Ministério da Educação
Universidade Tecnológica Federal do Paraná
Pró-Reitoria de Pesquisa e Pós-Graduação

RELATÓRIO FINAL DE ATIVIDADES DE INICIAÇÃO CIENTÍFICA

Classificação de sequências biológicas usando Máxima Entropia

vinculado ao projeto

Inferência de GRNs a partir da integração de dados multiníveis

Murilo Montanini Breve
Bolsista da UTFPR
Engenharia de Controle e Automação
Data de ingresso no programa: 08/2020
Dr(a). Fabrício Martins Lopes

Área do Conhecimento: Engenharia de Computação 3.00.00.00-9 — Engenharia

CAMPUS CORNÉLIO PROCÓPIO, 2021

MURILO MONTANINI BREVE
FABRÍCIO MARTINS LOPES

CLASSIFICAÇÃO DE SEQUÊNCIAS BIOLÓGICAS USANDO MÁXIMA ENTROPIA

Relatório de Pesquisa do Programa de Iniciação Científica da Universidade Tecnológica Federal do Paraná.

CORNÉLIO PROCÓPIO, 2021

SUMÁRIO

INTRODUÇÃO

No século XIX, o bioquímico suíço Friedrich Miescher (1844-1895) isolou de uma célula, um ácido que continha fósforo e nitrogênio, e após 20 anos desta descoberta, seu discípulo, Richard Altmann estabeleceu o nome de ácido nucleico a este composto, como conhecemos hoje [1]. E devido a estes avanços ciêntificos, em 1953, James D. Watson e Francis H. Crick, viriam a publicar "A Structure for Deoxyribose Nucleic Acid", [2] a primeira menção da estrutura DNA no campo cientifico, e viria ser considerada umas das mais importantes contribuições a Biologia.

Devido a sua indispensável participação na manutenção e criação da vida em nosso planeta, os ácidos nucleicos ganharam espaço e notoriedade em diversos campos do conhecimento, como por exemplo na informática, já que a quantidade de informação presente nos aglomerados destas partículas é enorme, o que torna impraticável a análise dos dados de forma manual. Desde de que Phage-X174 foi sequenciado em 1977 [3], uma quantidade enorme de organismos vêm sendo sequenciados e armazenados em databases. Programas de computador como o BLAST, são usados rotineiramente para pesquisar sequências e a partir de 2008, mais de 260.000 organismos foram analisados, contendo mais de 190 bilhões de nucleotídeos [4] [5].

Por isso, no começo dos anos 70, a necessidade de uma aplicação computadorizada na biologia criou um novo termo, a Bioinformática [6], com o objetivo de definir o estudo de processos computacionais nos sistemas bióticos. Desde então, a bioinformática tornou-se uma parte muito importante para o gerenciamento de dados na medicina moderna [7]. Novas técnicas de bioinformática, como imagem e processamento de sinais, fez possível a extração de resultados significativos a partir de grandes quantidades de dados brutos [8].

No campo da genética e genômica, a bioinformática participa no sequenciamento e anotação de conjuntos de DNA de um organismo, e suas mutações observadas [9]. O DNA é um tipo de ácido nucleico que armazena informações genéticas por meio da combinação de quatro tipos de bases nitrogenadas (adenina (A), timina (T), guanina (G) e citosina (C)), que irão formar moléculas de DNA distintas conforme a sequência. As informações gênicas que coordenam o desenvolvimento e funcionamento de todas as formas de vidas conhecidas no planeta e de alguns vírus estão presentes nos DNA [10], as características hereditárias também são passadas por meio dessa molécula [11].

Por outro lado, o RNA é o outro tipo de ácido nucleico composto por quatro tipos diferentes de subunidades nucleotídicas unidas entre si por ligações fosfodiéster, e têm como uma de suas funções servir de molde para a formação de proteínas. As sequências de RNA são analisadas para determinar quais genes codificam proteínas e, também, para comparar genes dentro de uma espécie ou entre espécies diferentes, o que pode mostrar semelhanças entre funções proteicas ou relações entre espécies. Há várias classes de RNA, como os RNAs mensageiros (mRNAs) que são codificantes de proteínas e os RNAs não codificantes (ncRNAs) que são subdivididos em duas categorias, pequenos e longos [12].

Os RNAs longos não codificantes (lncRNAs) estão emergindo como novos participantes no paradigma do câncer, demonstrando papéis potenciais nas vias oncogênicas e supressoras de tumor [13]. Os RNAs pequenos não codificantes (sncRNAs) fazem parte de oligo-nucleotídeos reguladores não codificantes com amplas funções fisiológicas e morfológicas. Eles controlam a programação genética das células e podem modular processos de diferenciação e morte [14].

De fato cada classe de RNA desempenha um papel ativo e distinto dentro das células, desde catalisar reações biológicas e até controlar a expressão gênica nos seres vivos [15]. Diante disto, diferenciar classes de RNA entre mRNA, lncRNA e sncRNA, pode contribuir para uma maior compreensão de suas funcionalidades e mecanismos.

Objetivo Geral. Este trabalho tem como intuito a produção de um método eficiente e eficaz para a classificação de diferentes classes de RNA, o que permite a análise de dados brutos e em enorme quantidade.

Objetivos Específicos. Para alcançar o objetivo geral, serão considerados os seguintes objetivos específicos:

• Realizar um estudo sobre os métodos de classificação de RNAs, PLEK [16], BASiNET [17] e CPC2 [18].

• Desenvolver um método para a substituição do uso de thresholds do trabalho BASiNET [17].

• Realizar a revisão bibliográfica de artigos relacionados a grafos e produção de dados biológicos em especial, DNA e RNA.

• Realizar estudo geral das aplicações no software R [19].

• Realizar um estudo sobre as ferramentas para produção de grafos e rede complexas.

• Realizar estudo sobre a implementação de um filtro de arestas baseado no método de máxima entropia [20].

• Realizar um estudo sobre a ferramenta de classificação Random Forest [21].

FUNDAMENTAÇÃO TEÓRICA

O objetivo desta seção é apresentar alguns conceitos importantes para um melhor entendimento do trabalho. Na seção Ácidos Nucleicos será exposto informações sobre sequências de DNA e RNA, sua importância para organismos, como elas funcionam e algumas características importantes para este projeto. Em seguida, na seção Redes Complexas será apresentada as estruturas dos grafos e também o funcionamento das redes complexas. Além disso, nesta seção está disponível as Medidas Topológicas que são adotadas neste trabalho para abstrair informações das redes complexas. Por fim, a seção Entropia apresentará o conceito de entropia.

Ácidos Nucleicos. Membros da família de biopolímeros, os ácidos nucleicos são macromoléculas de enorme importância biológica, compostos de nucleotídeos, estes possuem três componentes: um açúcar de 5 carbonos (pentose), um grupo fosfato e uma base nitrogenada. Aliás, as moléculas de DNA são provavelmente as maiores moléculas individuais conhecidas. Por exemplo, algumas moléculas de ácido nucleico biológicas estudadas variam em tamanho de 21 nucleotídeos a grandes cromossomos, enquanto isso o cromossomo humano 1 é uma única molécula que contém 247 milhões de pares de bases nitrogenadas. [22]

Na Tabela 1 é possível observar as bases nitrogenadas que estão presentes nos dois tipos de ácidos nucleicos, a princípio, a Timina (T) pertence apenas ao DNA, bem como a Citosina (C) está presente apenas no RNA, as restantes estão presente em ambos RNA e DNA, sendo estas as diferenças e semelhanças em relação as bases nitrogenadas, no que se diz respeito aos tipos de ácidos nucleicos.

Os ácidos nucleicos são a base essencial para qualquer organismo vivo, uma vez que codificam e armazenam informações de todas as células de todas as formas de vida na Terra. Eles funcionam para permitir o fluxo e expressar as informação de dentro e fora do núcleo da célula para as operações que possibilitam o funcionamento da célula. A informação codificada é amazenada e transmitida através de sequências de bases nitrogenadas, que fornece a ordenação em "escada" de nucleotídeos dentro das moléculas de RNA e DNA. Eles desempenham um papel especialmente importante no direcionamento da síntese de proteínas como é possível observar na Figura 2 e na Tabela 2.

Tabela 1. Bases nitrogenadas contidas nas sequências de DNA e RNA.

	DNA	RNA
Bases Nitrogenadas	Uracila (U)	Citosina (C)
	Timina (T)	Uracila (U)
	Adenina (A)	Adenina (A)
	Guanina (G)	Guanina (G)

Fonte: Autoria Própria.

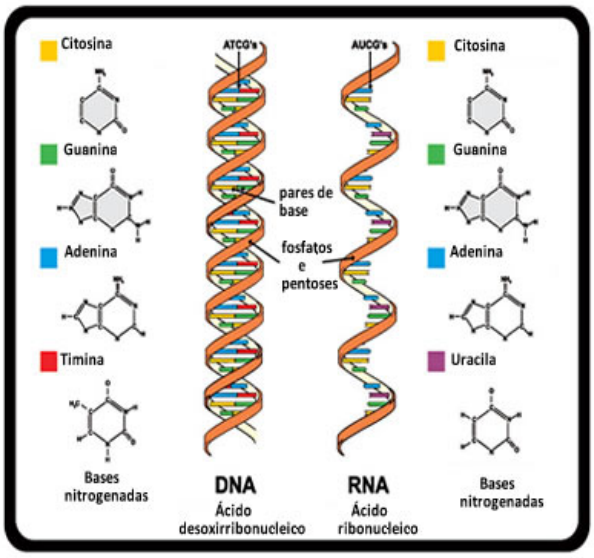

Figura 1. Características do RNA e DNA.

Fonte: [23].

O processo que realiza síntese de proteína em todos os seres vivos é descrito no "Dogma central da biologia molecular", esta tese foi declarada pela primeira vez por Francis Crick em 1957, e publicada em 1958 [24]. Segundo Crick, "O dogma central da biologia molecular lida com a transferência detalhada de resíduo por resíduo da informação sequencial. Afirma que tal informação não pode ser transferida de volta da proteína para proteína ou ácido nucleico"[25]. Ou seja, o dogma é uma explicação para a compreensão da transferência de informações de sequência para a síntese de proteína em organismos vivos.

Como pode ser observado na Figura 2, o DNA pode ser replicado (replicação de DNA), e suas informações podem ser copiadas para um mRNA (transcrição) e as proteínas podem ser codificadas com as informações transmitidas pelo mRNA como modelo (tradução). Além disso, outros tipos de transferências como o RNA sendo copiado de um outro RNA (replicação de RNA), e um DNA sendo sintetizado usando um modelo de outro RNA (transcrição reversa), são processos observados e essenciais nos seres vivos. Diferente dos mRNAs, os ncRNAs não codificam proteínas, contudo possuem outras funções importantes, como a regulação de genes alvos principalmente através da interação ncRNA-mRNA [26] e a transcrição generalizada (vide Tabela 2).

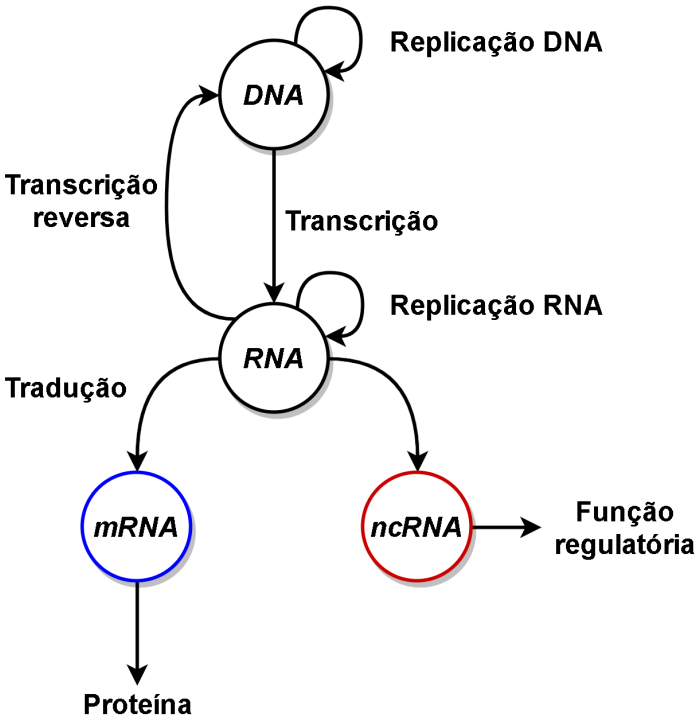

Figura 2. Dogma central da biologia molecular.

Fonte: Autoria própria.

Tabela 2. Classes RNAs e suas funções.

Tipo	Classe	Função
Codificante	mRNA	Contém a sequência para a síntese das proteínas [12]
Não Codificante	sncRNA	Controlam a programação genética das células [14]
	lncRNA	Regulação de genes alvo [15]

Fonte: Autoria Própria.

Redes Complexas. Inter-disciplinares, as redes complexas são grafos com topologias não triviais, que possuem um conjunto de vértices (nós) que são ligados através de arestas [27]. De fato, muitas interações em nosso mundo possuem ligações que podem ser formadas de várias maneiras, por exemplo nas redes sociais, onde pessoas se ligam com outras pessoas e formam-se grupos que também se ligam a outros grupos de pessoas.

Resumidamente, um grafo pode ser visto como um conjunto de pontos, chamados de vértices e quando um vértice se junta a outro vértice é formada uma ligação, definida como aresta. Dependendo da aplicação, arestas podem ou não ter direção, pode ser permitido ou não arestas ligarem um vértice a ele próprio e vértices e/ou arestas podem ter um peso (numérico) associado.

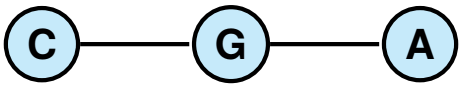

Figura 3. Grafo gerado a partir da sequência: CGA.

Fonte: Autoria Própria.

Tabela 3. Matriz de adjacências do grafo da Figura 3.

	A	***C***	**T**	**G**
A	0	0	0	1
C	0	0	0	1
T	0	0	0	0
G	1	1	0	0

Ao existir uma aresta entre dois vértices, estes são considerados adjacentes. No grafo da Figura 3 os vértices C e A não são adjacentes, entretanto os vértices C e G, e, G e A são. Na computação, um grafo direcionado ou não-direcionado é geralmente representado por sua matriz de adjacência, como na Tabela 3. Isto é, uma matriz n-por-n cujo valor na linha i e coluna j fornece o peso de arestas do i-ésimo ao j-ésimo vértices [28].

Medidas Topológicas. As redes complexas possuem topologias bem definidas que descrevem a estrutura da rede, diante disso é possível extrair medidas para abstrair tais características [29]. Para a abstração das características, neste trabalho foram separadas 10 medidas usadas na literatura atual, cada uma das medidas abaixo serão aplicadas nas redes, permitindo a classificação das sequências.

- Caminho Mínimo Médio: é a média de todos os caminhos mínimos de uma rede [29]. O valor para o caminho mínimo médio é baixo para redes muito concentradas.
- Coeficiente de Cluster: ou transitividade, tem como propósito calcular a probabilidade de vértices que estão conectados a um outro vértice estarem também conectados entre si. Por exemplo, a aplicação dessa medida em uma rede social teria como efeito o cálculo da probabilidade de duas pessoa que conhecem uma terceira pessoa também se conhecerem. Em outras palavras, a transitividade pode ser vista como a formação de triângulos nas redes [29].
- Máximo: é a medida que apresenta o vértice com o maior número de arestas conectadas.
- Mínimo: análoga a medida acima, contudo apresenta o vértice com o menor número de arestas conectadas.
- Degree: tem como objetivo o retorno da média de conexões que os vértices da rede possuem.
- Assortatividade: é a probabilidade de vértices com graus parecidos estarem conectados. Valores de assortatividade positivos significam que vértices de graus semelhantes tendem a se conectar e valores negativos de assortatividade significa o contrário [30][31].
- Desvio Padrão: cada nó da sequência possui um número de ligações que varia de sequência para sequência. Sabendo disso, o desvio padrão é calculado pelo número de ligações feitas por cada nó, valores altos de desvio padrão mostram que a sequência possui números de ligações desbalanceadas, ou seja, há nós que se conectam mais que outros.
- Intermediação: um vértice qualquer, independente dos seus número de ligações, pode ter

grande importância na formação de uma rede, pois um vértice pode estar posicionado nos caminhos geodésicos, isto é, na menor distância que une dois pontos. Sabendo disso a intermediação tem como objetivo fazer a medição da centralidade de cada nó em uma rede [32].

- Motivo de tamanho 3: também chamado de motifs em inglês, motivos são sub-redes com vários formatos que existem dentro de uma rede. Esta medida apresenta a contagem de quantos motifs de tamanho 3 ocorrem dentro de uma rede.
- Motivo de tamanho 4: esta medida apresenta a contagem de quantos motifs de tamanho 4 ocorrem dentro de uma rede.

Entropia. O conceito de entropia vem sendo usado em diversos campos do conhecimento, desde a termodinâmica clássica, onde foi inicialmente descoberto, até a física estatística e os princípios da teoria da informação. Com o tempo, este termo, encontrou aplicações de longo alcance na química e na física, e atualmente a medida entropia também participa no estudo de sistemas biológicos e sua relação com a vida.

Em 1948, o cientista Claude Shannon, desenvolveu conceitos estatísticos semelhantes de medição da incerteza microscópica e multiplicidade ao problema de perdas aleatórias de informação em sinais de telecomunicação [33]. Shannon chamou essa entidade de informação ausente de maneira semelhante ao seu uso na mecânica estatística como entropia, e deu origem ao campo da teoria da informação [33]. Esta descrição foi proposta como uma definição universal do conceito de entropia, e também muitas vezes chamada de entropia de Shannom, em sua homenagem.

A entropia prediz que certos processos são irreversíveis ou impossíveis, além da exigência de não violar a conservação da energia, sendo esta última expressa na primeira lei da termodinâmica. A entropia é central para a segunda lei da termodinâmica, que afirma que a entropia dos sistemas isolados deixados para a evolução espontânea não pode diminuir com o tempo, pois eles sempre chegam a um estado de equilíbrio termodinâmico, onde a entropia é maior.

Com o tempo, surgiram diversos métodos com diferentes funcionalidades usando o conceito de entropia, um deles foi o desenvolvimento do método da máxima entropia (ME) pelo físico Edwin Thompson Jaynes. Ele mostrou que maximizar estatisticamente a entropia com o objetivo de observar o modo como as moléculas de gás estavam distribuídas seria equivalente à simples maximização da entropia de Shannon [33] [34]. Esta observação conduziu a utilização do método da máxima entropia para atribuir probabilidades dentro de um sistema.

A máxima entropia procura identificar a distribuição das probabilidades de um sistema. Com isso, é possível encontrar o ponto onde a entropia está maximizada, indicando a posição da separabilidade de duas partes distintas de um conjunto dados ou de informações [34].

Em 1985, Kapur [20] desenvolveu um algoritmo de limiarização por máxima entropia, que considera um histograma de uma imagem como uma distribuição de probabilidades. O limiar é escolhido como a intensidade que maximiza a soma das entropias de duas partes distintas (objeto e fundo) da imagem [35] (vide Figuras 4, 5 e 6). Este valor não está relacionada com a disposição espacial dos pixels na imagem, o que significa que os pixels poderiam ser reorganizados de qualquer outra forma, e o resultado permaneceria o mesmo.

Figura 4. Foto original.

Figura 5. Foto limiarizada em $T = 75$.

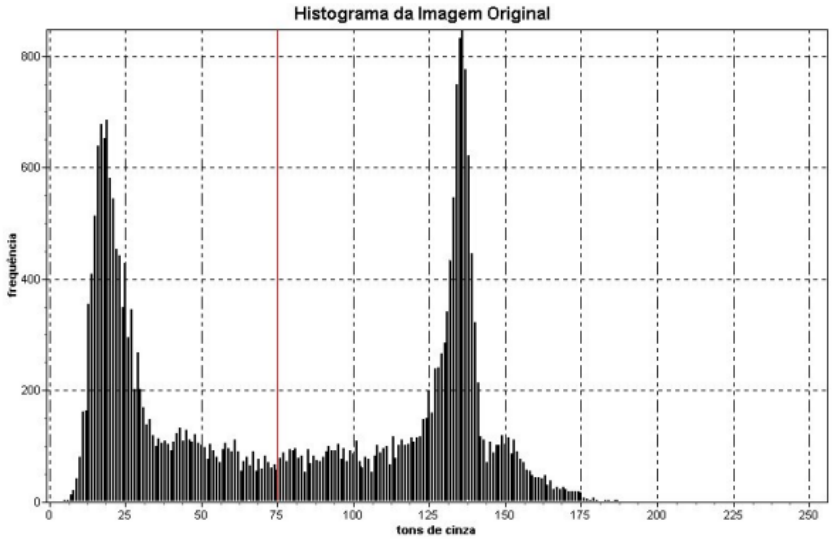

Figura 6. Exemplo da aplicação do método de Máxima entropia na limiarização de uma imagem.

Fonte: [35].

Trabalhos relacionados. Foram desenvolvidos diversos métodos para a classificação de sequências de RNA na literatura, entre eles estão o PLEK [16], o CPC2 [18], e o mais recente, o BASiNET [17].

O CPC2 [18] teve como objetivo diferenciar rapidamente e de uma forma bastante precisa as transcrições de RNA. Seu método se baseia na seleção de características hierárquicas das sequências, usando-as para identificar as características mais efetivas com os classificadores Random Forest (com validação cruzada). Além disso, a sua proposta foi a neutralidade em relação a espécie, tornando o método viável para transcrições de organismos não modelo, que por sua vez estão em constante crescimento.

Por outro lado, o método PLEK [16] teve como proposta uma tecnologia de sequenciamento de transcritos de RNA, com o objetivo de descobrir novos transcritos codificadores e não codificantes de proteínas, com enfoque na identificação de RNAs longos não codificantes (lncRNA). Teve como modelo de implementação o uso de k-mer e janelas deslizantes com

um comprimento de etapa de um nucleotídeo. Seu código pode ser baixado gratuitamente em https://sourceforge.net/projects/plek/files/.

O mais recente dentre os citados, BASiNET [17] desenvolveu uma ferramenta sem alinhamento para classificar sequências biológicas baseada no uso de thresholds. Uma técnica que consiste em cortar repetidamente uma rede complexa até o corte da aresta mais conectada (máximo de 200 cortes). Estas redes complexas são utilizadas para a extração de 10 medidas topológicas distintas, sendo que, para cada corte da rede, as mesmas 10 medidas devem ser extraídas, produzindo assim, até 2000 diferentes características para cada sequência. Ao ser utilizada para classificar as sequências de RNAs, essa quantidade massiva de dados amplifica o tempo de processamento e de memória necessários.

MATERIAS E MÉTODOS

O objetivo desta seção é demonstrar a metodologia usada neste trabalho. Em Conjunto de Dados mostrará os dados utilizados e suas referentes quantidades por meio das Tabelas 4 e 5. Na seção Método será apresentado os passos que foram necessários para o desenvolvimento do método proposto neste trabalho.

Conjunto de Dados. Os Datasets de RNAs selecionadas para os testes foram as mesmas dos trabalhos CPC2 [18] e PLEK [16]. Os quais obtiveram RNAs de diferentes espécies, como é possível ser visto nas Tabelas 4 e 5. Em relação as classes de RNA, foram selecionadas três classes, o RNA mensageiro (mRNA), o RNA longo não codificante (lncRNA) e o RNA pequeno não codificante (sncRNA).

Tabela 4. Dataset do CPC2 [18].

Espécies	*Classes*	*Transcrições*
Arabidopsis thaliana	mRNA	15931
	ncRNA	3853
Homo sapiens	mRNA	6142
	ncRNA	12015
Danio Rerio	mRNA	2344
	ncRNA	1528
Drosophila melanogaster	mRNA	3680
	ncRNA	3556
Mus musculus	mRNA	10638
	ncRNA	12251
Caenorhabditis elegans	mRNA	3551
	ncRNA	1582

Tabela 5. Dataset do PLEK [16].

Especies	*Classes*	*Transcrições*
Mus musculus	mRNA	26062
	ncRNA	2963
Danio rerio	mRNA	14493
	ncRNA	419
Xenopus tropicalis	mRNA	8874
	ncRNA	279
Bos taurus	mRNA	13190
	ncRNA	182
Pan troglodytes	mRNA	1906
	ncRNA	1166
Sus scrofa	mRNA	3978
	ncRNA	241
Macaca mulatta	mRNA	5709
	ncRNA	359
Gorilla gorilla	mRNA	33025
	ncRNA	367
Pongo abelii	mRNA	3401
	ncRNA	392

Método. Para o desenvolvimento deste trabalho foi necessário estudo e adaptação do código do BASiNET [17]. Por isso, foi utilizado grafos complexos (redes complexas), criados a partir de sequências de RNA (contidas nos Datasets de [16] e [18]) armazenadas em arquivos do tipo FASTA. Com o pacote Biostrings [36] no R, é possível carregar as sequências, como no Código 1

```
MRNA<-readBStringSet(mRNA)
LNCRNA<-readBStringSet(lncRNA)
SNCRNA<-readBStringSet(sncRNA)
```

Código 1. Leitura dos arquivos FASTA.

Para a produção destas redes complexas, foi necessário a configuração de dois parâmetros, o Passo e a Palavra. A função do passo é definir a distância que será percorrida na sequência depois que uma aresta foi formada para a formação de uma nova aresta. Enquanto isso, a Palavra se refere a quantos nucleotídeos serão agrupados em cada vértice. Este processo é melhor descrito na Figura 7. Neste projeto, semelhante ao trabalho BASiNET, usaremos $passo = 1$ e $palavra = 3$.

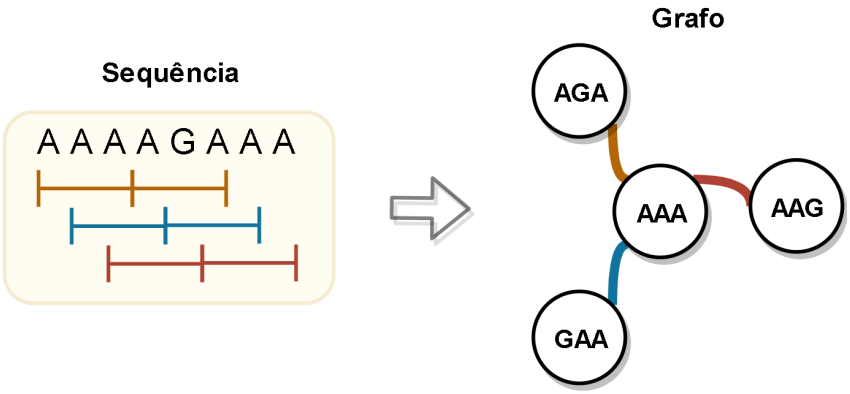

Figura 7. Esquema explicativo da conversão de sequências de RNA em grafos.

Fonte: Autoria Própria.

Com os grafos produzidos, foi necessário o desenvolvimento de uma ferramenta que selecionasse as arestas mais relevantes para a diferenciação das classes de RNA. O método de limiarização de Kapur [20] baseado na máxima entropia, ao ser aplicado a um histograma de uma imagem equalizada, encontra o limiar que maximiza a soma das entropias (Seção Entropia) de duas partes distintas (objeto e fundo) de uma imagem [35]. Análogo a este pensamento, para este trabalho, usaremos este método para a produção de um filtro de arestas, isto é, ao analisar um espectro de 4096 possibilidades de arestas, resultante de uma matriz de 64x64 vértices, separar o que é importante (objeto), e o que não é importante (fundo) para a classificação de classes de RNA.

Este espectro de 4096 possibilidades de arestas será convertido em um histograma, o qual será utilizado para o encontrar o limiar, ou seja, o ponto (aresta) com a máxima entropia no histograma. Para este propósito, é necessário o cálculo de P0 e P1, probabilidades das duas partes distintas, esquerda e direita respectivamente, de cada ponto no histograma (4096 pontos).

Ao obter P0 e P1 de cada ponto dos 4096 presentes no histograma, é também, calculado as entropias referentes (H0 e H1). Deste modo, para cada ponto, uma entropia total chamada de HTotal (H0 + H1) é encontrada, a qual é comparada com todas as outras entropias encontradas, selecionando o ponto com a maior entropia no histograma (vide Figura 8). Assim é possível identificar automaticamente o ponto de separação entre os dois subconjuntos de arestas, com maior e menor informação para a classe. Por exemplo, a Figura 9 apresenta a curva das somas das entropias referente a classe lncRNA da espécie *Gorilla gorilla*, no qual é encontrado o limiar no ponto 749. Diante disso, as arestas referentes aos pontos de 1 até 749 serão selecionadas, e as restantes serão descartadas.

As arestas selecionadas pela máxima entropia são organizadas em uma lista de matrizes, isto é, uma matriz para cada classe de RNA. Com isso, é possível filtrar os grafos das sequências desejadas de RNA, este processo de filtragem está explicado na Figura 10.

A função “Classificar” no Código 2 tem como objetivo diferenciar as sequências entre mRNA, lncRNA e sncRNA, as sequências de entrada por meio das matrizes de arestas que foram selecionados pela função “SelecionarArestas”. Deste modo é possível filtrar cada grafo de acordo com sua matriz de arestas correspondente, e assim, extrair as medidas topológicas dos grafos filtrados (vide Seção Medidas Topológicas) e organizá-las em um Dataframe. A metodologia geral está descrita no fluxograma da Figura 11.

```
ListaMatrizes<-SelecionarArestas(mRNA, lncRNA, sncRNA)
Resultado<-Classificar(mRNA, lncRNA, sncRNA ,ListaMatrizes)
```

Código 2. Funções “Classificar” e “SelecionarArestas”.

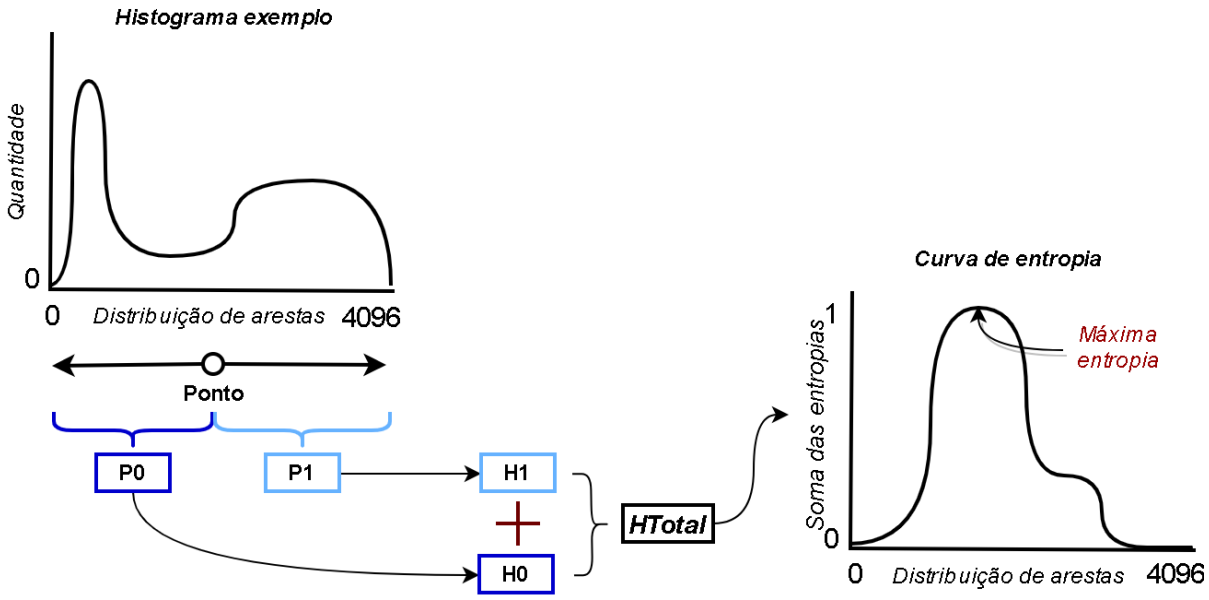

Figura 8. Exemplo da produção da curva de entropia.

Fonte: Autoria Própria.

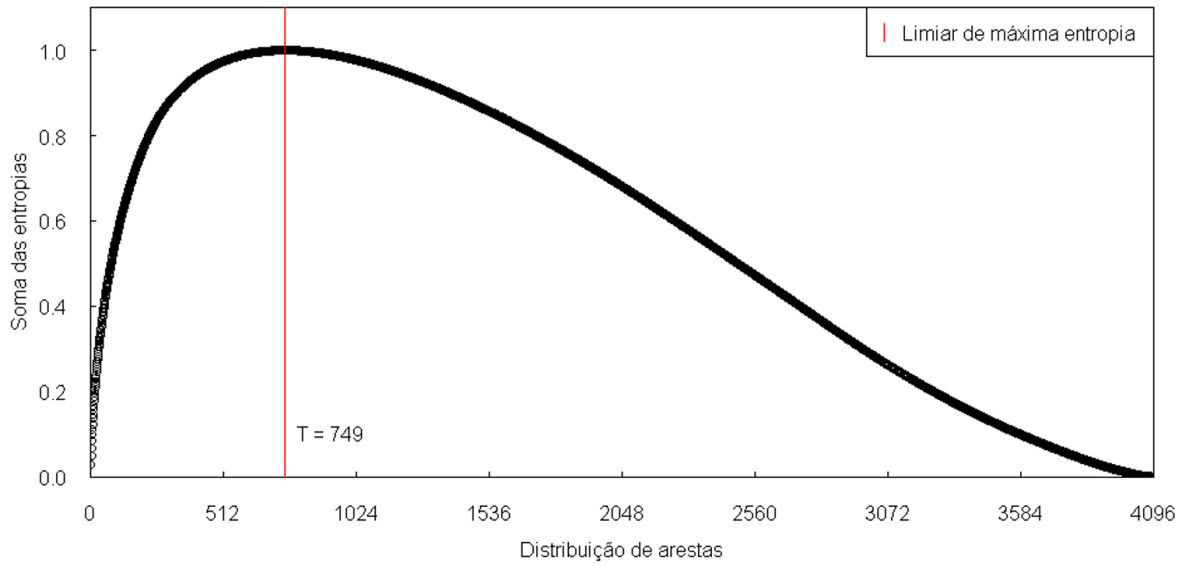

Figura 9. Curva da soma das entropia da classe lncRNA (*Gorilla gorilla*).

Fonte: Autoria Própria.

Dentro do Dataframe, cada medida produzida possui um alcance de valor diferente, por exemplo, a medida ASPL (Caminho Mínimo Médio) usualmente se encontra na escala da dezena.

Enquanto isso, outras medidas podem chegar a valores na escala de centenas ou até milhares, o que torna algumas medidas mais relevantes que outras para o classificador. Logo, por meio de um pré processamento, é feito um reajuste nos valores das medidas para que elas estejam entre 0 e 1, permitindo uma classificação mais consistente. A Equação 1 exemplifica como é reajustado os valores, sendo V referente ao valor, Min referente ao valor mínimo da classe e Max ao valor máximo da classe.

$$V\,(reajustado) = \frac{V - Min}{Max - Min} \tag{1}$$

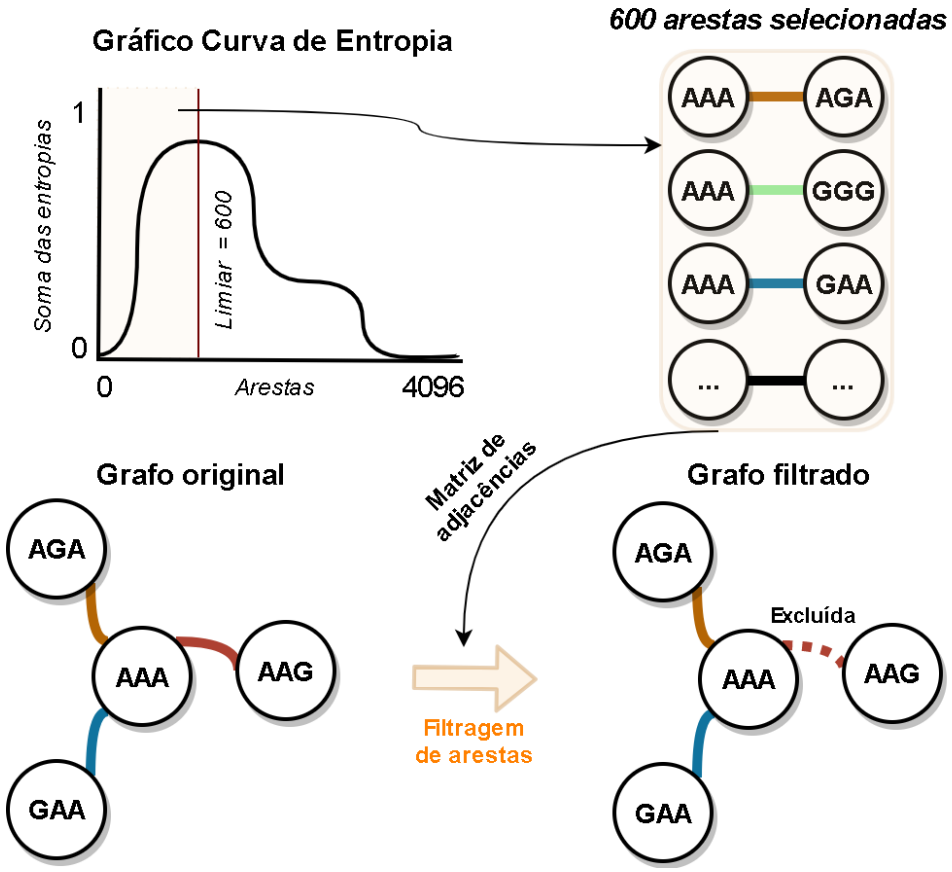

Figura 10. Esquema explicativo da filtragem das arestas através da máxima entropia.

Fonte: Autoria Própria.

A classificação das sequências por meio das medidas topológicas contidas no Dataframe foi executada pela ferramenta Random Forest, presente na biblioteca “randomForest” [21] no software R [19]. Bem como pelo pacote “rfUtilities” [37], com o objetivo de realizar a validação cruzada por meio de uma função previamente pronta “rf.crossvalidation” . Esta técnica consiste em executar divisões percentuais repetidas. Em outras palavras, dividir um conjunto de dados em 5 pedaços e usar um pedaço para teste e avaliar os 4 restantes juntos (até que todos os pedaços tenham sido testados). A validação cruzada tem como proposta a diminuição da estimativa de erro da classificação. Para este projeto foi usado o numero de pedaços igual a 10 ($n = 10$).

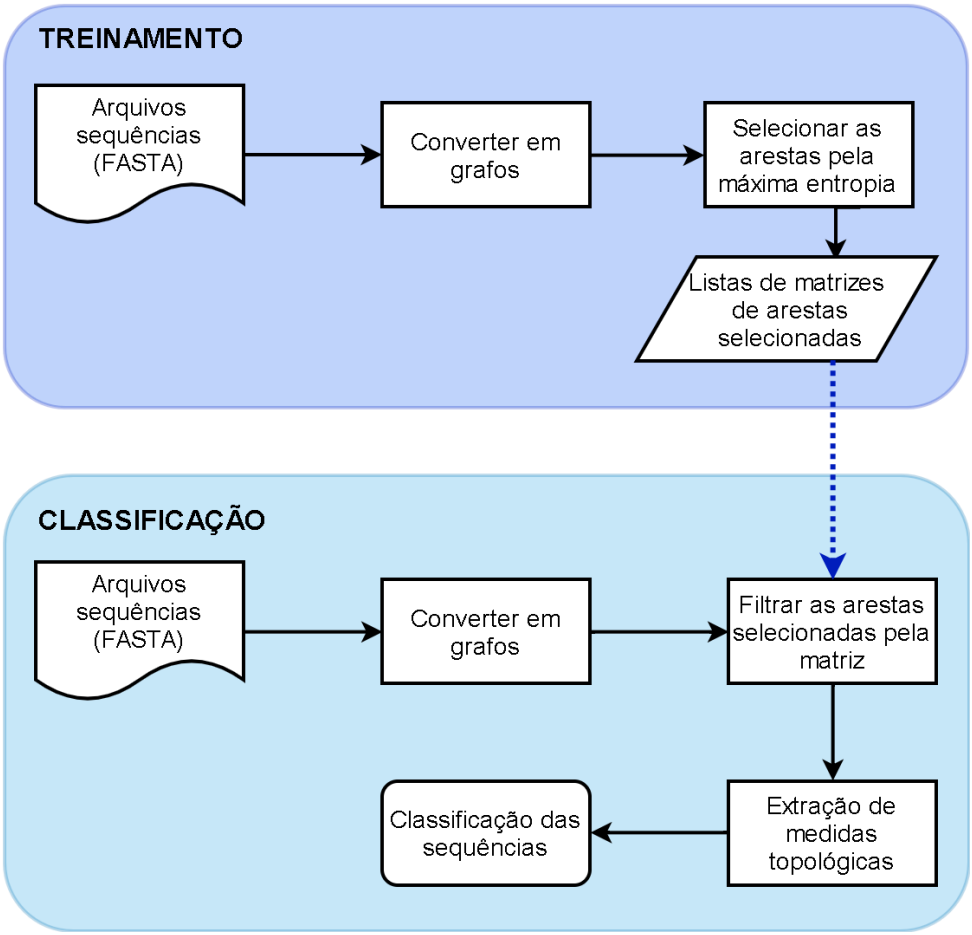

Figura 11. Fluxograma geral da metodologia.

Fonte: Autoria Própria.

RESULTADOS E DISCUSSÃO

Nesta seção será apresentado os resultados dos testes do método proposto por este trabalho. Sendo assim, foram utilizadas os dois datasets apresentadas na subseção Conjunto de Dados. Por último, serão apresentados os testes dois datasets em conjunto, comparando com o classificador PLEK [16].

A Tabela 6 apresenta os resultados das classificações referentes aos mRNAs, lncRNAs e sncRNAs, a base de dados usada foi a mesma utilizada no trabalho PLEK [16]. Porém, devido a limitação dos outros métodos, as classes foram organizadas entre mRNA e ncRNA. Os resultados do método proposto foram comparados com os resultados dos classificadores PLEK [16], CPC2 [18], BASiNET* (BASiNET sem threshold) e BASiNET [17].

Os resultados do método proposto em comparação aos classificadores CPC2, PLEK e BASiNET* foram satisfatórios, visto que, foi superior em acurácia tanto por espécies, quanto na média geral e no desvio padrão (vide a Tabela 6). Entretanto, ao comparar com o BASiNET [17], os resultados foram bastante semelhantes, com uma pequena vantagem para o método proposto na média geral e no desvio padrão na classe ncRNA, enquanto o BASiNET foi superior na classe mRNA. A Figura 12 apresenta um gráfico que melhor representa a comparação dos métodos em relação as espécies.

Tabela 6. Taxas de acerto na classificação das sequências de mRNA, ncRNA, por meio do método proposto para diferentes espécies (dataset PLEK).

Espécies	Classe	PLEK	CPC2	BASiNET*	BASiNET	Método Proposto
Mus musculus	mRNA	88.1	94.7	79.9	100.0	99,9
	ncRNA	89.9	99.9	75.9	99.9	100
Danio rerio	mRNA	91.3	96.6	99.8	100.0	100
	ncRNA	90.9	94.0	47.9	98.9	100
Xenopus tropicalis	mRNA	94.5	96.5	99.1	100.0	100
	ncRNA	100.0	100.0	95.0	100.0	100
Bos taurus	mRNA	94.8	95.9	91.2	100.0	100
	ncRNA	99.5	100.0	99.8	98.9	99.5
Pan troglodytes	mRNA	87.1	93.9	97.7	100.0	100
	ncRNA	99.9	100.0	88.3	99.8	100
Sus scrofa	mRNA	85.1	94.9	99.3	99.9	100
	ncRNA	98.3	98.3	76.3	99.6	100
Macaca mulatta	mRNA	85.0	94.2	99.1	100.0	100
	ncRNA	100.0	100.0	94.7	100.0	100
Gorilla gorilla	mRNA	83.8	91.6	97.6	100.0	99,7
	ncRNA	99.7	100.0	97.6	100.0	100
Pongo abelii	mRNA	98.0	94.4	99.9	100.0	99,9
	ncRNA	100.0	100.0	98.7	99.2	100
Média por classe	mRNA	89.74	94.75	95.96	99.99	99.94
	ncRNA	97.58	99.13	86.02	99.59	99.94
Média geral	–	**93.66**	**96.94**	**91.20**	**99.79**	**99.94**
Desvio padrão	mRNA	4.81	1.45	6.21	0.03	0.09
	ncRNA	3.88	1.89	15.97	0.44	0.16

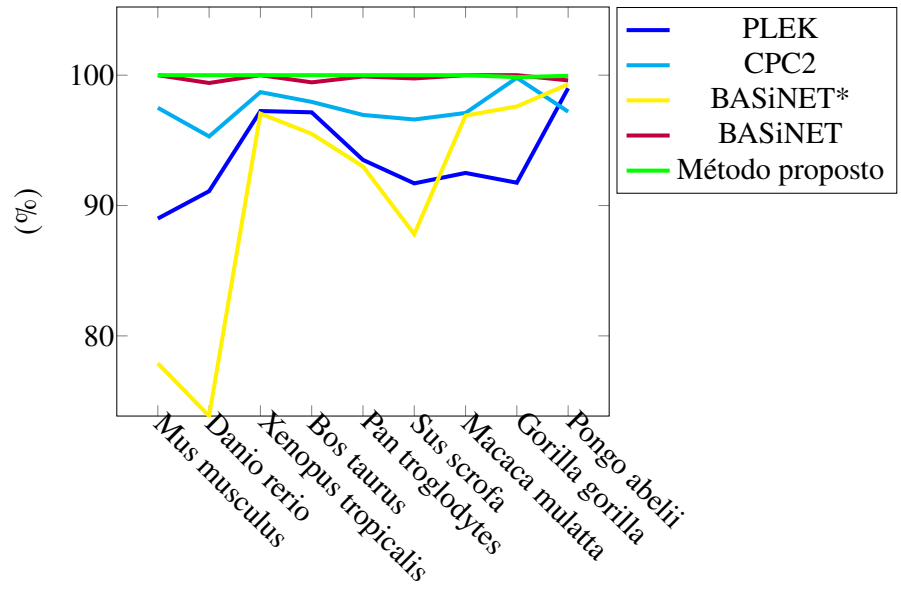

Figura 12. Gráfico para comparação dos classificadores por espécie (dataset PLEK).

Tabela 7. Taxas de acerto na classificação das sequências de mRNA, lncRNA e sncRNA, por meio do método proposto para diferentes espécies (dataset CPC2).

Espécies	Classe RNA	PLEK	CPC2	BASiNET *	BASiNET	Método proposto
Homo sapiens	mRNA	97.0	95.9	80.25	100.0	99.9
	lncRNA	97.6	92.8	58	100.0	99.4
	sncRNA	100.0	100.0	53.7	100.0	99.9
Mus musculus	mRNA	89.2	93.9	89.2	100.0	99.7
	lncRNA	91.7	95.0	99.8	99.9	100.0
	sncRNA	100.0	100.0	72.14	99.9	99.9
Danio rerio	mRNA	94.4	95.5	93.3	99.5	99.2
	lncRNA	79.2	88.1	99.5	98.9	100.0
	sncRNA	100.0	100.0	79.2	98.7	98.7
Drosophila melanogaster	mRNA	82.8	94.6	99.9	98.5	99.0
	lncRNA	87.5	91.9	84.6	97.3	100.0
	sncRNA	100.0	100.0	77.4	99.7	98.8
Caenorhabditis elegans	mRNA	53.0	96.5	77.8	100.0	98.8
	lncRNA	98.4	99.9	100	99.4	99.7
	sncRNA	100.0	100.0	88.7	99.9	99.9
Arabidopsis thaliana	mRNA	63.1	99.7	96.8	99.7	97.8
	lncRNA	99.6	95.3	86.8	99.7	99.8
	sncRNA	100.0	100.0	97.3	100.0	97.8
Média geral	**mRNA**	**79.92**	**96.02**	**89.52**	**99.62**	**99.07**
	ncRNA	**96.17**	**96.92**	**83.09**	**99.45**	**99.49**
Desvio padrão	mRNA	17.92	2.03	8.15	0.58	0.68
	ncRNA	6.67	4.18	15.13	0.81	0.67

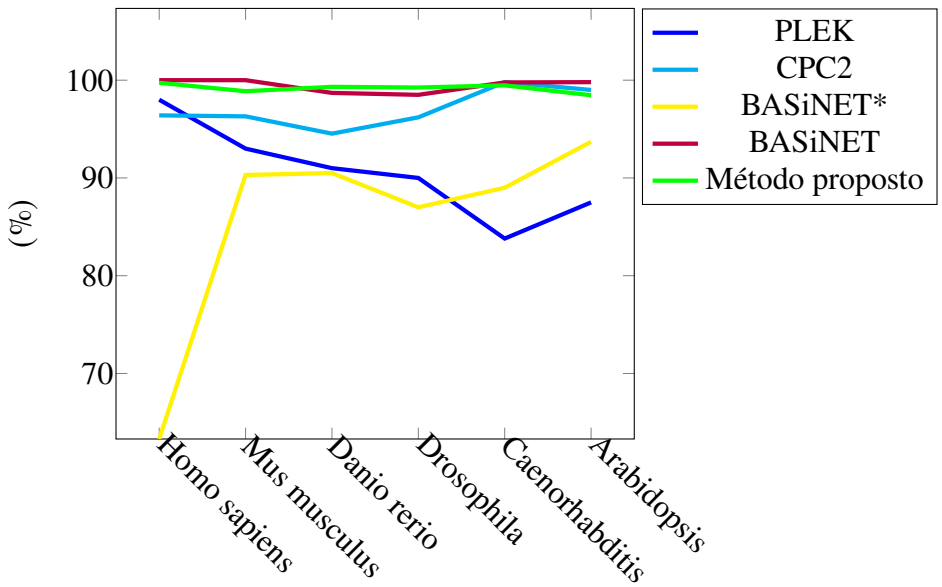

Figura 13. Gráfico para comparação dos classificadores por espécie (dataset CPC2).

A Tabela 7 apresenta os resultados das classificações referentes aos mRNAs, lncRNAs e sncRNAs, a base de dados usada foi a mesma utilizada no trabalho CPC2 [18]. Os resultados do método proposto foram comparados com os resultados dos classificadores PLEK [16], CPC2 [18], BASiNET sem threshold (BASiNET*), e BASiNET [17].

Novamente, os resultados do método proposto foram superiores em acurácia na comparação entre os classificadores CPC2, PLEK e BASiNET* (vide a Tabela 7 e a Figura 13). Contudo, ao comparar com o BASiNET [17], os resultados foram bastante semelhantes, com uma pequena vantagem para o BASiNET na média geral e no desvio padrão na classe mRNA, enquanto o método proposto foi superior na classe ncRNA.

Quanto ao tempo de processamento, no dataset do CPC2, o método proposto executou a classificação das espécies em 246,6 minutos, em média. Uma redução de 5,3% em comparação ao BASiNET, que executou a classificação das mesmas sequências em 260,4 minutos, sendo que o BASiNET sem threholds executou em 199,8 minutos.

Isso se deve principalmente à relação do número de características que cada método utiliza (10 em contraste com até 2000 características do BASiNET) (vide Subseção Trabalhos relacionados). Indicando que a maximização de entropia reduz a complexidade em termos da dimensionalidade das características, o que simplifica o problema e mantém uma alta assertividade na classificação. Logo, o método de máxima entropia foi adequado para diminuir o tempo de processamento e manter as taxas de acurácias em comparação ao uso de thresholds pelo BASiNET.

Para finalizar os testes, foi escolhida uma metodologia de validação cruzada entre datasets, isto é, produzir as matrizes das arestas selecionadas (treinamento) com o o dataset do CPC2 e realizar a classificação com o dataset do PLEK, com o objetivo de verificar o funcionamento da fase de treinamento. Contudo, para o teste foram selecionadas apenas as espécies que se intersectam entre os datasets, sendo assim, as espécies selecionadas foram a *Danio rerio*, a *Mus musculus* e apesar de não ser apresentada nos testes da dataset do PLEK, a *Homo sapiens*, que está contida nos arquivos em `https://sourceforge.net/projects/plek/files/`.

Na Tabela 8 é possível observar que o método foi superior nas médias por classe, na média geral e no desvio padrão em comparação ao PLEK. Por isso, o método de treinamento deste trabalho foi consolidado, já que mesmo utilizando sequências de diferentes datasets, foi possível classificar as sequências com uma alta taxa de acurácia.

Tabela 8. Taxas de acerto na classificação das sequências de mRNA e ncRNA, usando o dataset do CPC2 para produzir as matrizes com as arestas selecionadas, e classificando com o dataset do PLEK.

Espécies	Classe	PLEK	Método proposto
Homo sapiens	mRNA	90,0	99,7
	ncRNA	55,0	100,0
Danio Rerio	mRNA	100,0	99,7
	ncRNA	40,3	100.0
Mus musculus	mRNA	91,6	99,9
	ncRNA	95,8	97,0
Média por classe	mRNA	93.85	99.79
	ncRNA	63.70	98.98
Média geral	**–**	**78.78**	**99.39**
Desvio padrão	mRNA	4.40	0.11
	ncRNA	23.48	1.41

CONCLUSÃO

Com os resultados observados nos testes dos dois datasets PLEK e CPC2, bem como a comparação com outros métodos com objetivos semelhantes, fica evidente que a metodologia adotada neste trabalho foi adequada para seu propósito. Ainda assim, o método do BASiNET teve taxas da acurácia semelhantes, porém com um tempo de processamento maior. Por isso, a adaptação dos métodos de máxima entropia e de limiarização de Kapur, para selecionar as arestas com maior importância, se provou eficaz e eficiente na classificação de diferentes classes de RNA em comparação ao uso de thresholds.

Do mesmo modo, a extração de medidas topológicas de uma rede complexa gerada por uma sequência, pode contribuir para um melhor entendimento e diferenciação dos RNAs. Além disso, este método pode ser aplicado a outras sequências biológicas como as de DNAs, que possuem uma estrutura semelhante aos RNAs, auxiliando assim, a análise dos crescentes dados gerados nos últimos anos.

AGRADECIMENTOS

Os autores agradecem a Universidade Tecnológica Federal do Paraná (UTFPR), pela bolsa de Iniciação Científica (PIBIC 2020/2021), concedida ao acadêmico Murilo Montanini Breve.

REFERÊNCIAS

[1] DAHM, R. Friedrich Miescher and the discovery of DNA. *Developmental Biology*, v. 278, n. 2, p. 274–288, 2005. ISSN 0012-1606. DOI: https://doi.org/10.1016/j.ydbio.2004.11.028. Disponível em: http://www.sciencedirect.com/science/article/pii/S0012160604008231.

[2] FEUGHELMAN, M. et al. Molecular Structure of Deoxyribose Nucleic Acid and Nucleoprotein. *Nature*, v. 175, n. 4463, p. 834–838, mai. 1955. ISSN 1476-4687. DOI: 10.1038/175834a0. Disponível em: https://doi.org/10.1038/175834a0.

[3] SANGER, F. et al. Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature*, v. 265, n. 5596, p. 687–695, fev. 1977. ISSN 1476-4687. DOI: 10.1038/265687a0. Disponível em: https://doi.org/10.1038/265687a0.

[4] BENSON, D. A. et al. GenBank. eng. *Nucleic acids research*, Oxford University Press, v. 36, Database issue, p. d25–d30, jan. 2008. gkm929[PII]. ISSN 1362-4962. DOI: 10.1093/nar/gkm929. Disponível em: https://doi.org/10.1093/nar/gkm929.

[5] JEFFARES, D. C.; POOLE, A. M.; PENNY, D. Relics from the RNA World. *Journal of Molecular Evolution*, v. 46, n. 1, p. 18–36, jan. 1998. ISSN 1432-1432. DOI: 10.1007/PL00006280. Disponível em: https://doi.org/10.1007/PL00006280.

[6] HOGEWEG, P. The Roots of Bioinformatics in Theoretical Biology. *PLOS Computational Biology*, Public Library of Science, v. 7, n. 3, p. 1–5, mar. 2011. DOI: 10.1371/journal.pcbi.1002021. Disponível em: https://doi.org/10.1371/journal.pcbi.1002021.

[7] JOAQUIM, L. M.; EL-HANI, C. N. A genética em transformação: crise e revisão do conceito de gene. pt. *Scientiae Studia*, scielo, v. 8, p. 93–128, mar. 2010. ISSN 1678-3166. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1678-31662010000100005&nrm=iso.

[8] BAXEVANIS, A. D.; BADER, G. D.; WISHART, D. S. *Bioinformatics*. [S.l.]: John Wiley & Sons, 2020.

[9] DAVIES, K. *Decifrando o Genoma*. [S.l.]: Companhia das Letras, 2001.

[10] VARKI, A. Comparing the human and chimpanzee genomes: Searching for needles in a haystack. *Genome Research*, Cold Spring Harbor Laboratory, v. 15, n. 12, p. 1746–1758, dez. 2005. DOI: 10.1101/gr.3737405. Disponível em: https://doi.org/10.1101/gr.3737405.

[11] WOLF, Y. I. et al. Genome trees and the tree of life. *Trends in Genetics*, Elsevier BV, v. 18, n. 9, p. 472–479, set. 2002. DOI: 10.1016/s0168-9525(02)02744-0. Disponível em: https://doi.org/10.1016/s0168-9525(02)02744-0.

[12] ALBERTS. *Biologia molecular da célula*. 6. ed. [S.l.]: Artmed Editora, 2009.

[13] GIBB, E. A.; BROWN, C. J.; LAM, W. L. The functional role of long non-coding RNA in human carcinomas. *Molecular Cancer*, v. 10, n. 1, p. 38, abr. 2011. ISSN 1476-4598. DOI: 10.1186/1476-4598-10-38. Disponível em: https://doi.org/10.1186/1476-4598-10-38.

[14] KLIMENKO, O. V. Small non-coding RNAs as regulators of structural evolution and carcinogenesis. *Non-coding RNA Research*, v. 2, n. 2, p. 88–92, 2017. ISSN 2468-0540. DOI: https://doi.org/10.1016/j.ncrna.2017.06.002. Disponível em: https://www.sciencedirect.com/science/article/pii/S2468054017300215.

[15] RINN, J. L.; CHANG, H. Y. Genome regulation by long noncoding RNAs. eng. *Annual review of biochemistry*, v. 81, p. 145–166, 2012. PMC3858397[pmcid]. ISSN 1545-4509. DOI: 10.1146/annurev-biochem-051410-092902. Disponível em: https://doi.org/10.1146/annurev-biochem-051410-092902.

[16] LI, A.; ZHANG, J.; ZHOU, Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics*, v. 15, n. 1, p. 311, set. 2014. ISSN 1471-2105. DOI: 10.1186/1471-2105-15-311. Disponível em: https://doi.org/10.1186/1471-2105-15-311.

[17] ITO, E. A. et al. BASiNET—BiologicAl Sequences NETwork: a case study on coding and non-coding RNAs identification. *Nucleic Acids Research*, v. 46, n. 16, e96–e96, jun. 2018. ISSN 0305-1048. DOI: 10.1093/nar/gky462. eprint: https://academic.oup.com/nar/article-pdf/46/16/e96/25802486/gky462.pdf. Disponível em: https://doi.org/10.1093/nar/gky462.

[18] KANG, Y.-J. et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research*, v. 45, W1, w12–w16, mai. 2017. ISSN 0305-1048. DOI: 10.1093/nar/gkx428. eprint: https://academic.oup.com/nar/article-pdf/45/W1/W12/23741208/gkx428.pdf. Disponível em: https://doi.org/10.1093/nar/gkx428.

[19] R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2014. Disponível em: http://www.R-project.org/.

[20] KAPUR, J.; SAHOO, P.; WONG, A. A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image Processing*, v. 29, n. 3, p. 273–285, 1985. ISSN 0734-189X. DOI: https://doi.org/10.1016/0734-189X(85)90125-2. Disponível em: https://www.sciencedirect.com/science/article/pii/0734189X85901252.

[21] LIAW, A.; WIENER, M. Classification and Regression by randomForest. *R News*, v. 2, n. 3, p. 18–22, 2002. Disponível em: https://CRAN.R-project.org/doc/Rnews/.

[22] GREGORY, S. G. et al. The DNA sequence and biological annotation of human chromosome 1. *Nature*, Springer Science e Business Media LLC, v. 441, n. 7091, p. 315–321, mai. 2006. DOI: 10.1038/nature04727. Disponível em: https://doi.org/10.1038/nature04727.

[23] MORAES, P. L. *Características do RNA. Tipos de RNA*. [S.l.]: Mundo Educação. Disponível em: https://mundoeducacao.uol.com.br/biologia/rna.htm.

[24] CRICK, F. H. On protein synthesis. *Symp Soc Exp Biol*, v. 12, p. 138–163, 1958.

[25] CRICK, F. Central Dogma of Molecular Biology. *Nature*, Springer Science e Business Media LLC, v. 227, n. 5258, p. 561–563, ago. 1970. DOI: 10.1038/227561a0. Disponível em: https://doi.org/10.1038/227561a0.

[26] VOGEL, J.; WAGNER, E. G. H. Target identification of small noncoding RNAs in bacteria. *Current Opinion in Microbiology*, Elsevier BV, v. 10, n. 3, p. 262–270, jun. 2007. DOI: 10.1016/j.mib.2007.06.001. Disponível em: https://doi.org/10.1016/j.mib.2007.06.001.

[27] BARABASI, A.-L. *Linked: How Everything Is Connected to Everything Else and What It Means*. [S.l.]: Plume, 2003. ISBN 0452284392.

[28] NETTO, P. O. B. *Grafos: Teoria, Modelos, Algoritmos*. 5. ed. [S.l.]: Bluncher, 2011.

[29] BOCCALETTI, S. et al. Complex networks: Structure and dynamics. *Physics Reports*, v. 424, n. 4, p. 175–308, 2006. ISSN 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2005.10.009. Disponível em: http://www.sciencedirect.com/science/article/pii/S037015730500462X.

[30] ARAÚJO, D.; BASTOS-FILHO, C.; MARTINS FILHO, J. Métricas de Redes Complexas para Análise de Redes Opticas. *Revista de Tecnologia da Informação e Comunicação*, v. 2, p. 1–8, jan. 2012. DOI: 10.12721/2237-5112.v02n01a01.

[31] PANWAR, B.; ARORA, A.; RAGHAVA, G. P. Prediction and classification of ncRNAs using structural information. *BMC Genomics*, v. 15, n. 1, p. 127, fev. 2014. ISSN 1471-2164. DOI: 10.1186/1471-2164-15-127. Disponível em: https://doi.org/10.1186/1471-2164-15-127.

[32] REGINA MARTELETO, M. T. e. Redes sociais: posições dos atores no fluxo da informação. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, v. 11, n. 1, p. 75–91, 2006. ISSN 1518-2924. DOI: 10.5007/1518-2924.2006v11nesp1p75. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2006v11nesp1p75.

[33] SHANNON, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, v. 27, n. 3, p. 379–423, 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[34] JAYNES, E. T. Information Theory and Statistical Mechanics. *Phys. Rev.*, American Physical Society, v. 106, p. 620–630, 4 mai. 1957. DOI: 10.1103/PhysRev.106.620. Disponível em: https://link.aps.org/doi/10.1103/PhysRev.106.620.

[35] LOPES, F. M. UM MODELO PERCEPTIVO DE LIMIARIZAÇÃO DE IMAGENS DIGITAIS. *Universidade Estadual de Maringá*, 2003.

[36] REIMERS, M.; CAREY, V. J. [8] Bioconductor: An Open Source Framework for Bioinformatics and Computational Biology. In: DNA Microarrays, Part B: Databases and Statistics. [S.l.]: Academic Press, 2006. v. 411. [Methods in Enzymology]. P. 119–134. DOI: https://doi.org/10.1016/S0076-6879(06)11008-3. Disponível em: http://www.sciencedirect.com/science/article/pii/S0076687906110083.

[37] EVANS, J. S.; MURPHY, M. A. *rfUtilities*. [S.l.], 2018. R package version 2.1-3. Disponível em: https://cran.r-project.org/package=rfUtilities.

Fabrício Martins Lopes

Murilo Montanini Breve