia, precisão, revocação e a pontuação F1. Na análise comparativa entre CodAnCirc (GHMM) e FragGeneScan, que utiliza Modelos Ocultos de Markov (HMM), observamos que o CodAnCirc alcançou uma pontuação F1 de 83%, em comparação com 16% do FragGeneScan. Estes resultados indicam que o uso de GHMM, como implementado no CodAnCirc, poderia ser uma abordagem eficaz para a predição de regiões codificadoras em RNA circulares eucarióticos.



**Palavras-chave:** rna circular; predição de regiões codificadora; transcirc; codancirc; ghmm.

**ABSTRACT**

In this work, we approach the prediction of coding regions in eukaryotic circular RNA transcripts, applying the TransCirc database and Generalized Hidden Markov Models (GHMM). For this purpose, we developed CodAnCirc, an adapted and optimized version for circular RNA of the original CodAn tool (Coding sequence Annotator), which was conceived for linear RNA. From a selection of 15 thousand circular RNA from a total set of 320 thousand, we focused on specimens that contained a single protein-coding region. Using computational tools, we prepared the sample and conducted the analysis of circular RNA translation. We evaluated the performance of translation predictions using accuracy, precision, recall, and F1 score metrics. In the comparative analysis between CodAnCirc (GHMM) and FragGeneScan, which uses Hidden Markov Models (HMM), we observed that CodAnCirc achieved an F1 score of 83%, compared to 16% for FragGeneScan. These results indicate that the use of GHMM, as implemented in CodAnCirc, could be an effective approach for predicting coding regions in eukaryotic circular RNA.

**Keywords:** circular rna.; prediction of coding region; transcirc; codancirc; ghmm.

## LISTA DE ILUSTRAÇÕES

## LISTA DE TABELAS

**SUMÁRIO**

# 1 INTRODUÇÃO

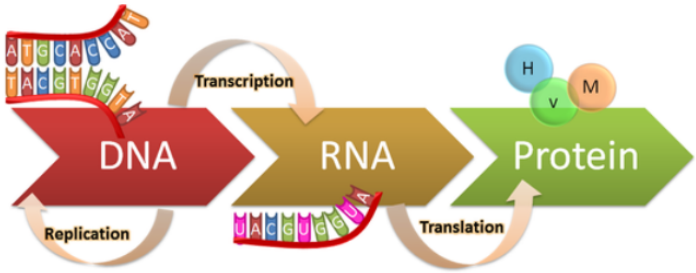
Os ácidos nucleicos são moléculas muitíssimo importantes na biologia, pois são responsáveis por armazenar e transmitir a informação genética de uma célula para outra. Existem dois tipos principais de ácidos nucleicos: o DNA (ácido desoxirribonucleico) e o RNA (ácido ribonucleico)[1, 2].

O RNA é uma molécula geralmente menor e menos estável do que o DNA, e é encontrado em diversas formas e funções no organismo. Uma dessas formas é o RNA linear, que é uma molécula de RNA com uma estrutura simples, sem qualquer ligação covalente entre os extremos. Ele é encontrado em vários tipos de células e pode ter funções diferentes, como a síntese de proteínas, a regulação da expressão do gene, e a replicação do DNA. Outra forma que podemos identificar o RNA é circular, qual é uma molécula de RNA com uma estrutura circular, formada pela ligação covalente entre os extremos[1, 2].

Os RNA circulares são gerados por meio de um processo chamado *back-splicing*, no qual os exons se ligam fora da ordem genômica original, formando uma estrutura circular. Isso pode resultar em uma maior variabilidade nas sequências de nucleotídeos dos RNA circulares, o que pode ampliar as possibilidades de proteínas viáveis produzidas a partir desses RNA[1, 2].

Há décadas propaga-se pelo Dogma Central da Biologia, que RNA canônicos, aqui compreendidos lineares, traduzem proteínas, conforme ilustrado na Figura 1. Mas como disse Dobzhansky em seu artigo de 1973, "Nothing in Biology Makes Sense Except in the Light of Evolution"[3]. Nessa linha de pensamento, para a surpresa de muitos, pesquisas mais recentes têm demonstrado que os RNAs circulares, sob determinadas condições, podem produzir incríveis 811% mais proteína do que os RNA lineares com *cap* 5' e cauda poliadelilada[4, 5]. Essa descoberta revela um aspecto intrigante, pouco lógico, e prafraseando Dobzhansky, até mesmo sem sentido quanto ao funcionamento dessas moléculas.

**Figura 1 – Dogma Central da Biologia**




**Fonte: 6 A figura ilustra o Dogma Central da Biologia de forma clássica, mas há outras representações, incluindo avanços da Biologia Molecular contemporânea. Contudo, ainda não se incluíram as características do RNA circular, o que sugere que novas revisitas advirão.**

Neste trabalho, abordamos a escassez de ferramentas em bioinformática para prever genes e proteínas em RNA circulares, propondo a criação e avaliação da ferramenta CodAnCirc.

Utilizamos o CodAn[7], voltado para RNA lineares, e seu componente flexível ToPs[8], permitindo adaptar o modelo GHMM[8] e criar o VERT_circ. Assim, desenvolvemos o CodAnCirc, uma versão específica da ferramenta CodAn original.

## 1.1 OBJETIVO GERAL

O objetivo geral deste trabalho é validar a ferramenta CodAnCirc para a previsão precisa e confiável de regiões codificadoras de proteínas em RNA circulares eucarióticos, como uma adaptação da ferramenta CodAn(Coding sequence Annotator)[7], utilizando Modelos Ocultos Generalizados de Markov, GHMM(*Generalized Hidden Markov Model*).

## 1.2 OBJETIVOS ESPECÍFICOS

- Reconhecer regiões codificadoras em transcritos eucarióticos de RNA circular, valendo-se de uma metodologia baseada em modelo probabilístico GHMM (Generalized Hidden Markov Model), configurado pelo framework probabilístico ToPS[8], qual integra a ferramenta CodAn(Coding sequence Annotator)[7] utilizada a princípio para RNA lineares com extremidades 5' e 3'.

- Adaptar a ferramenta CodAn[7] através da criação de um modelo específico para RNA circulares, nomeando-a de CodAnCirc.

- Comparar as performances entre si das ferramentas CodAnCirc adptada do CodAn[7], e FragGeneScan[9] (qual também utiliza Modelos Ocultos de Markov - HMM), para efeito de validação dos resultados obtidos.

## 1.3 JUSTIFICATIVA

Diante da escassez de ferramentas em bioinformática na predição de genes/proteínas em RNA circulares, propusemos, neste trabalho, a criação e avaliação de desempenho da ferramenta CodAnCirc, desenvolvida para RNA circulares, considerados os prováveis maestros da regulação gênica, incluindo seu papel protagonista como biomarcadores de doenças humanas[4].

## 1.4 ESTRUTURA DO TRABALHO

A estrutura do trabalho organizamos da seguinte forma:

No início, apresentamos nossa investigação para identificar regiões codificadoras em transcrições de RNA circular usando o banco de dados TransCirc. Explicamos como empregamos duas ferramentas na análise dos dados: CodAnCirc (uma versão adaptada do CodAn[7] para RNA circular e FragGeneScan[9], quais são todas baseadas em modelos probabilísticos, ou seja Modelos Ocultos de Markov (GHMM e HMM, respectivamente).

Em seguida, discutimos os objetivos gerais e específicos do trabalho, bem como a justificativa e a estrutura do estudo. Comparamos nossa pesquisa a trabalhos relacionados, mencionando a ferramenta CirCode[10], que utiliza FragGeneScan, e fornecemos uma visão geral da ferramenta CodAn.

Focamos no banco de dados TransCirc[11], que fundamenta nossa pesquisa, e passamos à fundamentação teórica sobre RNA circulares, destacando as semelhanças e diferenças na tradução entre circulares e lineares.

Prosseguimos com uma discussão detalhada sobre os Modelos Ocultos Generalizados de Markov, sem esgotar o tema. Com essa base de sustentação teórica partimos para a metodologia utilizada começando pela formação da amostra, os *scripts* em *Python* criados e a adaptação do modelo `VERT_full` para o `VERT_circ` qual possibilitou o surgimento de uma nova ferramenta específica para RNA circulares: O CodAnCirc, ainda na fase embrionária, mas já apresentando os primeiros resultados práticos.

Nesse passeio pela estrutura do trabalho chegamos a conclusão após concretos resultados positivos.

Por fim especulamos sobre o alcance dos objetivos dentro de nossas limitações, e sobre as recomendações de trabalhos futuros frente ao avanço da Inteligência Artificial no nosso cotidiano.

# 2 TRABALHOS RELACIONADOS

Inicialmente o artigo base, qual inspirou em grande parte este trabalho, foi uma revisão sobre as publicações de aplicativos de anotação e identificação de RNA circulares, considerando seus tipos e funções biológicas (*The bioinformatics toolbox for discovery and analysis*) [12]. Na seção a seguir discorremos sucintamente sobre os trabalhos que mais contribuíram para a efetivação de nossa pesquisa, sejam ferramentas de predição de proteínas ou bancos de dados de RNA circulares.

## 2.1 CIRCCODE, FRAGGENESCAN E OUTRAS FERRAMENTAS

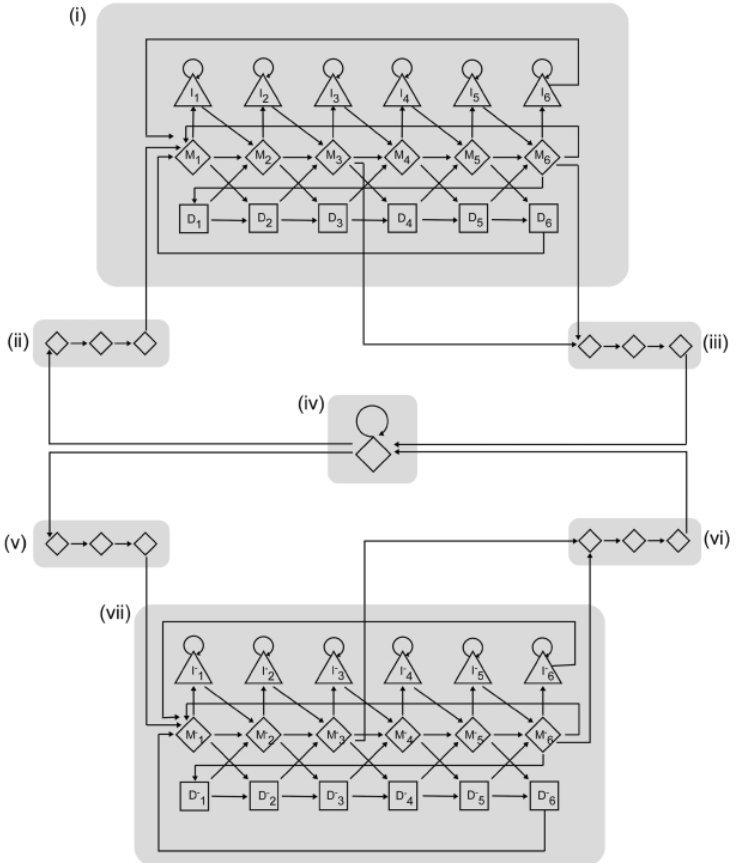
Num primeiro momento destacou-se a publicação *CircCode: A Powerful Tool for Identifying Coding Ability*[10]. E ainda nesse sentido, vale a pena ressaltar que o CircCode vai muito além das ferramentas para a previsão e identificação de RNA circulares, como CIRI[13], CIRCexplorer [14], CircPro[15], Circtools[16] entre outras pelo fato de estender seu escopo ao potencial de tradução de proteínas, e portanto contribuiu para a criação de uma massa crítica neste nosso trabalho. Analisando o CircCode mais à miude, verifica-se que se vale do uso do BASiNET[17], qual se fundamenta em aprendizado de máquina (*random forest*) para classificar RNA obtendo dados positivos, os RNA codificadores, e dados negativos, os RNA não codificantes. Esses dados, tanto positivos quanto negativos passam a compor redes complexas, e destas se extraem as medidas topológicas de um vetor de características. Após treinamento do modelo, usa-se para classificar a capacidade de codificação de RNA circulares. Mas o ponto mais importante dentro do ponto de vista do escopo deste trabalho, salienta-se pelo fato que o CircCode utiliza o FragGeneScan[9], qual emprega um modelo probalístico, tipo Modelo Oculto de Markov - HMM. Dessa forma, CircCode vai muito além das ferramentas para a previsão e identificação de RNA circulares, como CIRI[13], CIRCexplorer [14], CircPro[15], circtools[16] entre outras pelo fato de estender seu objetivo ao potencial de tradução de proteínas.

Os resultados alcançados pelo CircCode resumem-se aos de outras poucas ferramentas que estão disponíveis para identificar o potencial de tradução de RNA circulares. Na conclusão, os autores do artigo (*CircCode: a powerful tool for identifying coding ability*) afirmam que a ferramenta linha de comando CircCode é altamente sensível para identificar transcritos de RNA circulares a partir de sequencias de Ribo-seq com alta acurácia[10]. Tal feito é devido a conter a FragGeneScan, e portanto foi determinante que a escolhêssemos para termos de comparação com as nossas ferramentas. A seguir mais características da FragGeneScan.

## 2.2 FRAGGENESCAN

O FragGeneScan é construído sobre um modelo oculto de Markov (HMM), conforme ilustrado na Figura 2, que incorpora viés de uso de códons, modelos de erros de sequenciamento e padrões de códons de início/parada em um modelo unificado. Dada uma leitura curta (ou um genoma completo), o problema de predição do gene é encontrar o melhor caminho de estados ocultos com maior probabilidade de gerar a sequência de nucleotídeos observada, que pode ser resolvida pelo algoritmo de Viterbi [9]. E assim, o FragGeneScan relata os genes que atendem às três condições a seguir: (i) o comprimento dos genes é maior que 60 pb, (ii) os genes começam em um estado inicial (códon inicial) ou em um estado correspondente (região interna dos genes) e (iii) os genes terminam em um estado *stop* (*stop codon*) ou em um *match state* (região interna dos genes) [9]. Portanto, o FragGeneScan pode prever genes completos, bem como genes parciais (fragmentados) sem códons de início e/ou parada [9].

**Figura 2 – Modelo HMM do FragGenScan**




**Fonte: [9] O HMM do FragGeneScan trabalha com sete superestados. Superestados representam regiões gênicas por caixas sombreadas: (i) códons de início (ii) e códons de parada (iii) para ambas as fitas para frente (i – iii) e para trás (v – vii), e regiões que não codificam (iv). Os estados das regiões gênicas (ie vii) consistem em seis estados de correspondência consecutivos representados por losangos, estados de inserção por triângulos e estados de deleção por quadrados, que juntos correspondem a um HMM não homogêneo de seis períodos.**

## 2.3 CODAN

Em termos de identificação de regiões codificadoras em transcritos eucarióticos de RNA lineares há uma ferramenta que se destaca frente as demais pelo uso de modelo probabilístico GHMM(*Generalized Hidden Markov Model*). Trata-se da ferramenta CodAn, bem descrita no artigo *CodAn: predictive models for precise identification of coding regions in eukaryotic transcripts* [7].

Essa ferramenta foi originalmente projetada para trabalhar com RNA lineares. O CodAn, até então possuia quatro modelos probabilísticos, cada um para um grupo específico de eucariotos: vertebrados, invertebrados, plantas e fungos. Vale salientar que o CodAn tem se destacado por previsões altamente confiáveis de regiões CDS (região de codificação de um gene em inglês "*coding DNA sequence*", abreviada por CDS), e UTR(regiões não traduzidas) completas não apenas em sequências de transcrição completa específicas de fita, mas também em sequências cegas e parciais a uma taxa muito maior do que outros *softwares* disponíveis[7]

O CodAn apresenta-se com base num *framework* adaptativo chamado ToPS[8], qual em seu artigo de divulgação científica *ToPS: A Framework to Manipulate Probabilistic Models of Sequence Data* define-se como uma implementação nova e flexível de decodificação em GHMMs[8].

## 2.4 TRANSCIRC

Devido ao avanço das pesquisas de RNA circulares, uma boa referência em banco de dados nos dias atuais engloba aqueles considerados bons bancos de dados de outrora, e provavelmente esse ciclo se repetirá num futuro próximo. Nesse sentido poder-se-ia citar inúmeros outros bancos de dados, como por exemplo o CircRNADb[18], com seus 32.914 de RNA circulares exônicos de origem humana, porém hoje temos um banco de dados em especial, o TransCirc, que abarca não apenas o CircRNADb, como inúmeros outros, e tem catalogados mais de 320.000 RNA circulares humanos, muito bem documentados e atualizados. Ao TransCirc, um banco de dados especializado em RNA circulares de origem humana como dito acima, foram integradas evidências multi-ômicas para prever regiões codificadoras em ORFs(em inglês *open reading frame*, quais são quadros de leitura com o potencial de ser transcrito e traduzido).

As evidências multi-ômicas do TransCirc são as seguintes:

- RP/PP *ribosome/polysome* - evidências de ligação ribossomo / polissomo;
- TIS *ranslation initiation sites* - sítios de iniciação de tradução mapeados experimentalmente em RNA circulares;

- IRES *internal ribosome entry site*- local de entrada do ribossomo interno em RNA circulares;
- m6A *N-6-methyladenosine modification* - dados de modificação de N-6-metiladenosina publicados em que promovem o início da tradução;
- ORF *open reading frames* - comprimentos das ORFs específicas de ;
- SeqComp *sequence composition* - sequenciar pontuações de composição a partir de uma previsão de aprendizado de máquina de todas ORFs;
- MS *mass spectrometry*- dados de espectrometria de massa que suportam diretamente os peptídeos codificados por em junções de back-splice.

Além dos sete tipos de evidências para a tradução de RNA circulares listadas acima, ou ilustradas na Figura 3, temos ainda mais informações como *StarCodon*, *StopCodon*, TC(*translation cycles*), e demais fornecidas pelo *metadata* do TransCirc, como número de evidências, escore de evidências para cada RNA circular. Pode ser acessado em: https://www.biosino.org/transcirc/.

**Figura 3 – TransCirc**


**Fonte: [19] TransCirc: um banco de dados interativo para RNAs circulares traduzíveis com base em evidências multi-ômicas.**

Sete tipos de evidências para a tradução de RNA circulares foram incluídos: (i) evidências de ligação ribossomo / polissomo; (ii) sítios de iniciação de tradução mapeados experimentalmente em RNA circulares; (iii) local de entrada do ribossomo interno em RNA circulares; (iv) dados de modificação de N-6-metiladenosina publicados em que promovem o início da tradução((v) comprimentos das ORFs específicas de ; (vi) sequenciar pontuações de composição a partir de uma previsão de aprendizado de máquina de todos os potenciais quadros de leitura abertos; (vii) dados de espectrometria de massa que suportam diretamente os peptídeos codificados por em junções de *back-splicing*.

Na Tabela 1 lista-se cinco ferramentas e banco de dados que foram pesquisadas por sua relevância, mas observa-se que duas delas encontram-se alternativamente embutidas respectivamente o FragGeneScan[9]no CirCode[10] e o CircRNADb[18] que é parte integrante do TransCirc [11] banco de dados de RNA circulares.

**Tabela 1 – Principais ferramentas pesquisadas**

| Nome | Características | Ano | Observação |
|---|---|---|---|
| CircCode | ML e HMM | 2019 | Classifica e prediz tradução em polipeptídeos |
| FragGeneScan | HMM | 2010 | Prediz tradução em polipeptídeos |
| CircRNADb | Database | 2016 | Evidências: ORFs e IRESs |
| TransCirc | Database | 2020 | Com sete evidências de tradução |
| CodAn v1.2 | GHMM | 2021 | Passível de adaptação p/ |

**Fonte: Autoria própria (2021) Cinco ferramentas e bancos de dados relevantes analisados, com destaque para a integração do FragGeneScan no CirCode e do CircRNADb como parte do banco de dados TransCirc de RNA circulares**

# 3 FUNDAMENTAÇÃO TEÓRICA

Vamos primeiramente relembrar os diversos tipos de RNA conhecidos atualmente. A saber:

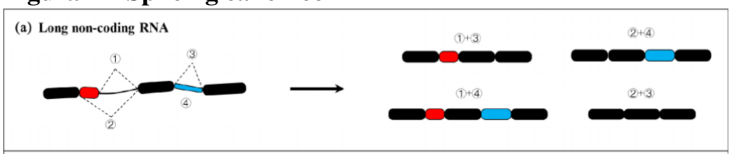
- RNA Mensageiro (mRNA) - Transfere as instruções do DNA no núcleo para o citoplasma, onde as proteínas são feitas[1].
- RNA Transportador (tRNA) - Transporta aminoácidos específicos para os ribossomos durante a síntese de proteínas[1].
- RNA Ribossômico (rRNA) - Forma a parte principal dos ribossomos e desempenha um papel na síntese de proteínas[1].
- RNA Circular (circRNA) - São estruturas fechadas que têm várias funções, incluindo a regulação da expressão gênica[2].
- RNA Interferente Pequeno (siRNA) - Interferem na expressão de genes específicos, levando à sua degradação[1].
- RNA Pequeno RNA nuclear (snRNA) - Atuam em uma série de processos nucleares, incluindo o *splicing* do pré-mRNA[1].
- RNA Pequeno RNA nucleolar (snoRNA) - Ajudam a processar e modificar quimicamente o rRNA[1].
- RNA Interagem com piwi (piRNA) - Protejem o DNA germinativo de sequências móveis que poderiam causar danos[1].
- RNA Longo Não Codificante (lncRNA) - Embora os lncRNAs não codifiquem proteínas, eles desempenham funções críticas na regulação gênica e em processos celulares[1].
- RNA Microssomal (miRNA) - Axiliam a regular a expressão gênica, ligando-se ao mRNA e influenciando sua estabilidade ou tradução.

Retomando nossa jornada, vamos doravante nos ater aos RNA lineares,como padrão de referência, e nos RNA circulares como foco do nosso trabalho.

Na biologia molecular, o RNA precursor, produzido a partir da fita molde de DNA por transcrição, pode ser processado em RNA mensageiro linear maduro por *splicing* de RNA, no qual os íntrons são removidos, conforme vê-se na Figura 4 enquanto os éxons se conectam em ordem genômica[1].

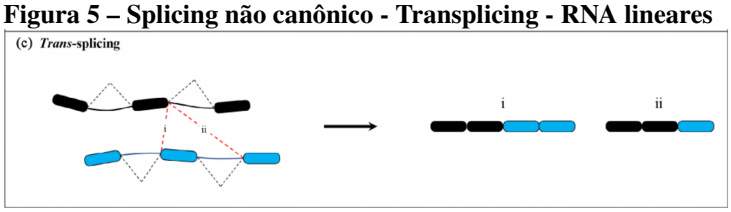
No entanto, também com os RNA lineares há o *splicing* alternativo, não canônico, denominado *transplicing*, onde como vê-se na Figura 5 há recombinação de éxons procedentes de RNA mensageiros distintos.

**Figura 4 – Splicing canônico**




**Fonte: Adaptado de [20] Processamento do RNA Precursor em RNA Mensageiro Linear Maduro: Remoção dos Íntrons e conexão dos Éxons em ordem genômica**

**Figura 5 – Splicing não canônico - Transplicing - RNA lineares**




**Fonte: Adaptado de [20] Transplicing: Recombinação de éxons de diferentes RNAs mensageiros através do splicing alternativo não canônico**

Com os RNA circulares ocorre o *back-splicing*, ilustrado na Figura 6, qual faz éxons formarem um círculo fora da ordem genômica original, e consequentemente provoca assim um aumento significativo na variabilidade nas suas sequencias de nucleotídeos, quais se traduzidos em polipeptídeos expandem ainda as possibilidades de proteínas probabilisticamente viáveis de serem produzidas. Este trabalho foca na identificação de regiões codificadoras em transcritos

**Figura 6 – Back-splicing - RNA circulares**


**Fonte: Adaptado de [20] Back-splicing em RNA circulares: Reorganiza éxons fora da ordem genômica, ampliando a diversidade de sequências nucleotídicas e potencializando a gama de proteínas produzíveis**

eucarióticos de RNA circular advindos dessa forma de *splicing* não canônico, também chamados de *back-splicing*[18, 5].

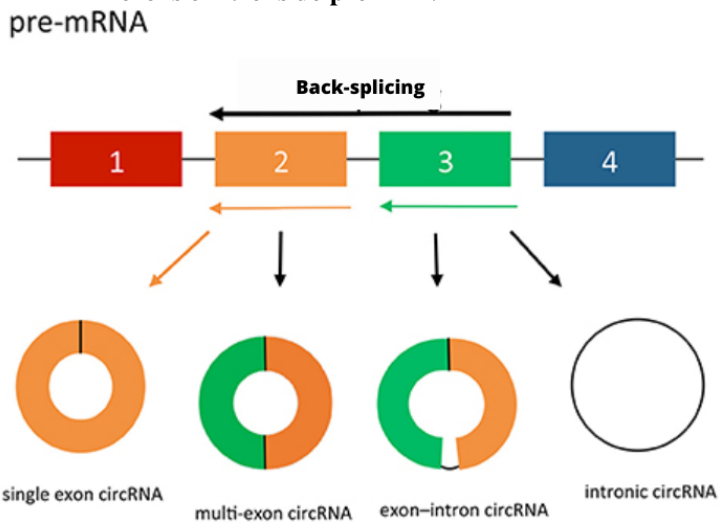
## 3.1 RNA CIRCULARES, O QUE SÃO?

RNA circulares podem ser constituídas por bases púricas A e G ou pelas pirimídicas C e U. A circularização do RNA se dá através de ligação covalente da extremidade 5' com a 3'[21].

Os RNA circulares podem conter um ou mais éxons, conter éxon e íntron, ou apenas íntron(vide Figura 7). Com esses conceitos em pauta, nosso foco será os RNA circulares exônicos, ou sejam os constituídos por um ou mais éxons[21, 22].

Essas moléculas de RNA ocorrem em eucariotos, apresentam forma de anel, e podem

**Figura 7 – RNA circulares - formados a partir de emendas de éxons e íntrons do pré-mRNA**




**Fonte: [4] Os RNA circulares podem conter um ou mais éxons, conter éxon e íntron, ou apenas íntron - Barras coloridas simbolizam éxons e linhas pretas, os íntrons**

ser identificadas com base em suas junções de *back-splicing*, frequentemente produzidas a partir de éxons, em que as extremidades 5' e 3' são covalentemente unidas para formar uma junção denominada BSJ(*back-splicing junction*). Foi demonstrado que, em humanos, os RNA circulares são a isoforma predominante de eventos de desordenamento de éxons, e que a circularização é uma característica generalizada[23, 10].

Na Figura 8 logo abaixo, pode-se observar de cima para baixo, em camadas ilustrativas as seguintes características dos RNA circulares: À medida que o migra do núcleo celular para fora da célula, temos os modelos de formação, os tipos existentes, as funções biológicas assumidas, a saída da célula através dos exossomos, fato este que lhe permite atuar como biomarcador.

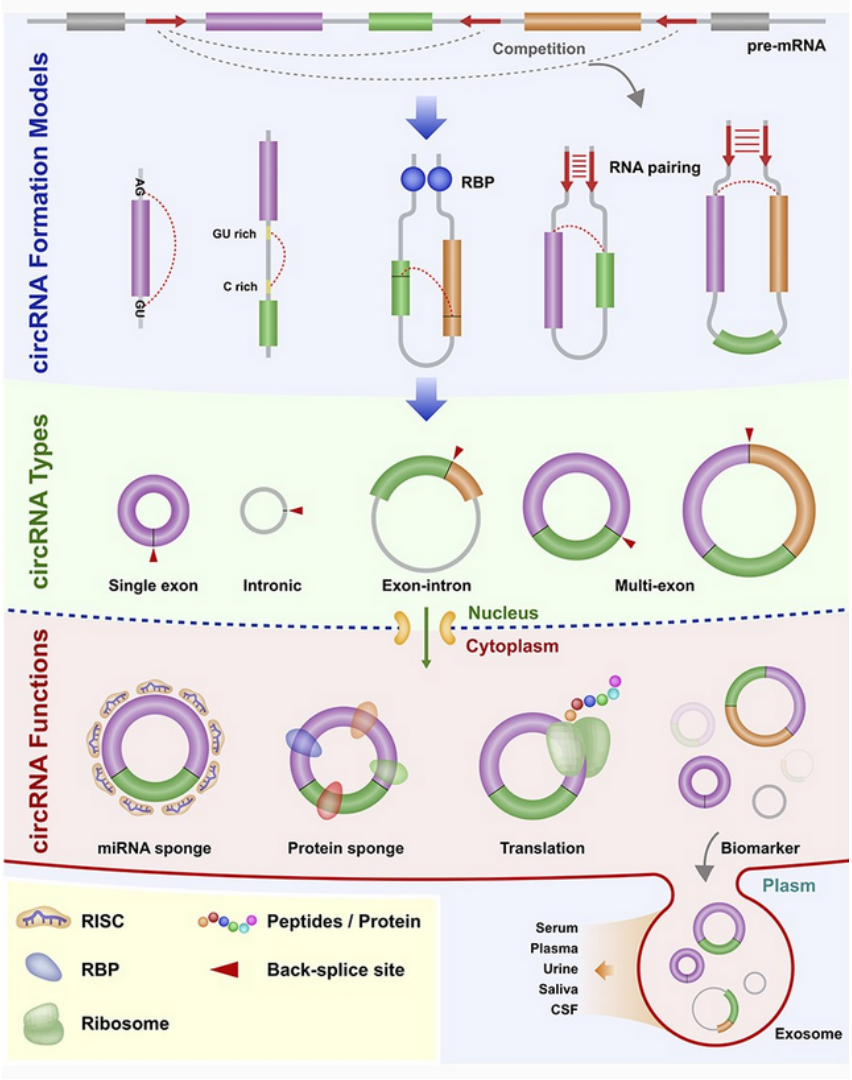
O aprofundamento de pesquisas sobre RNA circulares poderá viabilizar processos biotecnológicos na indústria de alimentos com a produção de proteínas específicas, consolidar e avançar mais ainda, a perspectiva de transferência de genes para o desenvolvimento de novas terapias genéticas para o tratamento de doenças humanas[5].

## 3.2 SEMELHANÇAS E DIFERENÇAS NA TRADUÇÃO ENTRE RNA CIRCULARES E RNA LINEARES

Ilustra-se na Figura 9, a formação e a relação entre RNA circulares e lineares provenientes de um mesmo pré-mRNA. Ou seja a informação genética original num determinado gene pode ser expressa tanto através de RNA circulares como lineares provocando assim resultados e funcionalidades distintas entre si.

É importante ressaltar que os códons representados na Figura 10 são formados pelas bases do RNA-mensageiro obtido após o processo de transcrição do DNA[2]. Podemos notar na

**Figura 8 – RNA circular, modelos de formação, tipos e funções**




**Fonte: [22]**
**Observa-se de cima para baixo em camadas ilustrativas a medida que migra do núcleo celular para fora da célula, os modelos de formação, os tipos existentes,as funções biológicas assumidas e a presença de vesículas na membrana celular contendo RNA circulares deixando a célula através de exossomos, o que permite atuar como biomarcador**

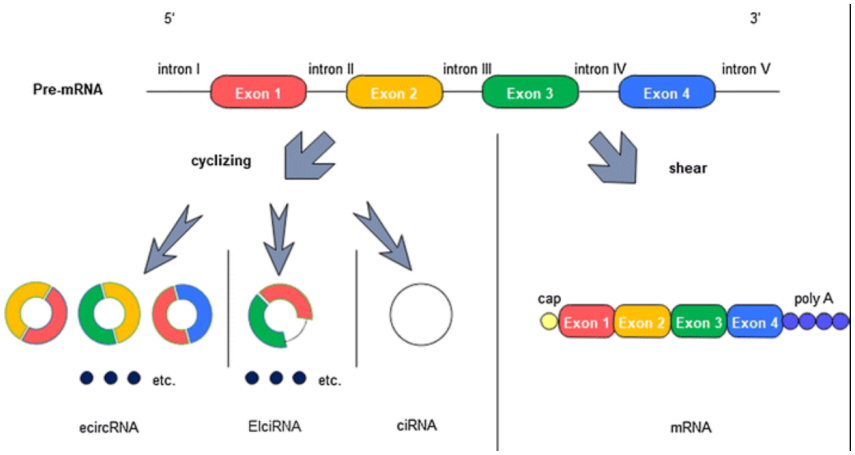
Figura 10 a relação entre os códons de RNAm (código de três letras) e os aminoácidos que eles codificam. O código genético é degenerado, o que significa que vários códons diferentes podem codificar o mesmo aminoácido.

Os códons de iniciação de cadeia polipeptídica são os códons que indicam onde a síntese de proteínas deve começar. Eles são geralmente AUG (em que A é adenina, U é uracila, G é guanina).

Os códons de término de cadeia polipeptídica são os códons que indicam onde a síntese de proteínas deve terminar. Eles são geralmente três códons diferentes, UAA, UAG e UGA, que não codificam aminoácidos.

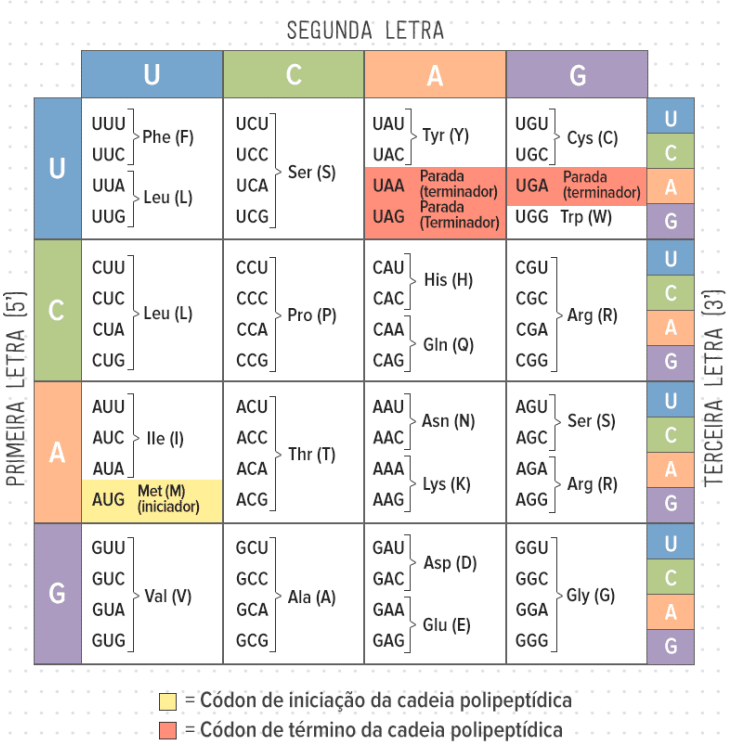
Recordando a tradução de RNA linear em proteína, vê-se na Figura 11 o fluxo da infor-

**Figura 9 – RNA lineares e circulares: Formação e relações entre si.**




**Fonte: [24]**
**Formação de RNA circulares e mRNA. RNA circulares são gerados no processo de *splicing* de pré-mRNA e competem com mRNA. RNA exônicos (ERNA circulares) são os mais comuns.**

**Figura 10 – Código Genético**

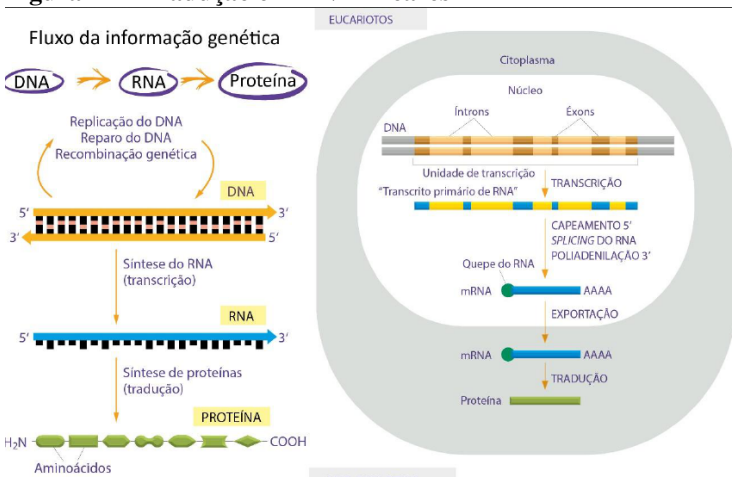



**Fonte: [25]Códons no RNA mensageiro são sequências de três bases resultantes da transcrição do DNA. Esses códons determinam os aminoácidos correspondentes na síntese proteica. Múltiplos códons podem codificar um único aminoácido devido à degenerescência do código genético. O códon de iniciação é geralmente AUG, enquanto os códons de término incluem UAA, UAG e UGA, que não se traduzem em aminoácidos**

mação genética do DNA para o RNA e sua tradução em proteínas, nas seguintes fases:

- Transcrição.

**Figura 11 – Tradução em RNA lineares**



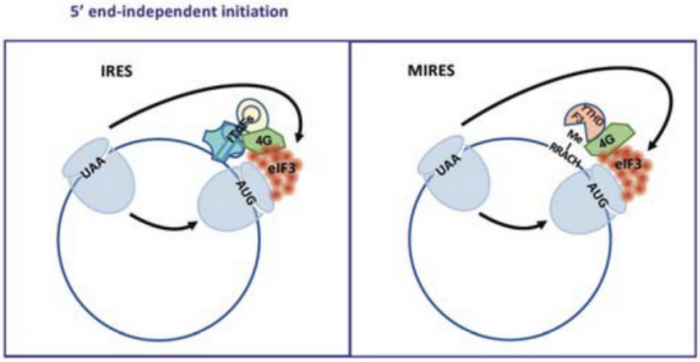

**Fonte: [1] Fluxo de informação genética do DNA ao RNA em eucariotos.**

- Processamento do RNA.
- Exportação do RNA para fora do núcleo celular.
- Tradução do RNA com Quepe 5' e cauda poliadenilada 3'.
- Proteína pronta.

Em contraste, uma nova família de RNA, os chamados circulares(RNA circulares) revela o papel central da tradução independente de extremidade 5'.

Observa-se na Figura 12 o iniciar da tradução nos RNA circulares: à esquerda por I.R.E.S. (*internal ribosome entry sites*) - locais de entrada de ribossomo interno e à direita por M.I.R.E.S.(*m6A-induced ribosome engagement sites*) locais de engajamento de ribossomos induzidos por m6A promovem o início da tradução independente de extremidade 5'.

**Figura 12 – IRES e MIRES**

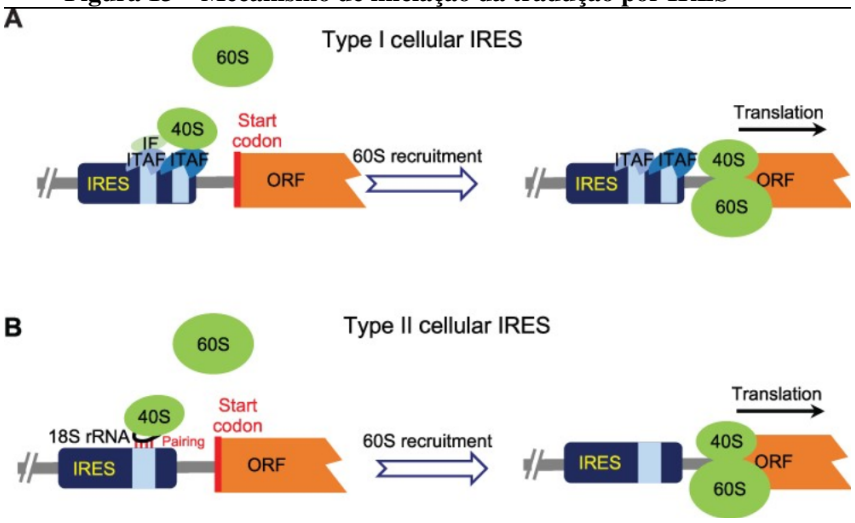



**Fonte: Adaptado de [5] Iniciação da tradução em RNA circulares: à esquerda, através dos I.R.E.S. (locais de entrada de ribossomo interno) e à direita, pelos M.I.R.E.S. (locais de engajamento induzidos por m6A), ambos promovendo a tradução sem depender da extremidade 5'**

RNA circulares têm um forte impacto no controle translacional por meio de sua função de "esponja"e formam uma nova família de mRNA à medida que são traduzidos em proteínas com funções fisiopatológicas[4].

Até a década de 1990, pensava-se que a máquina de tradução eucariótica era incapaz de traduzir um RNA circular. No entanto, como ilustrado na Figura 13 locais de entrada de ribossomo interno (IRESs) e locais de engajamento de ribossomos induzidos por m6A (MIRESs) foram descobertos, promovendo o início da tradução independente de extremidade 5'[5].

**Figura 13 – Mecanismo de iniciação da tradução por IRES**




**Fonte: [26] Mecanismo da tradução conduzida por IRES celular. O mecanismo dos IRESs virais traduzindo proteínas já está bem esclarecido; no entanto, IRESs celulares ainda precisam de evidências mais confiáveis para aclarar seus mecanismos. Na figura temos:(A) IRES celular Tipo I. (B) IRES celular Tipo II. Nos dois tipos, esses IRESs celulares podem interagir diretamente com a subunidade ribossômica 40S e recrutar a subunidade ribossômica 60S para iniciar a tradução."**

E ainda mais recentemente, mecanismos adicionais de interação 3'- 5' foram relatados, incluindo modificação de m6A, e que a circularização funcional melhora a tradução por meio da reciclagem do ribossomo, aumentando a taxa de iniciação da tradução[5]. Este cenário de tradução de mRNA circular, fechada covalentemente e não covalentemente, mostra que o RNA circular pode ser a regra para a tradução com um impacto importante no desenvolvimento de doenças e nas possíveis aplicações biotecnológicas[5].

## 3.3 MODELOS OCULTOS GENERALIZADOS DE MARKOV - GHMM

Inúmeras aplicações foram desenvolvidos para encontrar genes usando o modelo oculto generalizado de Markov (GHMM). Este tipo de modelo é preciso e oferece uma interpretação fácil do problema de localização de genes. Um GHMM é um modelo que cria sequências de bases, como éxons ou íntrons, a partir de cada estado. Este modelo tradicionalmente não inclui informações de homologia[27].

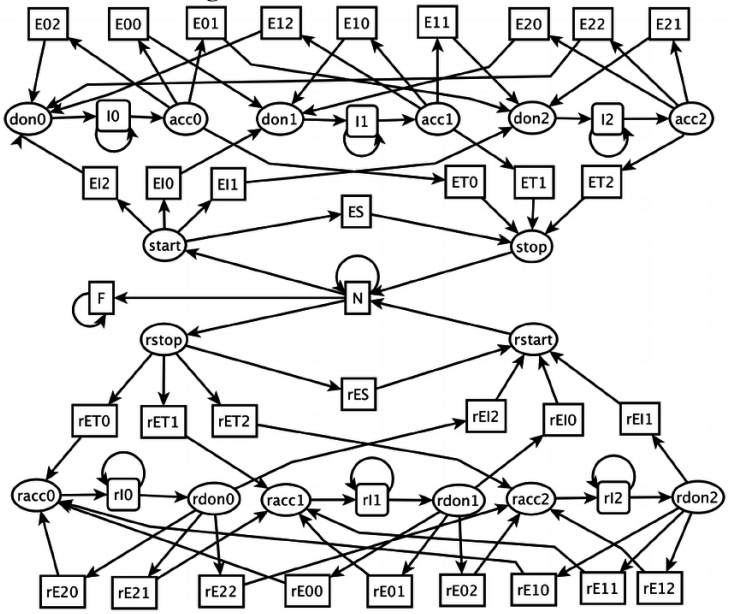
Seja $\alpha = \{A, T, C, G\}$, $\alpha^n = \{s \mid s$ é uma string sobre $\alpha$ de comprimento $n\}$, $\alpha^* = \bigcup_n(\alpha^n)$, $\mathbb{N}$ os inteiros não negativos e $\mathbb{R}$ os reais. Um GHMM é a rigor definido como uma 6-tupla $(Q, P_t, P_d, P_e, \pi_0, \pi_f)$. Há conjuntos de probabilidades de transição de estado, probabilidades de duração ou comprimento e probabilidades de emissão, ou seja, um conjunto de estados $Q$ com estado inicial designado $\pi_0$ e estado final $\pi_f$ (ambos silenciosos), um conjunto de probabilidades de transição de estado $P_t : Q \times Q \to \mathbb{R}$, um conjunto de probabilidades de comprimento ou 'duração' $P_d : \mathbb{N} \times Q \to \mathbb{R}$ condicional no estado e um conjunto de probabilidades de emissão $P_e : \alpha^* \times Q \times \mathbb{N} \to \mathbb{R}$ condicional ao estado e duração. Uma única execução $n$-state do GHMM começa no estado $q_0 = \pi_0$, transita estocasticamente do estado $q_{i-1}$ para o estado $q_i$ para $1 \leq i < n$ de acordo com $P_t(q_i \mid q_{i-1})$, e termina no estado $q_{n-1} = \pi_f$. Em cada estado $q_i$ o GHMM escolhe estocasticamente a duração $d$ de acordo com $P_d(d \mid q_i)$ e emite string $S_i \in \alpha^d$ de acordo com $P_e(S_i \mid q_i, d)$. Observe que $P_d(0 \mid \pi_0) = P_d(0 \mid \pi_f) = 1$ [27]. :

$$\phi_{max} = \underset{\phi}{argmax}\ P(\phi|\, S) = \underset{\phi}{argmax}\ \tfrac{P(\phi,S)}{P(S)}$$
$$= \underset{\phi}{argmax}\ P(\phi, S) = \underset{\phi}{argmax}\ P(S|\, \phi)P(\phi)\,,$$

(1)

onde cada $\phi = \{(q_i, d_i)|0i < n\}$ especifica uma série ordenada no tempo de estados (recursos) e durações inteiras (comprimentos de recursos) durante uma única execução do GHMM. $P(S|\phi)$ pode ser fatorado de acordo com os estados em $\phi$ para produzir $P_e(S_i \mid q_i, d_i)$, onde a fórmula precisa para cada $P_e(S_i \mid q_i, d_i)$ é definido separadamente para cada estado dependendo do tipo de modelo usado no estado (uma cadeia de Markov, matriz de peso específica de posição, etc.). $P(\phi)$ também pode ser decomposto por estado em um produto de probabilidades de transição e duração para produzir onde a concatenação $S_0, \ldots, S_{n1}$ de recursos individuais forma a sequência de entrada $S$. Esta etapa de maximização pode ser calculada eficientemente usando um algoritmo de decodificação de Viterbi modificado [28] [27]

**Figura 14 – Modelos ocultos generalizados de Markov**




**Fonte: [8] Os Modelos Ocultos Generalizados de Markov (GHMM) são definidos formalmente como uma tupla de seis elementos, incluindo o conjunto de estados, probabilidades de transição de estado, probabilidades de duração, probabilidades de emissão e estados iniciais e finais.**

# 4 METODOLOGIA

Neste trabalho, após acurada análise, optamos pelo TransCirc[11] por reunir demais relevantes bancos de dados de RNA circulares, e pelo número maior de evidências de regiões codificadoras, ou seja, pelo maior número de evidências de Espectometria de Massa(MS) entre outras evidências também muito importantes, e pelo maior número de RNA circulares, e que muitos são originários de outros bancos, e desse forma o TransCirc aglutina demais iniciativas, conservando o ID original e adicionando um ID próprio, somando mais de 320 mil RNA circulares de origem humana.

### 4.0.1 Detalhamento da obtenção da amostra.

Com base nos dados fornecidos pelo TransCirc através dos arquivos de metadados e arquivos de sequencia no formato fasta dos RNA circulares constantes do TransCirc, ilustrado na Figura 15, criamos um *script* em *Python* de pre-processamento.

**Figura 15 – RNA circular linearizado no arquivo FASTA**


**Fonte: Própria autoria (2021)**
**Observa-se neste arquivo FASTA que na sequencia de cada representação de RNA circular, pode haver inúmeras ORFs(Quadro Aberto de Leitura), onde encontram-se as possíveis regiões codificantes**

O TransCirc disponibiliza arquivos fasta de RNA circulares com seus respectivos metadados, e nosso interesse convergiu nos RNA circulares com os maiores escores de evidência

de tradução de proteínas. Para tal, aplicamos alguns filtros para selecionar os RNA circulares com maiores escores de evidência de tradução em proteínas e também para retirar ORFs com mesma posisão de códon de início e códon de parada, como também mesmo comprimento de ORFs para evitar redundâncias desnecessárias.

Iniciamos definindo o número da amostra de RNA circulares em 15 mil exemplares para processamento dentro do número total de RNA circulares do TransCirc com mais de 320 mil exemplares de RNA circulares humanos.

Primeiramente, realizamos leitura completa do arquivo de metadados constante do TransCirc, e a partir de então começamos os procedimentos a seguir:

- Remover as sequências que não contém nenhuma ORF.
- Eliminar as sequências que possuem mais de uma ORF.
- Ordenar as sequências de forma decrescente com base nos escores de evidências de tradução.
- Selecionar os primeiros 15 mil como definido anteriormente. Observa-se que essa limitação dos primeiros 15 mil foi arbitrada por questões práticas, pela demora do processamento, mas pelo fato de ser baseada nos maiores escores de evidência de tradução não compromete em absoluto o resultado do experimento. Temos até então 15 mil RNA circulares selecionados com os maiores escores de tradução em proteínas
- Localizar a posição inicial da ORF com base nessa seleção acima e o arquivo de sequencias fasta obtém-se a posição que a ORF começa em cada um dos RNA circulares selecionados.
- Calcular os detalhes da ORF considerando as sequências indexadas a partir de 0(zero). Os detalhes são onde a ORF inicia, onde acaba, quantas voltas são dadas e o tamanho da sequencia da ORF. Por voltas seja entendido o número de ciclos de translações até um limite de máximo 3, pois trata-se de transcritos de RNA circulares que em geral podem dar até três voltas na maioria dos casos para encontrar um códon de parada.
- Definir detalhes técnicos importantes como o retorno das posições genômicas, pois o início das sequencias no Python começa por 0(zero) e nos dados do TransCirc inicia-se por 1(um).

- Remover RNA circulares com ORFs com mesmos códon de início, códon de parada e mesmo comprimento (repetidos). Trata-se de uma depuração da amostra a ser utilizada.
- Por fim registrar em arquivo de metadados com os 12.800 RNA circulares restantes após as remoções e redundâncias, e também registrar os arquivos fasta com as sequencias referentes aos RNA circulares selecionados obtendo-se a amostra preparada para análise posterior.

Assim temos agora a amostra com as informações complementares de início e final das ORFs, número de ciclos de translações, e o comprimento final de cada um dos RNA circulares, ou seja temos a amostra preparada para análise, representada pelos arquivos metadados e fasta.

### 4.0.2 O experimento que possibilitou o comparativo entre as ferramentas: CodanCirc e FragGeneScan

Para implementação do experimento em si implementamos mais dois scripts Python, um que utiliza o resultado do pré-processamento anterior e auxilia a execução da ferramente em estudo, e outro que nos fornece as métricas do experimento para as comparações das ferramentas no encerramento do experimento.

Em seguida, ele itera pelas previsões e compara-as com as posições reais de início de códon dos RNA circulares para calcular verdadeiros positivos (TP), falsos positivos (FP), verdadeiros negativos (TN) e falsos negativos (FN).

A função então calcula as métricas de desempenho como acurácia, precisão, revocação e f1_score usando esses valores.

Em resumo, este script automatiza o cálculo de métricas de desempenho para o resultado de uma análise de tradução de RNA circulares comparando a previsão com a posição real de *start codon* dos RNA circulares.

Os procedimentos descritos acima são para a ferramenta CodAn com o modelo `VERT_circ`. No caso da ferramenta FragGeneScan o procedimento é outro. Por ser concebido com outras tecnologias ficou mais prático fornecer os *inputs* fora do ambiente *Python* citado anteriormente, e submeter seus *outputs* para calcular as métricas pelo mesmo *scrip* eval.py separadamente.

### 4.0.3 E como surgiu a ferramenta CodAnCirc?

Como RNA circulares não possuem *cap* e cauda poliadenilada, e sim um formato característico circular, foi necessário realizar uma adaptação do modelo para este tipo específico

de RNA. Para isso, foram feitas algumas modificações no modelo, dando origem a uma versão específica do CodAn para RNA circulares, denominada CodAnCirc.

A ferramenta ToPs[8], a qual o CodAn utiliza internamente, é bastante flexível, o que permitiu a adaptação do modelo GHMM, criando assim o modelo VERT_circ para trabalhar com os transcritos de RNA circulares. Graças a essa adaptação, é possível trabalhar com esses transcritos de maneira mais precisa e eficiente.

#### 4.0.3.1 A adaptação do modelo VERT_full para VERT_circ

Originalmente partimos do modelo VERT_full, qual contém internamente o arquivo ghmm_intronless.model, onde situa a configuração geral da GHMM.

A seguir descrevemos sucintamente o modelo do ToPS, o GHMM, qual é um modelo oculto de Markov que tem vários estados e que diferentemente do HMM, modelo oculto de Markov simples, cada estado da GHMM pode emitir mais de um símbolo. Ou seja, podem representar várias características ou fases do processo de transcrição e tradução. Estes incluem "begin", "end", "CDS", "start", "stop", "CDS0", "CDS1"e "CDS2". Os símbolos de observação "A", "C", "G"e "T" representam as quatro bases nucleotídicas do RNA. O modelo especifica probabilidades iniciais para cada estado e probabilidades de transição entre estados, refletindo a sequência de eventos na transcrição e tradução do RNA. Os modelos específicos para os estados "start", "stop", "cds"e "noncoding" são então referenciados e representam submodelos que descrevem com mais detalhes as características desses estados, como a composição de base e padrões de codificação. A duração de cada estado "CDS" é definida por um modelo "Phased Run Length Distribution", que pode ser utilizado para representar a duração variável de segmentos de codificação em diferentes fases de tradução. Além disso, cada estado tem uma configuração associada que define um modelo de observação, a duração (se aplicável), e as fases de entrada e saída. É preciso notar que o estado "CDS"inclui uma extensão de emissão, que poderia ser usada para modelar a influência do contexto de sequência na emissão de um estado. Essas configurações e parâmetros podem ser ajustados para refletir a realidade dos RNA circulares, fornecendo uma ferramenta poderosa para explorar a estrutura e a função dessas moléculas foco de nosso estudo.

GHMMs são modelos probabilísticos muito flexíveis que podem ser integrados a outros modelos para descrever uma arquitetura complexa que foge ao escopo deste trabalho. No entanto essas descrições acima expostas objetivaram apenas mostrar como podemos modificar os parâmetros internos do modelo para se adequar a realidade dos RNA circulares.

# 5 RESULTADOS E DISCUSSÃO

## 5.1 COMPARATIVO ENTRE AS FERRAMENTAS - CODANCIRC E FRAGGENESCAN

Tanto o CodAnCirc e o FragGeneScan[9] foram alimentados com a mesma amostra e foram utilizadas as mesmas métricas de avaliação, conforme descrito na metodologia, e os resultados foram adicionados na tabela 2 a seguir:

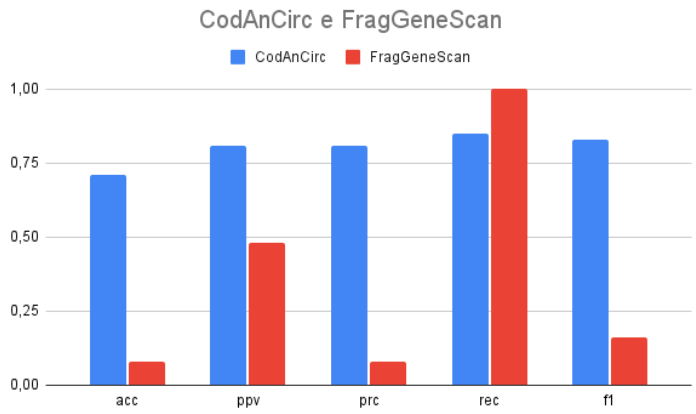
**Tabela 2 – Tabela comparativa CodAnCirc e FragGeneScan**

| Nome | Amostra | TP | TN | FP | FN | acc | ppv | prc | rec | f1 |
|---|---|---|---|---|---|---|---|---|---|---|
| CodAnCirc | 12.800 | 9.086 | 0 | 2.086 | 1.628 | 0.71 | 0.81 | 0.81 | 0.85 | 0.83 |
| FragGeneScan | 12.756 | 1.085 | 0 | 11.671 | 0 | 0.08 | 0.48 | 0.08 | 1 | 0.16 |

**Fonte: Autoria própria (2023)**

Conforme observa-se no gráfico da Figura 16, foi realizada uma comparação entre CodAnCirc e FragGeneScan. O FragGeneScan apresentou resultados consideravelmente menores em todas as métricas, exceto no *recall*. Isso se deve ao número de falsos negativos ser igual a zero. Como a métrica de *recall* é calculada pelo número de verdadeiros positivos dividido pela soma de verdadeiros positivos e falsos negativos, a razão se torna TP/TP, ou seja, uma unidade ou 100%.

**Figura 16 – Comparativo entre ferramentas**




**Fonte: Autoria própria (2023) FragGeneScan teve desempenho menor em todas as métricas, exceto no recall. Isso ocorre porque não apresentou falsos negativos. Portanto, a fórmula do recall (verdadeiros positivos divididos pela soma de verdadeiros positivos e falsos negativos) resulta em 100%**

Na análise comparativa entre CodAnCirc (baseado em GHMM) e FragGeneScan (que utiliza Modelos Ocultos de Markov - HMM), nota-se que o *f1-score* obtido por nossa ferramenta foi de 83%, enquanto a ferramenta concorrente alcançou apenas 16%. Essa diferença observada sugere uma eficácia maior do GHMM para esta tarefa específica

# 6 CONSIDERAÇÕES FINAIS

## 6.1 CONCLUSÕES DA PESQUISA

Concluindo em resumo, os objetivos propostos foram alcançados com sucesso, através da identificação de regiões codificadoras em transcritos eucarióticos de RNA circular, e da adaptação da ferramenta CodAn para RNA circulares criando-se a nova ferramenta CodAnCirc, e pela comparação das performances entre as ferramentas utilizadas.

## 6.2 ALCANCE DOS OBJETIVOS

Neste trabalho, foi atingido o objetivo geral de identificar regiões codificadoras de proteínas nos RNA circulares aplicando-se Modelos Ocultos Generalizados de Markov. Especificamente, alcançamos os seguintes objetivos:

- Foram identificadas regiões codificadoras em transcritos eucarióticos de RNA circular, utilizando uma metodologia baseada em um modelo probabilístico GHMM (Generalized Hidden Markov Model), configurado pelo *framework* probabilístico ToPS[8], que integra a ferramenta CodAn (Coding sequence Annotator)[7], originalmente utilizada para RNA lineares com extremidades 5' e 3', qual adaptamos através da criação de um modelo específico para RNA circulares, e a nomeamos CodAnCirc (Foi criado um *fork* da publicação original do CodAn no GitHub em:

  `https://github.com/rodolphojunior/CodAnCirc`).

- Compararamos as performances entre si das ferramentas CodAnCirc e FragGeneScan [9] (Os scripts utilizados encontram-se disponíveis no GitHub em:

  `https://github.com/rodolphojunior/bioinfo-master`).

## 6.3 RECOMENDAÇÕES DE TRABALHOS FUTUROS

O CodAnCirc, conforme os resultados obtidos provaram, pode ser apenas a ponta do *iceberg* de novas pesquisas. Com algumas alterações no modelo probabilístico GHMM através da criação do modelo específico `VERT_circ` pudemos originar de fato uma nova ferramenta, aqui anunciada desde 2021 como CodAnCirc. Dessa forma, comprovamos a expressiva vantagem dos

métodos probalísticos baseados em GHMM do CodAnCirc em relação as demais ferramentas na mesma categoria.

No entanto, em meio a tendência atual do avanço das tecnologias de Inteligência Artificial, nos permitimos a mergulhar no ingrediente mais importante da criação intelectual: a imaginação! Com esse propósito, vemos que novas tecnologias como *Transformers*[29], talvez possam ser amplamente aplicáveis em biologia. Por se basear em NLP, processamento de linguagem natural, em bioinformática poderiam ser usados para processamento de sequências biológicas, como RNA e proteínas. Assim poderiam incluir tarefas como classificação de sequências, identificação ou previsão de regiões codificadoras em transcritos de RNA circular, entre outros. Além disso, as técnicas de aprendizado profundo dos *Transformers* poderiam ser combinadas com contextos dos modelos biológicos e probabilísticos, e assim melhorar a precisão dessas análises. Diante das possibilidades apresentadas, pode-se inferir que há um potencial para o desenvolvimento de novas ferramentas bioinformáticas inteligentes.

# REFERÊNCIAS


[1] ALBERTS, B. et al. *Biologia Molecular da Célula*. [S.l.: s.n.], 2021. ISBN 9788582714232.

[2] SNUSTAD, S. *Fundamentos de Genética*. [S.l.]: Guanabara, 2017.

[3] DOBZHANSKY, T. Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher*, [University of California Press, National Association of Biology Teachers], v. 35, n. 3, p. 125–129, 1973. ISSN 00027685, 19384211. Disponível em: <http://www.jstor.org/stable/4444260>.

[4] GREENE, J. et al. Circular rnas: biogenesis, function and role in human diseases. *Frontiers in molecular biosciences*, Frontiers, v. 4, p. 38, 2017.

[5] PRATS, A.-C. et al. Circular rna, the key for translation. *International Journal of Molecular Sciences*, v. 21, n. 22, 2020. ISSN 1422-0067. Disponível em: <https://www.mdpi.com/1422-0067/21/22/8591>.

[6] GENIUS, B. *Central Dogma*. 2021(accessed September 7, 2021). <https://genius.com/Biology-genius-the-central-dogma-annotated>.

[7] NACHTIGALL, P. G.; KASHIWABARA, A. Y.; DURHAM, A. M. Codan: predictive models for precise identification of coding regions in eukaryotic transcripts. *Briefings in bioinformatics*, Oxford University Press, v. 22, n. 3, p. bbaa045, 2021.

[8] KASHIWABARA, A. Y. et al. Tops: a framework to manipulate probabilistic models of sequence data. *PLoS computational biology*, Public Library of Science San Francisco, USA, v. 9, n. 10, p. e1003234, 2013.

[9] RHO, M.; TANG, H.; YE, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research*, v. 38, n. 20, p. e191–e191, 08 2010. ISSN 0305-1048. Disponível em: <https://doi.org/10.1093/nar/gkq747>.

[10] SUN, P.; LI, G. Circcode: a powerful tool for identifying circrna coding ability. *Frontiers in genetics*, Frontiers, v. 10, p. 981, 2019.

[11] HUANG, W. et al. Transcirc: an interactive database for translatable circular rnas based on multi-omics evidence. *Nucleic acids research*, Oxford University Press, v. 49, n. D1, p. D236–D242, 2021.

[12] CHEN, L. et al. The bioinformatics toolbox for circrna discovery and analysis. *Briefings in bioinformatics*, Oxford University Press, v. 22, n. 2, p. 1706–1728, 2021.

[13] GAO, Y.; WANG, J.; ZHAO, F. Ciri: an efficient and unbiased algorithm for de novo circular rna identification. *Genome biology*, BioMed Central, v. 16, n. 1, p. 1–16, 2015.

[14] MA, X.-K. et al. Circexplorer pipelines for circrna annotation and quantification from non-polyadenylated rna-seq datasets. *Methods*, Elsevier, 2021.

[15] MENG, X. et al. Circpro: an integrated tool for the identification of circrnas with protein-coding potential. *Bioinformatics*, Oxford University Press, v. 33, n. 20, p. 3314–3316, 2017.

[16] JAKOBI, T.; UVAROVSKII, A.; DIETERICH, C. circtools—a one-stop software solution for circular rna research. *Bioinformatics*, Oxford University Press, v. 35, n. 13, p. 2326–2328, 2019.

[17] ITO, E. A. et al. Basinet—biological sequences network: a case study on coding and non-coding rnas identification. *Nucleic acids research*, Oxford University Press, v. 46, n. 16, p. e96–e96, 2018.

[18] CHEN, X. et al. circrnadb: a comprehensive database for human circular rnas with protein-coding annotations. *Scientific reports*, Nature Publishing Group, v. 6, n. 1, p. 1–6, 2016.

[19] TRANSCIRC. 2021. Disponível em: <https://www.biosino.org/transcirc/>.

[20] XU, B.; MENG, Y.; JIN, Y. Rna structures in alternative splicing and back-splicing. *Wiley Interdisciplinary Reviews: RNA*, Wiley Online Library, v. 12, n. 1, p. e1626, 2021.

[21] DIALLO201945. *How are circRNAs translated by non-canonical initiation mechanisms?* 2020 (accessed October 10, 2021). <https://doi.org/10.1016/j.biochi.2019.06.015>.

[22] ZENG, X. et al. A comprehensive overview and evaluation of circular rna detection tools. *PLoS computational biology*, Public Library of Science San Francisco, CA USA, v. 13, n. 6, p. e1005420, 2017.

[23] GLAŽAR, P.; PAPAVASILEIOU, P.; RAJEWSKY, N. circbase: a database for circular rnas. *Rna*, Cold Spring Harbor Lab, v. 20, n. 11, p. 1666–1670, 2014.

[24] HOU, J. chen et al. Circular rnas and exosomes in cancer: a mysterious connection. *Clinical and Translational Oncology*, v. 20, p. 1109–1116, 2018.

[25] GROUP, C. M. *Código genético*. 2010(accessed September 7, 2021). <https://www.todoestudo.com.br/biologia/codigo-genetico>.

[26] YANG, Y.; WANG, Z. Ires-mediated cap-independent translation, a path leading to hidden proteome. *Journal of molecular cell biology*, Oxford University Press, v. 11, n. 10, p. 911–919, 2019.

[27] MAJOROS, W. H. et al. Efficient decoding algorithms for generalized hidden markov model gene finders. *BMC bioinformatics*, BioMed Central, v. 6, n. 1, p. 1–13, 2005.

[28] BURGE, C. B. *Identification of genes in human genomic DNA*. [S.l.]: Stanford University, 1997.

[29] VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.