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

Relatório Final de Atividades

Reconhecimento de padrões em sequências de mRNA e lncRNA: um estudo de caso utilizando redes complexas

vinculado ao projeto

Reconhecimento de padrões em sequências genômicas

Eric Augusto Ito

Bolsista CNPq

Engenharia de Computação

Data de ingresso no programa: 08/2016

Prof. Dr. Fabrício Martins Lopes

Área do Conhecimento: 1.03.00.00-7 - ciência da computação

CAMPUS CORNÉLIO PROCÓPIO, 2017

ERIC AUGUSTO ITO
FABRÍCIO MARTINS LOPES

RECONHECIMENTO DE PADRÕES EM SEQUÊNCIAS DE MRNA E LNCRNA: UM ESTUDO DE CASO UTILIZANDO REDES COMPLEXAS

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

CAMPUS CORNÉLIO PROCÓPIO, 2017

SUMÁRIO

INTRODUÇÃO

Com o passar dos anos a biologia realizou grandes descobertas na área do genoma humano, visto que houve a necessidade de entender melhor como funcionam as sequências biológicas, para assim entender melhor o funcionamento de organismo vivos. Em consequências a este avanços muitas coisas mudaram, como a acessibilidade a tais sequências, que ocasionou uma explosão de dados a serem analisadas. Por tanto um dos problemas que a biologia enfrenta é a análise de enormes conjuntos de dados, que precisam ser analisados com precisão e na maioria da vezes com velocidade também. Por este motivo há vários trabalhos que visam a classificação de dados, que é o caso do artigo[1].

Umas das vertentes da bioinfomática é a criação de métodos para a análise de dados, como dito anteriormente há enormes conjuntos de dados para serem analisados os quais só é possível com o uso de computadores e programas específicos para o tratamento desses dados. Este trabalho tem como objetivo abordar uma metodologia para a classificação de mRNA e lncRNA utilizando a teoria de redes complexas.

O RNA longo não-codificante também chamado de LncRNA, são tipo de RNAs os quais ainda não há uma definição exata que caracteriza este RNA, por muito tempo achou que este RNA não era importante para o entendimento da biologia, sendo chamada até de RNA lixo. Mas estudos recentes mostram que o LncRNA tem funções importantes para o funcionamento do organismos na formação de proteínas[2].

As redes estão presentes em vários lugares, desde a internet à formação de textos. E como se sabe as sequências de DNA e RNA são formadas por bases nitrogenadas que por sua vez formam os códons com o objetivo de formar proteínas. Como as proteínas são as responsáveis pelo funcionamento de organismos vivos, logo o entendimento das proteínas podem trazer muitos esclarecimentos sobre doenças por exemplo. Tendo como foco os códons, a partir das sequências deles serão formadas redes, com isso o objetivo deste trabalho é abstrair características dessas redes para então fazer a classificação entre as sequências de RNA mensageiro e RNA longo não codificante.

Em suma a idéia gerar redes a partir de sequências de RNA, após isso abstrair possíveis características da rede como número médio de conexões, e assim classificar as sequências baseados nas características da rede.

REVISÃO BIBLIOGRÁFICA

Redes complexas. O matemática Leonhard Euler em 1736 deu início a teoria dos grafos ao resolver o problema de Königsberg, após isso houveram vários estudos sobre redes complexas e como elas são formadas, dando destaque ao Erdös e Rényi que originaram a teoria de redes aleatórias as quais são redes que como o nome diz são formadas de forma aleatória como um jogo de dados. Mas há outros tipos de redes que descrevem melhor as redes formadas nesse trabalho, que é o caso das redes livre de escala do matemático Albert-László Barabási, que descreve as redes complexas sendo grafos com topologias não triviais, que possuem um conjunto de vértices(nós) que são ligados através de arestas[3]. Na Figura 1 pode ser visto um exemplo de rede complexa.

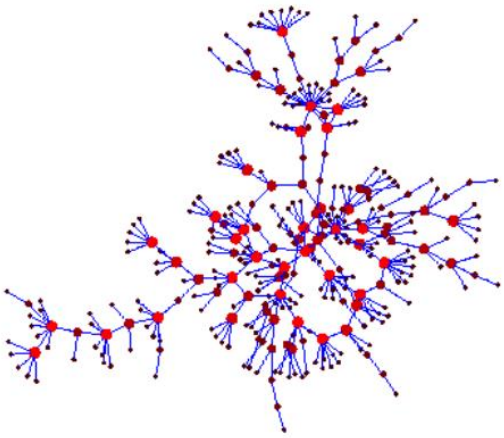

Figura 1. Exemplo de rede complexa[4]

Sequências biológicas. Com os estudos sobre a molécula ácido desoxirribonucleico(DNA), foi aberto um leque de conhecimentos sobre o funcionamento biológico de organismos vivos, o DNA é responsável pelas instruções para a produção de proteínas de um organismo[5]. Toda a informação que gera um organismo encontra-se em uma sequência linear de quatro bases, sendo elas, Timina (T), Adenina (A), Citosina (C) e Guanina (G), no RNA o Timina da lugar a Uracila (U). O RNA é sintetizado por transcrição a partir da molécula de DNA, e a partir da sequência de RNA é feito a tradução do códons, que são sequência de três bases e cada um desses códons é responsável pela formação de uma proteína, como pode ser visto na Figura 2. Cabe a célula, armazenar, obter e traduzir as instruções genéticas para que o organismo se mantenha vivo[5][6].

Com o avanços dos estudos sobre o RNA foi encontrado um RNA não-codificante, que se caracteriza por ter mais que 200 nucleotídeos de comprimento e não codificam proteínas. Este tipo de RNA possui duas classes principais, sendo elas, Pequeno RNA não-codificante (SncRNA) e Longo RNA não-codificante (LncRNA). Atualmente não há um definição clara que possa caracterizar lncRNA e poucas moléculas de LncRNAs foram documentadas, mas segundo[2] os estudos atuais sugerem que os LncRNAs podem estar relacionados a quase todos as fases da regulação de expressão gênica.

MATÉRIAS E MÉTODOS

Matérias utilizados. Para a realização dos testes aqui apresentados foi utilizado um desktop com um processador Intel i7 4790k (Stock), Memórias HyperX 4x4gb 1600MHz, SSD 120GB Samsung Evo 840.

Software R. Ferramenta criada por Ross Ihaka e Robert Gentleman, muito utilizada para cálculos estatísticos, o software R também suporta o uso de pacotes criados por desenvolvedores independentes que acrescentam novas funções ao Software R, alguns desses pacotes são os Igraph e o Rmcfs, o primeiro tem como finalidade o trabalho com grafos, possuindo não só funções para criação de grafos mas assim como funções de representação e medidas que são fundamentais para este trabalho. Já o pacote Rmcfs será

utilizado para gerar arquivos no formato .arff com o valores abstraídos das redes complexas. Por fim, o ultimo pacote utilizado foi o RGL para a criação de gráfico tridimensionais.

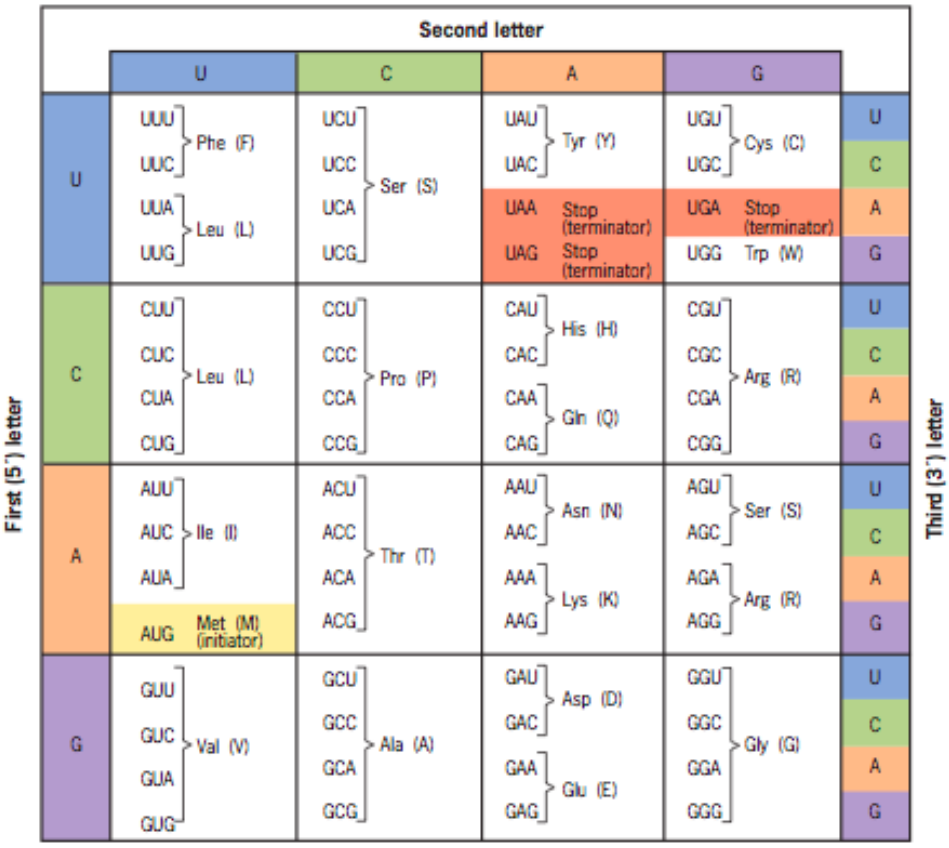

Figura 2. Tradução dos Códons[7].

WEKA. O WEKA é um ferramenta criada pela universidade de Waikato com vários recursos implementados para a classificação de dados, muito utilizado em vários trabalhos acadêmicos pela credibilidade de seus resultados, apresentar uma interface e possuir a licença GNU General Public License, que o torna um ótimo software para trabalhos. Entre suas inúmeras funções duas serão utilizadas neste trabalho, a primeira são os classificadores Random Forest, NaiveBayes e J48 que já estão implementados no WEKA, e outra função que será utilizada neste trabalho é a seleção de recursos, esta função tem como objetivo fazer a seleção de atributos que tem a maior porcentagem de contribuição para a classificação dos resultados, dessa forma diminuindo atributos que acabam atrapalhando na hora da classificação

A primeira etapa para a aplicação da metodologia escolhida foi o estudo sobre redes e como fazer a criação delas, dessa forma foi realizado a criação de redes complexas a partir de algumas sequências aleatórias. Para a criação destas redes foi feito uma aplicação em JAVA que tem 2 parâmetros como entrada, o primeiro é o tamanho da palavra e o segundo o passo. Como já dito redes são formadas por arestas e vértices, dada um sequências ACCTGACA se o tamanho da palavra for 1, então cada vértice será constituído de uma letra, como pode ser visto na Figura 3, se o tamanho da palavra for 2, logo cada vértice será constituído de duas letras, como na Figura 4.

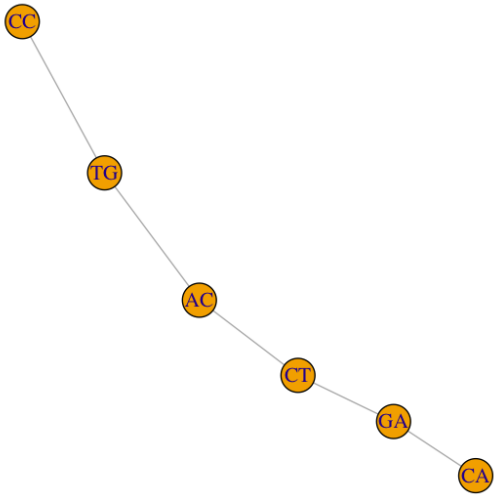

Figura 3. Rede com vértices formados com palavras de tamanho 1.

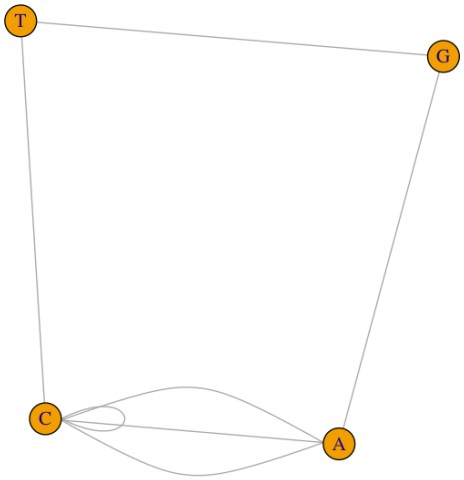

Figura 4. Rede com vértices formados com palavras de tamanho 2.

O parâmetro de passo defini a posição da janela das ligações entre os vértices, por exemplo, se o tamanho da palavra for igual a 2, teremos duas janelas de tamanho 2, de forma que a ligação entre essas duas janelas formara uma aresta entre dois vértices. O passo defini a distância que essas duas janelas se deslocam pela sequência. Na Figura 5 é possível ver como é formada as ligações quando o passo é igual a 1, a janela azul será a primeira ligação a ser formada, com a ligação entre os vértices AC e TG, a segunda ligação será formada com o deslocamento da janela em uma posição para a direita, desse modo a próxima ligação seria CT com GC. Já na Figura 6 a ligação foi formada com parâmetros de passo igual a 2, o que diferencia essa formação de rede é que para a segunda ligação há um deslocamento de duas posições para a direita, dessa forma a segunda ligação não seria mais CT e CG, e sim, TG e CT.

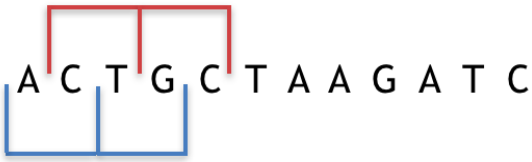

Figura 5. Formação da rede com passo igual a 1.

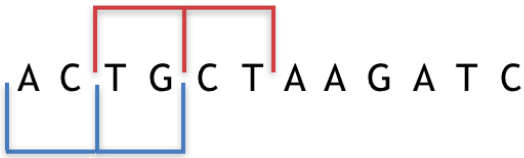

Figura 6. Formação da rede com passo igual a 2.

Ainda na aplicação JAVA foi feito um método para a leitura de arquivos no formato FASTA, que é comum ser encontrado na bioinformática e é utilizado para representar sequências de nucleotídeos. Como pode ser visto abaixo, há 4 linhas de arquivo no formato FASTA, linhas q se referem a duas sequências da espécie *Mus musculus*. Em arquivos FASTA cada sequência possui duas linhas, na primeira há uma descrição sobre a sequência e a segunda linha se trata da própria sequência de nucleotídeos.

>ENSMUST00000104738 ncRNA#: miRNA# chromosome: GRCm38: 17:35960730: 35960814: -1 gene: ENSMUSG00000077931 gene_biotype: miRNA# transcript_biotype: miRNA#
GTAGAGGAGATGGCGCAGGGGACACAAGGTAGGCCTTGCGGGTCTGTGGA
CCCTTGGACATGTGTCCTCTTCTCCCTCCTCCCAG
>ENSMUST00000175080 ncRNA#: miRNA# chromosome:GRCm38: 8: 109425644: 109425744: -1 gene: ENSMUSG00000092821 gene_biotype: miRNA# transcript_biotype: miRNA#
ATATATATATATATATATATATATGTGTGTGTGTGTGTGTGTGTGTGTATA
TATATATACACATACATACACACACTCATAGAATACATGCATATACAC

Dessa forma foi gerado várias redes as quais agora não eram mais aleatórias mas sim de um grupo específico de espécies, no total foram usadas 9 espécies diferentes, como mostra a tabela 1.

Estes conjunto de dados foram retirados do artigo[8], o qual usa este mesmo conjunto de dados para fazer a classificação utilizando duas ferramentas, PLEK e o CNCI.

Até então foi gerado grafos sem peso, os quais são grafos que não diferem nós com mais ou menos ligações, tal fato é um problema já que para compreender a estrutura da rede é necessário saber como cada vértice e aresta da rede se comporta, dessa forma foi necessário começar a trabalhar com grafos com peso, na Figura 7 é possível ver um grafo originado da sequência ACCTGACACACACACACA, o qual possui várias ligações entre os vértice A e C.

Tabela 1: Conjunto de dados

Espécies	Tipo de RNA	Número de nucleotídeos
Mus musculus	Mensageiro	26062
	Não-codificante	2963
Danio rerio	Mensageiro	14493
	Não-codificante	419
Xenopus tropicalis	Mensageiro	8874
	Não-codificante	279
Bos taurus	Mensageiro	13190
	Não-codificante	182
Pan troglodytes	Mensageiro	1906
	Não-codificante	1166
Sus scrofa	Mensageiro	3978
	Não-codificante	241
Macaca mulatta	Mensageiro	5709
	Não-codificante	359
Gorilla gorilla	Mensageiro	33025
	Não-codificante	367
Pongo abelii	Mensageiro	3401
	Não-codificante	392

Como pode ser visto no grafo da Figura 7, arestas mais grossas representam ligações que ocorrem mais vezes. E a partir desses grafos com peso também foi possível gerar matrizes que representam as ligações do grafo, como a tabela 2.

Tabela 2: representação das ligações feitas pelo grafo no formato de matriz

	A	C	T	G
A	.	13	.	1
C	13	1	1	.
T	.	1	.	1
G	1	.	1	.

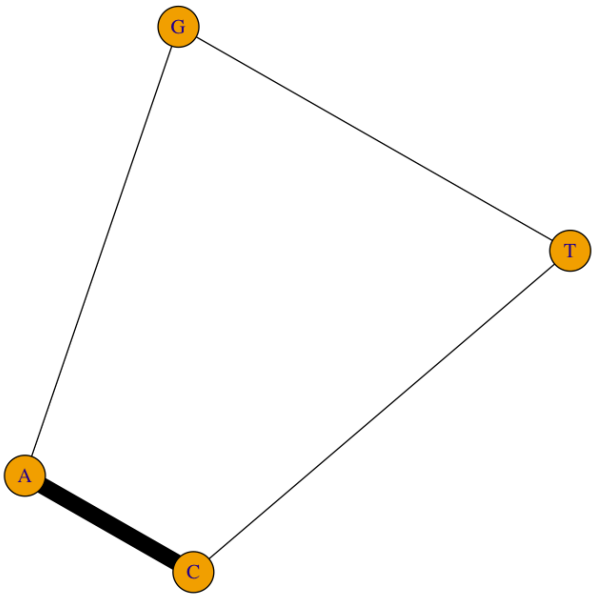

Figura 7. Exemplo de grafo com peso.

Para cada rede então formada pelos conjuntos de dados da tabela 1, foi abstraído alguma medidas, pelas quais caracterizam a estrutura da rede. As medidas selecionadas foram 9, todas listadas abaixo:

Desvio Padrão: Cada nó da sequência possui um número n de ligações que varia de sequência para sequência, tal característica pode ser determinante para a classificação de sequências. Sabendo disso será calculado o desvio padrão do número de ligações feitas por cada nó, valores altos de desvio padrão mostram que a sequência possui números de ligações não balanceadas, ou seja, há nós que se conectam mais que outros.

Mínimo: Em cada sequência há nós os quais fazem as suas ligações com outros nós, esta medida retorna o menor número de conexões feitas por um nó em uma sequência.

Máximo: Ao contrário da medida anterior agora será retornado o maior número de conexões feitas por um nó dado um sequência.

Degree: Em uma rede o número de conexões que os vértices fazem podem determinar comportamentos específicos de uma rede. Para isso será usado uma medida chamada de degree na qual retorna a média de conexões que os vértices da rede possuem[9].

Coeficiente de Cluster: Também chamada de transitividade, esta medida calcula a probabilidade de vértices que estão conectados a um outro vértice estarem também conectados entre si. 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. A transitividade também pode ser vista como a formação de triângulos nas redes[9]. A transitividade pode ser obtida através da seguinte equação:

Assortatividade: Tem como resultado 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 significam o

contrário[8]. Em uma sequência biológica na qual um certo tipo de padrão é recorrente, eles tendem a se agrupar, isto é, estarem todos associados.

Intermediação: A intermediação se tornou uma estratégia popular para lidar com redes complexas. As aplicações incluem redes sociais, redes computadores, redes biológicas, redes de transporte, entre outras[11]. Um vértice qualquer mesmo ele tendo poucas ligações pode ter grande importância na formação da rede, pois o vértice pode estar posicionado nos caminhos geodésicos entre outros pares de vértices na rede. Sabendo disso a intermediação tem como objetivo fazer a medição da centralidade de cada nó em uma rede[12].

Caminho Mínimo Médio: Em uma rede o caminho mínimo desempenha um papel importante no transporte e na comunicação. É por isso que na internet ou em qualquer outra rede na qual a solução geodésica representa uma solução ótima para um problema, o caminho mínimo desempenha um papel importante na caracterização da estrutura interna de um gráfico. Dado uma rede qualquer o caminho mínimo seria uma matriz M na qual dado uma linha e coluna se encontraria a distância mínima entre dois vértices. O caminho mínimo médio é a média de todos os caminhos mínimos de uma rede[9]. Para redes muito concentradas, o valor para o caminho mínimo médio é baixo, com isso pode se concluir que o grau das ligações são semelhantes.

Motivo: Também chamado de *motifs* em inglês, motivos são sub-redes com vários formatos que há dentro de uma rede. Nessa medida será feita a contagem de quantos *motifs* do tipo 3 e 4 ocorrem dentro de uma rede[13].

Para melhor entender a estrutura da rede e seu comportamento, foi tomada a estratégia de aplicar *Thresholds* na rede, que são cortes que acontecem sucessivamente na rede em 1 em 1 nas arestas mais fracas, com isso a cada *threshold* aplicado as arestas mais fracas presentes são descartadas, na Figura 8 é possível ver a rede original composta por todos as arestas, na Figura 9 foi executado um *threshold* nas arestas de peso 1, ou seja, as arestas de peso 1 foram apagadas, na Figura 10 foi executada um *threshold* nas arestas de peso 2. A cada *threshold* é feita a aplicação de todas as medidas acima listadas. Os *thresholds* acontecem até que não hajam mais arestas na rede.

Após a extração de todas as medidas para cada *threshold*, os resultados foram alocados em uma matriz, de forma que cada medida corresponde a uma matriz, totalizando 10 matrizes criadas. Dessa maneira cada matriz é constituída de Sequências por *Thresholds*, ficando como mostra a tabela 3.

Tabela 3: Exemplo da matriz formada para cada medida

Sequências	***Threshold* 0**	***Threshold* 1**	***Threshold* 2**
1	0,89	1,56	1,90
2	1,20	1,23	2,12
3	0,56	1,44	1,98
4	0,76	1,65	1,67

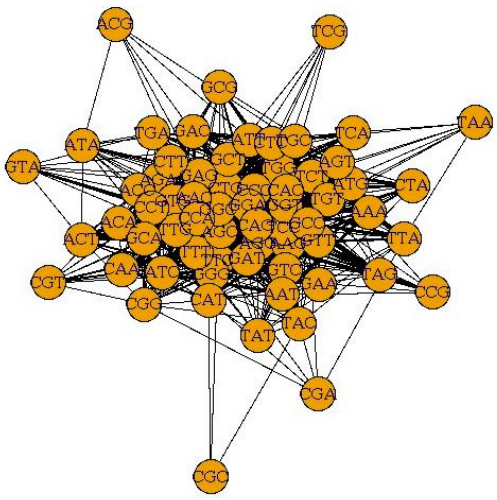

Figura 8: Rede original sem aplicação de *threshold*.

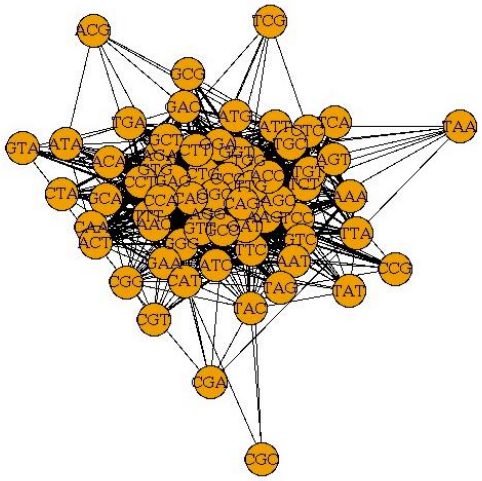

Figura 9: Rede com aplicação de *threshold* nas arestas de peso 1.

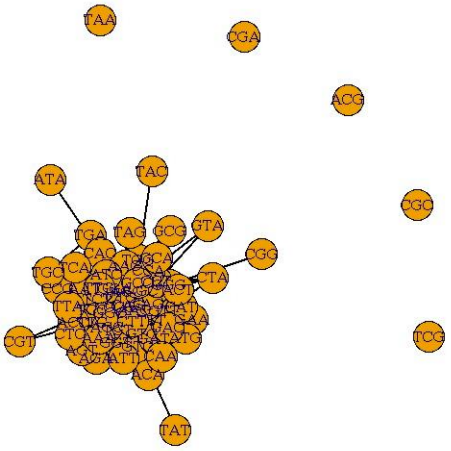

Figura 10: Rede com aplicação de *threshold* nas arestas de peso 2.

Para averiguar a variância dos tamanhos das sequências, foi criada uma outra aplicação em JAVA que também faz a leitura de arquivo do tipo FASTA, mas dessa vez essa aplicação tem como objetivo contar o tamanho de cada sequência e gerar um script para o software R, esse script tem como papel criar um boxplot para cada tipo de RNA. Na Figura 11 pode ser vista um exemplo de boxplot para o conjunto Pongo Abelli.

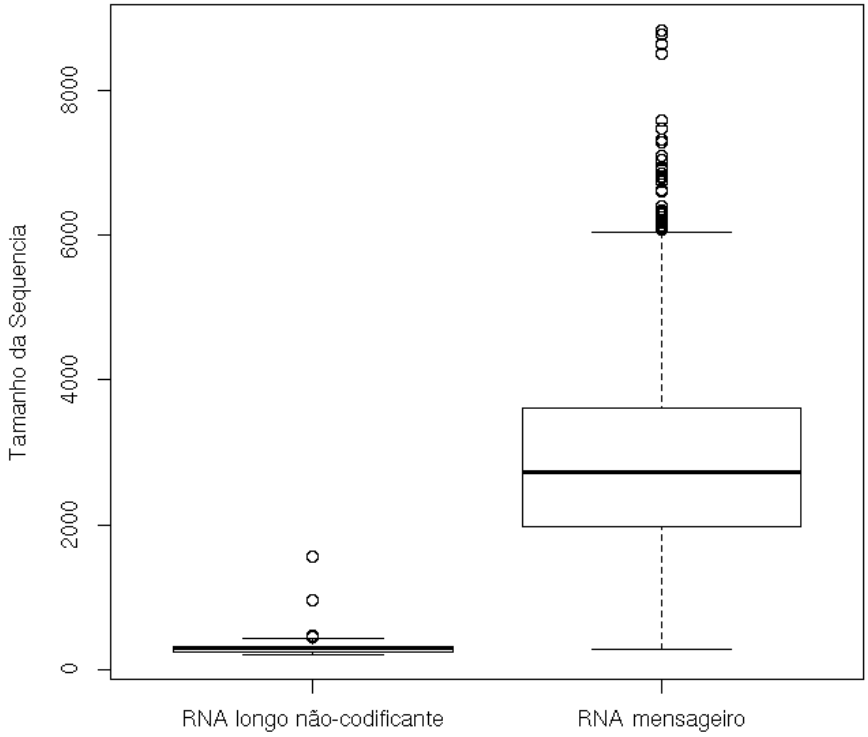

Figura 11. Exemplo de boxplot do conjunto Pongo Abelli.

Como pode ser visto na Figura 11 há uma grande variância de tamanhos nas sequências, tal fato pode ter influência direta nos resultados, e como não é interessante que o tamanho das medidas influencie nos resultados, foi aplicada um ajuste nas medidas para alterar a escala dos resultados conforme o tamanho da sequência. Dessa maneira cada resultado abstraído da rede por uma medida foi dividido pelo tamanho da sequência analisada.

Para cada matriz que corresponde as medidas analisadas, foi criado um gráfico(*Threshold* x Medida), em busca de alguma trajetória que possa ser uma assinatura para identificação do genoma, dessa forma foram gerados 10 gráficos, um para cada medida aqui analisada. A Figura 12 mostra um exemplo de gráfico para o caminho mínimo médio do conjunto Pongo Abelli.

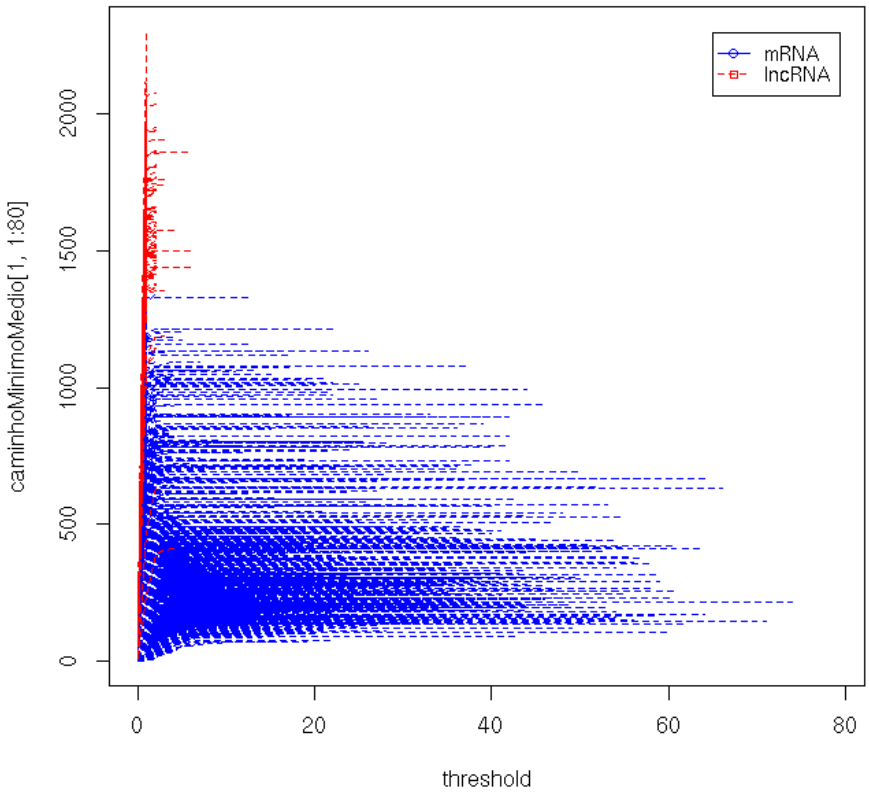

Figura 12. Gráfico *Threshold* X Caminho mínimo médio.

Uma outra abordagem feita neste trabalho, foi gerar gráficos tridimensionais, visando melhorar a discrepância entre os resultados de mRNA e lncRNA, fazendo uso de duas medidas simultaneamente, com isso os gráficos tridimensionais ficaram como mostra a Figura 13. No total foram gerados 100 gráficos distintos com todas as combinações possíveis com as 10 medidas analisadas neste trabalho.

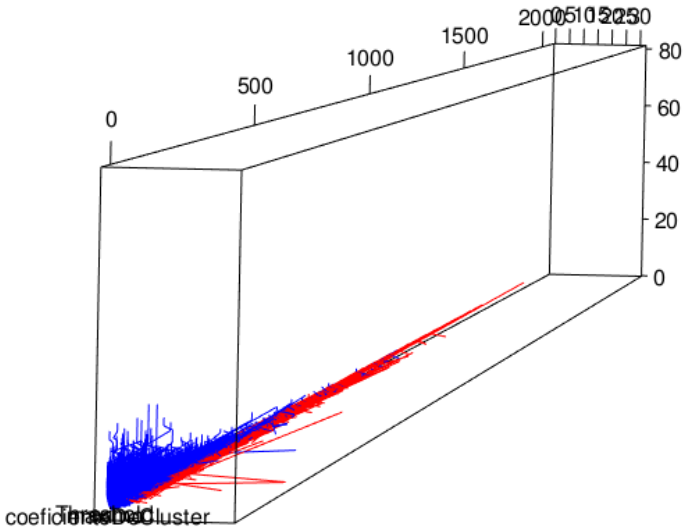

Figura 13. Exemplo de gráfico tridimensional.

Com todos as redes analisadas e resultados obtidos, foi gerado um arquivo do tipo arff pelo software R através do pacote RMCFS. Este arquivo foi dado como entrado no WEKA e posteriormente colocado para classificação pelos classificadores: Random Forest, J48 e Naive Bayes.

RESULTADOS E DISCUSSÕES

Todos os resultados apresentados neste trabalho foram realizados utilizando como parâmetros passo igual 3 e tamanho de palavra também igual a 3, está escolha se deve pela existência dos códons que como já dito correspondem a sequência de 3 bases que serão traduzidas para proteínas específicas.

Com a geração dos gráficos bidimensionais deu para perceber que a metodologia de classificar as redes a partir de dados gerados sobre a estrutura da rede era possível, já que visualmente deu para ver que o mRNA e o lncRNA se destoam, como pode ser visto na Figura 14, o gráfico da Figura 14 foi feito com base na espécie Pongo Abelli com o passo configurado para 3 e o tamanho da palavra sendo 3.

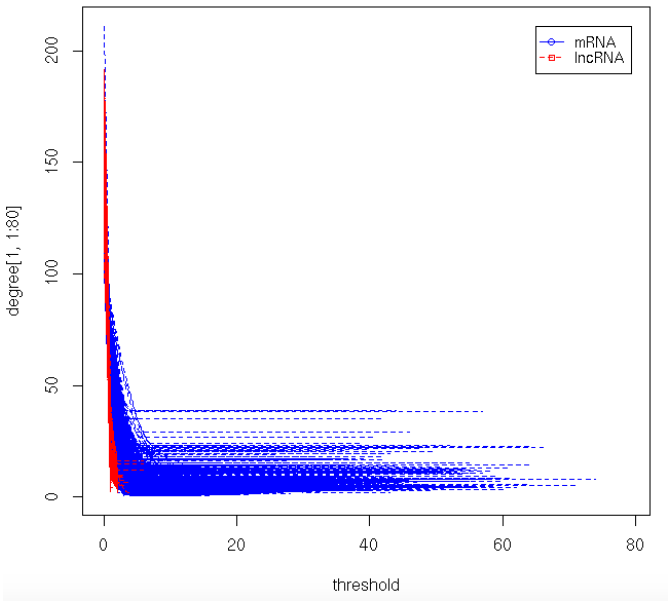

Figura 14: Gráfico degree x *threshold*.

Os gráficos tridimensionais não mostraram nenhuma informação relevante que possa mostrar alguma trajetória específica para os tipos de RNA testados neste trabalho. Na Figura 15 fica bem claro que não foi possível encontrar nenhum trajetória especifica independente da posição que o gráfico é posto.

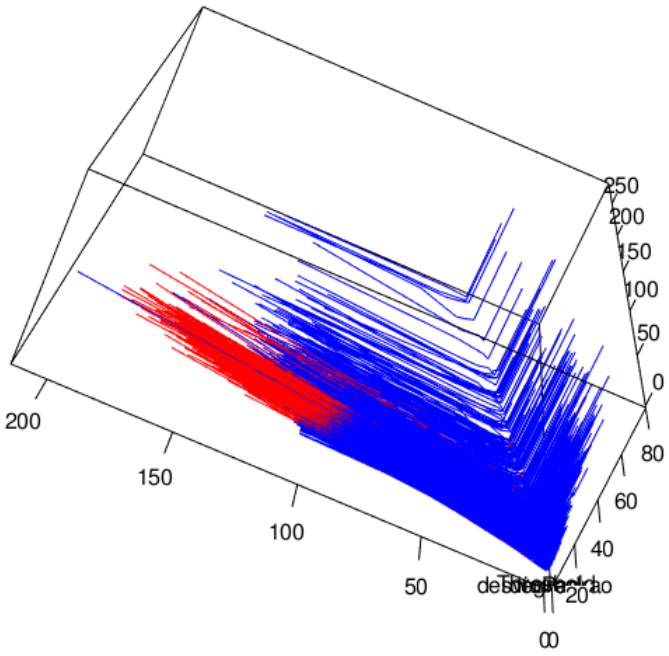

Figura 15. Gráfico tridimensional(*threshold* x desvio padrão x degree).

A classificação utilizando os classificadores Random Forest, Naive Bayes e J48 para todas as 9 espécies da tabela 1 podem ser vistos na tabela 4.

Para a classificação dos RNA mensageiros o Random Forest teve melhores resultados em 7 espécies mas se considerado a média geral o J48 conseguiu superar o Random Forest por 0,04%. Já para a classificação de lncRNA o Naive Bayes foi o melhor conseguindo na média geral ser 11,69% superior ao Random Forest que foi o segundo melhor. Na média total o classificador Naive Bayes obteve a maior porcentagens de acertos mesmo ele não sendo o melhor na classificação de mRNA.

Do artigo[8] foram retirados os resultados da classificação pelos classificadores CNCI e PLEK, como já mencionado o conjunto de dados testados neste trabalho foi a mesma utilizada no artigo[8], dessa forma foi feita uma comparação entre os resultados, que podem ser vistos também na tabela 4.

Se tratando da classificação de mRNA o método deste trabalho superou tanto o CNCI quanto o PLEK por uma boa margem de acertos. O J48 conseguiu ser superior ao CNCI por 6,9% e 9,69% ao PLEK. Na classificação de lncRNA o CNCI conseguiu melhores resultados, sendo 4,79% superior. Na média geral tanto o CNCI e o PLEK não conseguiram bater o número de acertos realizados pelo classificação Naive Bayes, obtendo resultados 3,45% superiores ao obtidos pelo PLEK e 1% superior ao CNCI.

Tabela 4: porcentagem de acertos na classificação.

Espécies	Tipo de RNA	Random Forest	Naive Bayes	J48	CNCI	PLEK
Mus musculus	mRNA	98,20	75,60	98,20	93,90	88,10
	lncRNA	38,10	69,70	38,10	97,10	89,90
Danio renio	mRNA	99,80	87,10	99,90	95,30	91,30
	lncRNA	44,60	89,70	40,30	89,30	90,90
Xenopus tropicalis	mRNA	100,00	98,80	100,00	92,90	94,50
Bos taurus	lncRNA	97,80	99,30	97,10	99,70	100,00
	mRNA	99,90	97,80	99,80	94,30	94,80
	lncRNA	92,30	98,90	90,10	100,00	99,50
Pan troglodytes	mRNA	98,00	99,80	98,60	90,20	87,10
	lncRNA	99,10	95,20	98,00	100,00	99,90
Sus scrofa	mRNA	99,50	87,60	99,50	93,40	85,10
	lncRNA	78,40	92,70	73,40	95,90	98,30
Macaca mulatta	mRNA	99,40	99,40	99,30	92,00	85,00
	lncRNA	95,30	95,70	95,00	99,70	100,00
Gorilla gorilla	mRNA	99,80	97,20	99,70	87,40	83,80
	lncRNA	88,30	99,50	84,20	99,70	99,70
Pongo abelii	mRNA	99,90	99,50	99,90	93,40	98,00
	lncRNA	99,00	97,40	98,50	99,80	100,00
Média: Mensageiro		99,39	93,64	99,43	92,53	89,74
Média: Não -codificante		81,43	93,12	79,41	97,91	97,58
Média total		90,41	93,38	89,42	92,38	89,93

CONCLUSÕES

O trabalho aqui proposta teve como meta desenvolver um método para classificar sequências de mRNA e lncRNA, mas que possivelmente pode ser utilizadas em outros tipos de sequência como a de DNA.

Neste trabalho foram realizados vários testes com diversas sequências de 9 espécies distintas, como resultado tanto o PLEK como o CNCI tiveram resultados inferiores na classificação de RNA mensageiro e também na média geral de acertos.

Com esses resultados fica claro a adequação da metodologia aqui demonstrada, com parâmetros de passo igual a 3 e tamanho da palavra também igual a 3, essa escolha tem ligação com os códons que possuem tamanho de 3 bases.

Assim, espera-se que o desenvolvimento deste projeto, bem como seus resultados sejam de grande valia para avançar esta área de pesquisa, pois pode auxiliar a Bioinformática no estudo sobre o entendimento sobre as sequências biológicas, que são primordiais para o funcionamento de organismos.

REFERÊNCIAS

[1] CONQUE, Bruno MM; KASHIWABARA, Andre Yoshiaki; LOPES, Fabricio Martins. A feature extraction approach based on complex networks for genomic sequences recognition. In: Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), International Congress on. IEEE, 2016. p. 1803-1807.

[2] DOBBLER, Priscila. Rna longo não-codificante: mecanismos, características e funcionalidades do dna "lixo". Trabalho de Conclusão de Curso(Graduação)-Universidade Federal do Pampa, 2015.

[3] BARABÁSI, Albert-László. Linked: The new science of networks. 2003.

[4] METZ, Jean et al. Redes Complexas: conceitos e aplicações. Relatórios Técnicos do ICMC-USP São Carlos, 2007.

[5] ALBERTS, Bruce et al. Biologia molecular da célula. Artmed Editora, 2010.

[6] PASSAGLIA, L. M. P.; ZAHA, Arnaldo. Biologia molecular básica. Porto Alegre: Mercado Aberto, 2003.

[7] SNUSTAD, D. Peter. Principles of Genetics 6E Binder Ready Version with WileyPlus. John Wiley & Sons, 2012.

[8] LI, Aimin; ZHANG, Junying; ZHOU, Zhongyin. 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, 2014.

[9] DE ARAÚJO, Danilo RB; BASTOS-FILHO, Carmelo JA; MARTINS-FILHO, Joaquim F. Métricas de Redes Complexas para Análise de RedesOpticas. 2014.

[10] BOCCALETTI, Stefano et al. Complex networks: Structure and dynamics. Physics reports, v. 424, n. 4, p. 175-308, 2006.

[11] BALESTRIN, Alsones; VERSCHOORE, Jorge Renato; REYES JUNIOR, Edgar. O campo de estudo sobre redes de cooperação interorganizacional no Brasil. RAC-Revista de Administração Contemporânea, v. 14, n. 3, 2010.

[12] TOMAÉL, Maria Inês; MARTELETO, Regina Maria. Redes sociais: posições dos atores no fluxo da informação. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, n. Especial 1, 2006.

[13] MILO, Ron et al. Network motifs: simple building blocks of complex networks. Science, v. 298, n. 5594, p. 824-827, 2002.