UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

BRUNO MENDES MORO CONQUE

EXTRAÇÃO DE CARACTERÍSTICAS A PARTIR DE REDES COMPLEXAS: UM ESTUDO DE CASO NA CLASSIFICAÇÃO DE SEQUÊNCIAS GENÔMICAS

TRABALHO DE CONCLUSÃO DE CURSO

CORNÉLIO PROCÓPIO

2014

BRUNO MENDES MORO CONQUE

EXTRAÇÃO DE CARACTERÍSTICAS A PARTIR DE REDES COMPLEXAS: UM ESTUDO DE CASO NA CLASSIFICAÇÃO DE SEQUÊNCIAS GENÔMICAS

Trabalho de Conclusão de Curso apresentado ao Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do título de Tecnólogo.

Orientador: Fabrício Martins Lopes

Co-orientador: André Yoshiaki Kashiwabara

CORNÉLIO PROCÓPIO

2014

AGRADECIMENTOS

Em primeiro lugar agradeço a Deus, por ter me dado saúde, persistência e vontade para superar todas as barreiras que surgiram no decorrer deste trabalho e do curso.

Aos meus pais Ângela e Laertes por todo incentivo durante a minha vida frente aos estudos e pelo apoio incondicional em todas minhas decisões.

Ao meu irmão Laertes Junior por todo auxílio prestado de inúmeras formas durante todo meu caminho percorrido até aqui.

Ao meu irmão Leonardo Rhodan por sempre proporcionar distrações e risadas em momentos difíceis.

Ao professor Doutor Fabrício Martins Lopes pela oportunidade oferecida de iniciação científica e a minha orientação frente ao trabalho de diplomação, sendo esta cercada de atenção, paciência e ideias, elementos fundamentais para a conclusão deste trabalho.

Ao professor Doutor André Yoshiaki Kashiwabara por toda ajuda prestada.

A todos os professores do curso, que são contribuintes de todo conhecimento gerado e minha formação acadêmica.

A todos meus amigos por sua amizade, em especial ao Anderson Brilhador, Danillo de Oliveira, Higor Cotrim, João Antonio Dias Tonet, João Paulo de Lima, Mike Henrrique e Renan Martins por toda contribuição prestada de alguma forma durante a minha formação.

A todos vocês, meu muito obrigado!

RESUMO

MENDES, Bruno. Extração de características a partir de Redes Complexas: Um estudo de caso na classificação de Sequências Genômicas. 51 f. Trabalho de Conclusão de Curso – Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2014.

No âmbito da bioinformática, o reconhecimento de padrões dentro de sequências genômicas pode ser utilizado para classificar regiões (gênica, promotora, não-codificante) de um DNA. Neste sentido, caso uma boa classificação ocorra um modelo pode ser gerado para inferir sequências desconhecidas. Frente a essa perspectiva, medidas que representam particularidades dentro dessas sequências devem ser identificadas. Este trabalho propõe duas metodologias para caracterizar as sequências genômicas baseadas na teoria das redes complexas e teoria da informação. A teoria da informação lida com a frequência das ocorrências de nucleotídeos, dinucleotídeos e trinucleotídeos dentro de uma sequência para calcular entropia, soma de entropia e valor máximo da entropia para compor as características da mesma. As redes complexas por sua vez, retratam as sequências como uma rede através da ocorrência de encontro entre os nucleotídeos, dinucleotídeos e trinucleotídeos dentro da sequência. As medidas das metodologias são utilizadas na classificação com métodos classificadores como SVM, MultiLayerPerceptron, J48, IBK, NaiveBayes e RandomForest, para os quais foram obtidos resultados similares apresentando pouca diferença a favor das redes complexas, sendo que o RandomForest apresentou os melhores resultados com aproximadamente 86% de acurácia, seguido do J48 com 84% e do MultiLayerPerceptron com 82%. Os resultados obtidos indicam que através dessa abordagem de extração de características é possível alcançar bons níveis de classificação considerando a simplicidade dos métodos uma vez que são utilizadas somente as sequências genômicas sem nenhum outro conhecimento acerca delas.

Palavras-chave: redes complexas, sequências genômicas, extração de características, classificação, reconhecimento de padrões, bioinformática

ABSTRACT

MENDES, Bruno. Feature extraction from complex networks: A study case on classification of Genomic Sequences. 51 f. Trabalho de Conclusão de Curso – Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2014.

Within the scope of bioinformatics, pattern recognition in genomic sequences can be used to classify regions (gene, promoter, non-coding) of a DNA. In this sense, if a model a good classification occurs can be generated to infer unknown sequences. Faced with this prospect, measures that represent characteristics within these sequences must be identified. This paper proposes two methods to characterize the genomic sequences based on the theory of complex networks and information theory. Information theory deals with the frequency of occurrences of nucleotide, dinucleotide and trinucleotide within a sequence to calculate entropy, sum entropy and maximum entropy to compose the same characteristics. Complex networks in turn retrate the sequences as a network through the occurring of the nucleotides, dinucleotides and trinucleotides within the same. Measures of methodologies are used in the classification methods such as SVM classifiers, MultiLayerPerceptron, J48, IBK, and NaiveBayes RandomForest, where similar results were obtained among the methods, showing little difference in favor of the complex networks, wherein RandomForest showed the best results with approximately 86 % accuracy, followed by J48 with 84 % and MultiLayerPerceptron with 82 %. The results indicate that by such feature extraction approach can achieve good classification levels considering the simplicity of the methods used since they are only genomic sequences without any further knowledge about them.

Keywords: complex networks, genomics sequences, feature extraction, classification, pattern recognition, bioinformatics

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

DNA	Ácido Desoxirribonucleíco (*Deoxyribonucleic Acid*)
RNA	Ácido Ribonucleíco (*Ribonucleic Acid*)
WEKA	Waikato Environment for Knowledge Analysis
ARFF	*Attribute Relation File Format*
SVM	*Support Vector Machines*
DBTSS	*DataBase of Transcriptional Start Sites*
TP	Tamanho da palavra
P	Passo

SUMÁRIO

1 INTRODUÇÃO

O estudo dos sistemas biológicos e como seus componentes estão interligados é um grande desafio nos dias de hoje atraindo a atenção de pesquisadores de diversas áreas do conhecimento. Esta área de pesquisa cientifica é conhecida como Systems biology (biologia de sistemas), a qual é altamente interdisciplinar sendo o seu foco principal analisar o organismo em uma forma holística.

Nesse contexto, um dos cenários que se encontra fortemente envolvido é o da bioinformática. Dividida em diferentes áreas de pesquisa, tem sido explorada e buscado apoio computacional com aplicações para diferentes finalidades, tais como identificação de genes, inibidores de enzimas, organizar e relacionar informação biológica, prever a configuração de proteínas, simular células, agrupar proteínas, montar árvores filogenéticas, realizar análise de expressão gênica, entre outras (PROSDOCIMI et al., 2012).

Uma teoria muito utilizada para representar sistemas interligados são as redes complexas. Por ter caráter multidisciplinar, é utilizada em diversas áreas de pesquisas como no caso da biologia, onde é utilizada na modelagem de interações entre os componentes celulares em estudos das relações entre genes (LOPES et al., 2014; LOPES; JR; COSTA, 2011; SHEN-ORR et al., 2002; DIAMBRA; COSTA, 2005), proteínas (COSTA; RODRIGUES; TRAVIESO, 2006; JEONG et al., 2001), e relações metabólicas (JEONG et al., 2000).

Na maioria das vezes, o estudo em grupo de componentes de um sistema interligado é mais importante do que a análise individual deles, o que favorece a aplicação das redes complexas na biologia, como sugerido pelos autores em (VOGELSTEIN; LANE; LEVINE, 2000) onde realizar o estudo das conexões do gene p53 que é um supressor de tumores, é mais importante do que estuda-lo individualmente.

Através das redes complexas é possível extrair medidas que representem características em sistemas naturais e artificiais compostos de elementos que interagem. No entanto, apesar do grande sucesso obtido pela teoria de redes complexas, ainda há um enorme campo de pesquisa a ser explorado devido a limitações tais como a falta de medidas e métodos para analisar, carac-

terizar e classificar reais redes. Portanto, para obter uma caracterização mais precisa, é essencial considerar um vasto conjunto de medições não redundantes, o que pode ser alcançado com a utilização de técnicas de reconhecimento padrões e mineração de dados (COSTA et al., 2007).

No contexto de reconhecimento de padrões, a extração de características é uma forma de redução de dimensionalidade (THEODORIDIS; KOUTROUMBAS, 2008). Mais especificamente, é um ponto de vista reducionista que utiliza do método de extração de características para tentar representar uma amostra com um subconjunto de suas medidas, minimizando a perda de informações a partir desta amostra.

Por outro lado, a teoria da informação oferece um arcabouço importante, no qual existem medidas que geralmente são utilizadas para indicar a quantidade de informações de uma determinada fonte, como mostrado em aplicações de sucesso em uma miríade de problemas da bioinformática a fim de caracterizar as relações entre genes (BUTTE; KOHANE, 2000; LOPES; MARTINS; CESAR, 2008; LOPES; OLIVEIRA; CESAR, 2011; LOPES et al., 2014).

Frente a essas perspectivas, esse trabalho tem por objetivo explorar uma abordagem inédita para a classificação de sequências genômica por meio de medidas de redes complexas e teoria da informação. Um extrator de características será desenvolvido e aplicado em um determinado conjunto de sequências genômicas, afim de avaliar a eficácia da classificação e caracterização realizada pelas medidas extraídas em regiões encontradas no DNA.

1.1 OBJETIVOS

1.1.1 OBJETIVO GERAL

O Objetivo deste trabalho é apresentar duas metodologias para realizar a caracterização de sequências genômicas originárias de regiões distintas de um DNA. Nesse estudo é apresentado uma análise entre as metodologias propostas assim como seus resultados, avaliando os mesmos em relação a caracterização e classificação das sequências no âmbito da bioinformática.

1.1.2 OBJETIVOS ESPECÍFICOS

- Retratar uma sequência genômica como uma rede complexa
- Aplicar algoritmo para extração de medidas das metodologias.
- Realizar a classificação com as medidas extraídas.
- Analisar resultados.

1.2 JUSTIFICATIVA

A relação entre extração de características (medidas) e a classificação adequada é um tanto arbitrária. Um extrator de característica ideal seria produzir uma representação que torna o trabalho do classificador trivial, inversamente, um classificador onipotente não precisaria da ajuda de um sofisticado extrator característica. O objetivo tradicional do extrator característica é caracterizar um objeto a ser reconhecido por meio de medições cujos valores são muito semelhantes para os objetos de uma mesma categoria, e muito diferente para os objetos em diferentes categorias (DUDA; HART; STORK, 2012).

Olhando por este lado, a motivação deste trabalho é propor o estudo de um novo método de extração de medidas para a classificação de sequências genômicas. Para isso serão realizado dois experimentos. Em um primeiro momento serão extraídos as seguintes medidas das sequências para realizar a classificação: histograma, desvio padrão, entropia, soma de entropia e maximização de entropia. Logo após, serão extraídas e utilizadas medidas de redes complexas (centralidade, desvio padrão do grau, mínimo e máxima do grau, transitividade, caminho mínimo médio, número de comunidades e *motifs*) baseando-se na metodologia aplicada no artigo *A complex network-based approach for texture analysis* (BACKES; CASANOVA; BRUNO, 2010) onde ao invés de pixels, é utilizado os nucleotídeos de uma sequência genômica para formação da rede.

Este trabalho segue com a contextualização dos conceitos abordados através da revisão bibliográfica, a qual pode ser encontrada na seção 2. Os materiais e métodos são apresentados na seção 3. Na seção 4 é apresentado os resultados obtidos das metodologias aplicadas. Sequencialmente, sendo apresentado o cronograma realizado na seção 5 e as considerações e conclusão na seção 6.

2 REVISÃO BIBLIOGRÁFICA

2.1 GENÉTICA

Um nucleotídeo é uma molécula composta por um açúcar chamado pentose, um grupo de fosfato e uma base nitrogenada (GRIFFITHS, 2008). Nucleotídeos são diferenciados através da base nitrogenada que os compõem, podendo variar em: adenina, citosina, guanina, timina ou uracila. A pentose de um nucleotídeo pode se ligar ao grupo fostafo de um outro, formando uma cadeia. Uma cadeia formada por vários nucleotídeos é chamada de polinucleotídeo.

Existem dois tipos de polinucleotídeos que armazenam informações genéticas: o DNA e o RNA.

O DNA é formado por duas fitas de nucletotídeos, e a pentose que o o constitui é a desoxirribose. As duas fitas do DNA são unidas através de pontes de hidrogênio formadas entre a suas quatros bases nitrogenadas. A adenina sempre forma pontes de hidrogênio com a timina, e a citosina com a guanina. As duas fitas do DNA são ditas complementares, e sempre é possível construir uma fita a partir da outra. A sequência de bases nitrogenadas ao longo da cadeia de DNA constitui a informação genética.

O RNA é composto por apenas uma fita e sua pentose é a ribose. A base nitrogenada timina, exclusiva do DNA, é substituida pela uracila, exclusiva do RNA. Uma fita de RNA pode se dobrar de tal forma que parte de suas próprias bases nitrogenadas se pareiam umas com as outras. Esse pareamento intramolecular é um fator importante no formato tridimensional do RNA, que é capaz de assumir uma variedade maior de formas complexas do que a dupla hélice do DNA.

2.2 FORMATO FASTA

O formato FASTA é um formato baseado em texto para representar tanto sequencias de nucleótidos quanto sequencias de peptídeos, no qual os nucleotídeos ou aminoácidos são representados usando códigos de uma única letra (MARKEL; LEON, 2003).

Inicialmente criado por David J. Lipman and William R. Pearson em 1985 em um pacote de softwares chamado de FASTP, veio a se tornar o formato padrão usado na Bioinformática devido a sua simplicidade que torna fácil manipular e analisar sequências usando ferramentas de processamento de texto e linguagens de script como Python, Ruby, e Perl.

A primeira linha em um arquivo FASTA começa com um símbolo '>' (maior que) junto a uma palavra indicando o identificador da sequências, o restante é considerado descrição sendo este opcional. Após a linha inicial vem a sequência em si no padrão de código de uma letra (A, C, T, G). Outra coisa além de um código válido é ignorada (incluindo espaços, tabulações, asteriscos, etc ..).

Exemplo de um arquivo fasta:

>MCHU
ATGGCAGAGACCGCGGCCGGAGTGGGCCGCTTCAAGACCAACTAT*

2.3 HISTOGRAMA

Um histograma é a frequência de um conjunto de dados quantitativo na forma de um gráfico em barras. O histograma consiste basicamente em um plano cartesiano cujo no eixo x (horizontal) se encontra a amostra e no eixo y (vertical) sua frequência.

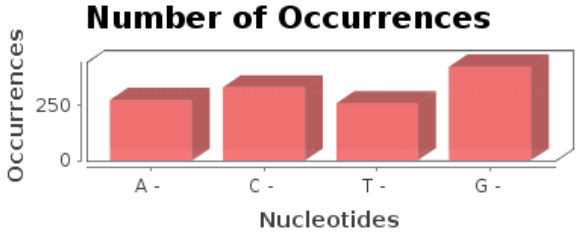


Figura 1: Exemplo de um Histograma de nucleotideos.

2.4 ENTROPIA E INFORMAÇÃO MÚTUA

Na metade do século XIX Rudolf Clausius através de seus estudos durante um ciclo de funcionamento da máquina térmica de Carnot, chegou a um quociente que denominou de entropia, dando origem a segunda lei da termodinâmica (COVENEY; HIGHFIELD, 1991). Por

volta de 1866, Ludwig Boltzmann definiu quantitativamente o conceito de entropia relacionado a desordem na termodinâmica estatística. Posteriormente, a entropia foi definida na Teoria da Informação por Claude Shannon (SHANNON, 2001), onde a mesma indicaria a quantidade de informação contida em uma determinada fonte, mas também, podendo graduar a desordem de um conjunto de dados (BISHOP, 1995), onde quanto maior a quantidade de símbolos incertos e padrões em um determinado conjunto de dados, maior a entropia de todo o conjunto (LOPES, 2003). Dentre as equações que definem entropia nos variados contextos a utilizada neste trabalho é a da Teoria da Informação, definida como:

$$H(X) = -\sum_{x \in X} P(x) log P(x) \tag{1}$$

2.4.1 SOMA E MAXIMIZAÇÃO DE ENTROPIA

Uma vez que a entropia mede a quantidade de informação em determinada fonte, a sua soma representa uma das etapas para se encontrar o ponto máximo da informação encontrada, sendo a soma dada por (JAYNES, 1957):

$$f(X) = \sum_{i=1}^{n} P_i f(x_i) \tag{2}$$

e a maximização por:

$$S_{max} = \lambda_0 + \lambda_1(f_1(x)) + ... + \lambda_m(f_m(x)) \tag{3}$$

2.5 REDES COMPLEXAS

O estudo dos grafos teve início em 1736, quando Euler ao se deparar com um conjunto de pontes em Königsberg propôs uma solução para atravessar todas as pontes sem repetir nenhuma, originando a teoria e possivelmente o primeiro grafo da história (BOLLOBÁS, 1998).

Considerada uma extensão da teoria dos grafos (COSTA et al., 2007), redes complexas é o termo que define um grafo cujo apresenta uma estrutura irregular composta por ligações constituídas de vértices (nós) interligados por arestas (BARABÁSI, 2002), sendo um tema multidisciplinar que abrange várias ciências, tais como a biologia, ciência da computação, física, matemática e sociologia (BARABÁSI, 2003).

Após a teoria de Euler os avanços dos estudos desencadearam aplicações práticas em di-

ferentes áreas. Como por exemplo a Sociologia, onde sociólogos a partir da década de 50 que tinham finalidade de estudar o comportamento e a relação entre as pessoas passaram a fazer uso de redes complexas em suas pesquisas. Tais pesquisas eram embasadas em características familiares das redes, como a centralidade (o vértice mais central) e a conectividade (vértices com maior número de conexões). Nesse cenário, a centralidade e a conectividade eram usadas para determinar quem melhor se relacionava com os demais ou identificar os indivíduos que passavam maior influência (WASSERMAN, 1994).

Com o tempo, devido ao avanço tecnológico, computadores e meios de comunicação permitiram a análise de grandes quantidades de informações. Tal acontecimento fez com que pesquisas aos quais estudavam apenas redes pequenas observando propriedades de vértices individuais passassem a observar propriedades em larga-escala, como comunidades identificadas dentro das redes.

Neste contexto, diversas abordagens do mundo real podem ser representadas por meio de redes complexas, como por exemplo, a conexão entre os aeroportos (GUIMERA; AMARAL, 2004), (GUIMERÀ et al., 2005); a rede da Internet conectada por *hyperlinks* em documentos (ALBERT; JEONG; BARABÁSI, 1999), (BARABÁSI; ALBERT; JEONG, 2000), (ADAMIC; HUBERMAN, 2000); a relação entre grupos de pessoas (WASSERMAN, 1994) e redes neurais (COSTA; SPORNS, 2005; RUBINOV; SPORNS, 2010).

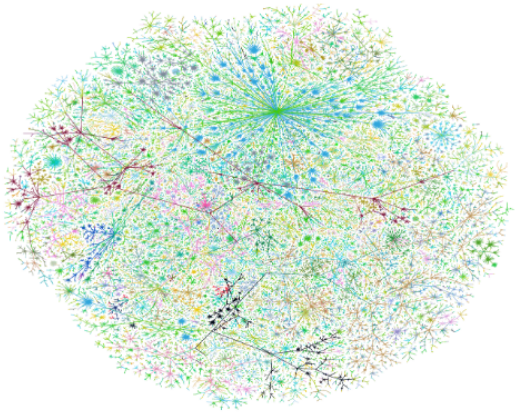


Figura 2: Representação da estrutura da Internet

Fonte: (NEWMAN, 2003)

Em termos computacionais, as redes complexas podem ser representadas como listas ou matrizes de adjacências. No caso da lista, apenas os vértices conectados são armazenados. Já

no caso da matriz de adjacências A, se dois vértices i e j estiverem conectados, o valor de a_{ij} será igual a 1, caso contrário 0. No caso de mais conexões entre vértices já conectados, então o valor de a_{ij} será respectivo ao número de ligações (peso da aresta).

Em uma rede, suas conexões podem ser dirigidas e não dirigidas. Quando dirigida, o sentido da ligação de um vértice a outro importa, caso contrário, será não-dirigida. Se as ligações possuem intensidade, então a cada aresta é atribuído um peso representando a mesma.

2.5.1 TIPOS DE REDES

Nesta seção são descritos de forma breve os três principais modelos teóricos de redes complexas: rede aleatória, rede pequeno-mundo e rede livre de escala.

Rede Aleatória. Esse modelo é conhecido como rede ER e é considerado o mais simples que uma rede complexas pode assumir. Proposto por Paul Erdös e Alfréd Rényi em 1959 (ERDŐS; RÉNYI, 1959), arestas não direcionadas são adicionadas aleatoriamente entre um número fixo de vértices. Com base nesse relato, Erdös e Rényi concluíram que todos os vértices de uma determinada rede tem a aproximadamente a mesma quantidade de conexões, e também, a mesma chance de receber novas conexões (BARABÁSI; ALBERT, 1999). Os autores afirmam também que quanto mais complexa a rede, maior a chance dela ser aleatória.

Figura 3: Rede complexa aleatória
Fonte: (ALBERT; BARABÁSI, 2002)

Rede Pequeno-Mundo.

Conhecida por rede WS, segundo os autores Watts e Strogatz (WATTS; STROGATZ, 1998), em muitas redes são encontrados padrões relacionados a conexão dos vértices cujo tendem a formar pequenas conexões em cada. Pensando dessa maneira, eles propuseram um modelo similar ao ER, baseando-se no fenômeno mundo pequêno (MILGRAM, 1967) ao qual supõe que a média de distância de conexão de uma pessoa a qualquer outra é de 6 pessoas, assim determinando

no contexto das redes complexas que grande parte das conexões são estabelecidas entre os vértices mais próximos. Deste modo, o modelo se assemelha ao de Erdös e Rény no sentido de que grande parte das conexões serão formadas entre vértices próximos uns dos outros, tendo como consequência, o valor da distância média entre quaisquer dois vértices aleatórios vir a ser um pequeno número de vértices. Para que isso ocorra, basta que algumas conexões aleatórias entre grupos sejam estabelecidas.

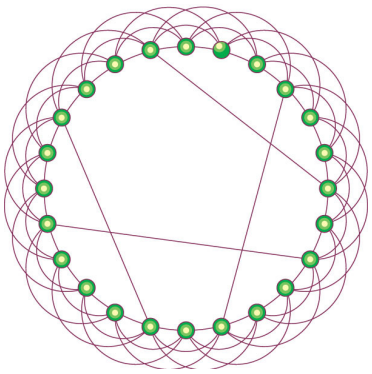


Figura 4: Rede complexa pequeno-mundo

Fonte: (STROGATZ, 2001)

Rede Livre De Escala.

O Modelo desenvolvido por Barabási e Albert (BA) (BARABÁSI; ALBERT, 1999) demonstram que as redes podem assumir uma ordem dinâmica de estruturação com características bem específicas, como a conexão preferencial, que é uma das principais características encontradas nesse tipo de rede ao qual representa a tendência de um novo vértice se conectar a um vértice que tem um grau elevado de conexões. Essa característica simboliza redes com poucos vértices altamente conectados, e muitos vértices com poucas conexões. Essas redes tem sido observadas em várias abordagens do mundo real como por exemplo, na internet demonstrada na Figura 2, em redes de metabolismo e redes de citações de artigos científicos.

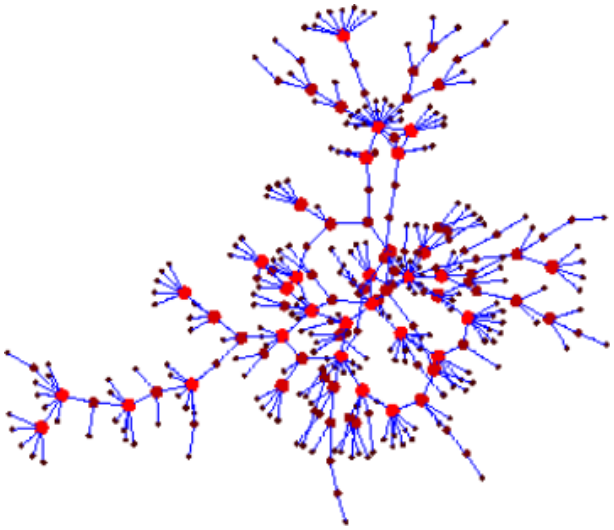


Figura 5: Rede complexa livre de escala

Fonte: (STROGATZ, 2001)

2.5.2 MEDIDAS

Dentre as medidas possíveis de se extrair de uma rede, as que foram usadas neste trabalho são:

1. **Caminho mínimo médio.** O comprimento do menor caminho entre dois vértices i e j, d_{ij}, é dado pela extensão de todos os caminhos que conectam estes vértices cujos comprimentos são mínimos (WATTS; STROGATZ, 1998). Sua determinação é importante para caracterizar a estrutura interna das redes e não investigação de efeitos dinâmicos relativos ao transporte e à comunicação (BOCCALETTI et al., 2006). Dado uma matriz de distâncias D, cujos elementos d_{ij} representam o valor do menor caminho entre os vértices i e j. A média entre os valores na matriz D expressa o menor caminho médio, sendo calculada por:

$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \tag{4}$$

2. **Coeficiente de Cluster.**

   O coeficiente de cluster ou transitividade é uma medida de aglomeração que representa a probabilidade entre os vértices adjacentes de um dado vértice estarem conectados, como por exemplo em uma rede social, pode ser representado como a probabilidade entre dois amigos (A e B) terem um amigo (C) em comum. Dependendo da topologia da rede, o valor da transitividade pode ser diferente. Segundo (BARRAT et al., 2004), a transitividade

pode ser obtida através da seguinte equação:

$$C^{\omega}(i) = \frac{1}{S_i(k_i - 1)} \sum_{j,h} \frac{e_{i,j} + e_{i,j}}{2} a_{i,j} a_{i,h} a_{j,h} \tag{5}$$

3. **Centralidade.** No âmbito da teoria dos grafos e redes complexas, existem diferentes tipos de centralidade, tais como:

   - Centralidade de Grau: A centralidade de grau é definida como o número de ligações incidentes sobre um vértice No caso de uma rede direcionada é de costume definir duas medidas distintas de centralidade de grau: *indegree* e textsloutdegree.

     Indegree é o número de ligações que são recebidas pelo vértice e *outdegree* é o número de ligações que parte do vértice para outros.

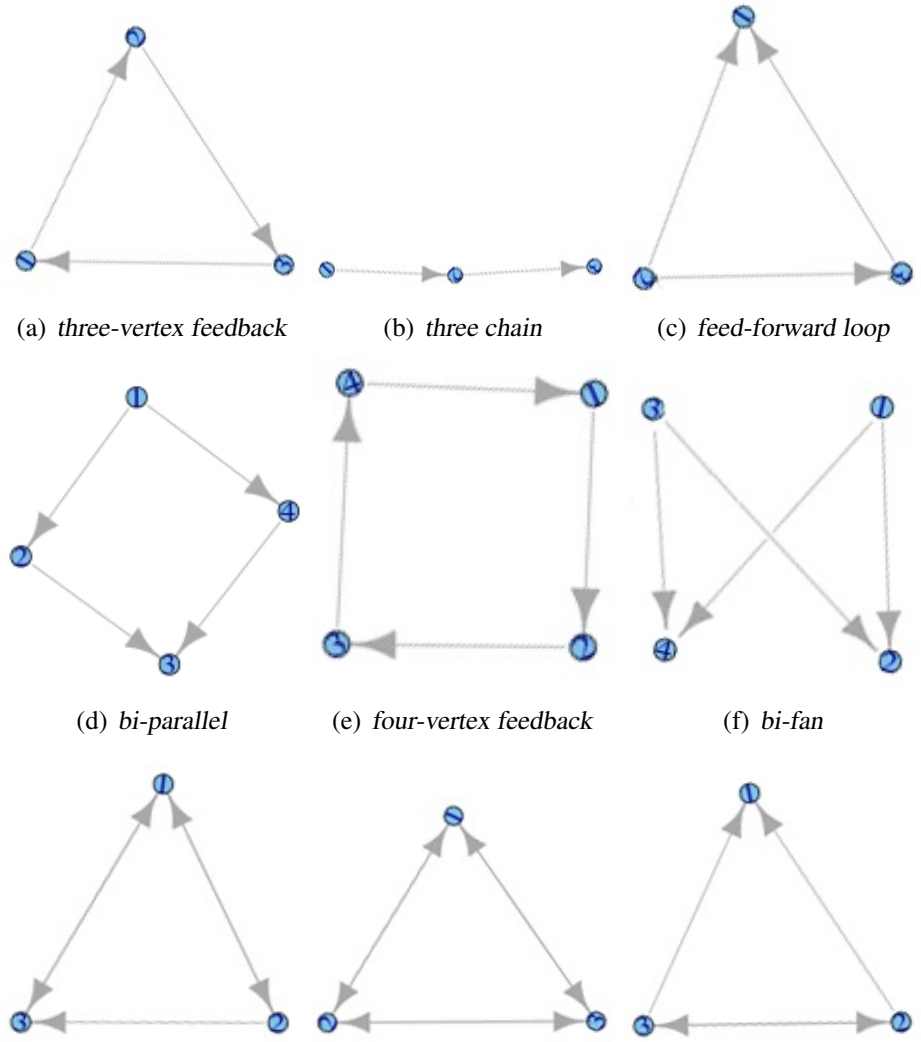
   - Centralidade de Proximidade: A centralidade de proximidade é a distância natural entre um nó a todos os outros. Ou seja, quanto mais central é o nó, menor a distância do seu total para todos os outros.

   - Centralidade de Intermediação: A Centralidade de intermediação é uma medida que quantifica o número de vezes que um vértice age como intermediário em um caminho entre dois nós (FREEMAN, 1977). Por exemplo, em uma rede social, o número de vezes que uma pessoa serve de ponte para duas outras se conhecer.

   - Centralidade de Eficiência: A centralidade de eficiência indica a excentricidade de um vértice em relação a outro, ou seja, indica o caminho mais rápido para se conectar com um outro vértice, onde quanto menor for sua excentricidade mais eficiente é o vértice.

4. **Grau médio.** O grau médio é a média aritmética de graus existentes dentro da rede, podendo ser obtido pela divisão do número de arestas pelo número de vértices.

5. ***Motifs.*** Os *motifs* são subgrafos identificados com grande frequência dentro de uma rede complexa. Segundo (MILO et al., 2002), *motifs* estão diretamente relacionados à estrutura e evolução das redes complexas.



(a) *three-vertex feedback* (b) *three chain* (c) *feed-forward loop*

(d) *bi-parallel* (e) *four-vertex feedback* (f) *bi-fan*

(g) *feedback with two mutual dyads* (h) *fully connected triad* (i) *uplinked mutual dyad*

Figura 6: Exemplos de *motifs*

Fonte: (COSTA et al., 2007)

6. **Número de comunidades.**

   A maioria das redes costumam ser modulares, isto é, as conexões são mais frequentes entre vértices que pertençam a um mesmo grupo e menos frequentes entre vértices de grupos distintos. Esses módulos são definidos como comunidades por terem seus vértices altamente conectados entre si e poucos conectados com o restante da rede (DANON et al., 2005).

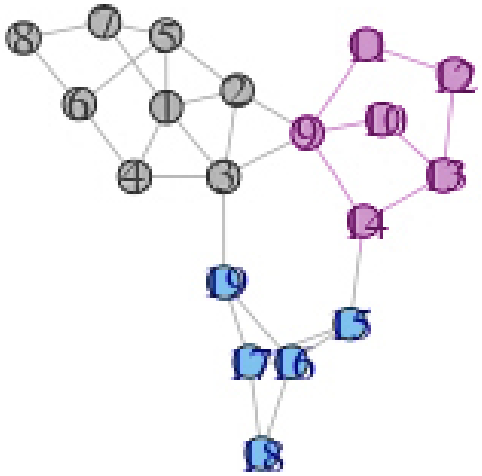


Figura 7: Exemplo de uma rede modular.

Fonte: Autoria própria.

2.6 WEKA

O WEKA (Waikato Environment for Knowledge Analysis) (HALL et al., 2009) é atualmente um sistema de referencia na área de aprendizado de máquina e mineração de dados com ampla aceitação nos meios acadêmicos e áreas afins.

Iniciado em 1993 pelo governo da Nova Zelândia na Universidade de Waikato, teve suas primeiras versões desenvolvidas em linguagem C mas devido à algumas dificuldades encontradas o sistema foi todo reescrito em JAVA.

A estrutura padrão para manipulação de dados do WEKA é um arquivo de texto do tipo ARFF. Um arquivo ARFF tem sua estrutura composta na seguinte ordem: O cabeçalho indicado com marcação @relation representando o identificador do arquivo. A sequência de atributos (um em cada linha) e as classes que as instâncias podem pertencer indicados com a marcação @attribute. Por fim os dados, que devem estar situados abaixo da marcação @data.

O vetor de característica que representará cada instância do conjunto de dados deve ser composto na mesma ordem da definição dos atributos, onde seus valores devem ser separados por uma vírgula como exibido na Figura 8.

```
@RELATION exemplo

@ATTRIBUTE tamanho  NUMERIC
@ATTRIBUTE largura  NUMERIC
@ATTRIBUTE class {classeA,classeB,classeC}

@DATA
5.1,3.5,ClasseA
4.9,3.0,ClasseA
4.7,3.2,ClasseA
4.6,3.1,ClasseB
5.0,3.6,ClasseB
5.4,3.9,ClasseB
4.6,3.4,ClasseC
5.0,3.4,ClasseC
4.4,2.9,ClasseC
```

Figura 8: Exemplo de um arquivo Arff.

Fonte: Autoria Própria.

O WEKA dispõe de um elevado número de classificadores, tais como: Árvore de Decisão induzida, Tabelas de decisão, Regressão local de pesos, Regressão lógica, Naive Bayes, SVM, etc.

Dentre os resultados gerados na saída pelo WEKA, são normalmente mais avaliados os seguintes valores:

Instâncias Classificadas Corretamente (***Correctly Classsified Instances***): porcentagem de registros que foram classificados corretamente durante a elaboração do modelo de classificação.

Instâncias Classificadas Incorretamente (***Incorrectly Classsified Instances***): porcentagem de registros que foram classificados incorretamente durante a elaboração do modelo de classificação.

Precisão e Recall (***Precision and Recall***): são métricas normalmente vistas na área de reconhecimento de padrões (BAEZA-YATES; RIBEIRO-NETO et al., 1999). Quando utilizadas, o conjunto de amostras é dividido em dois subconjuntos ao qual um dos é considerado relevante em relação ao objetivo da métrica. O Recall é representado como a porção de instâncias corretas dentro todas as instâncias do subconjunto relevante, e a Precisão é a porção de instâncias correta dentro as que o algoritmo indicou pertencer ao subconjunto relevante. Quanto mais próximo de um, melhor a classificação.

Falso Positivo (***False Positive***): é a porcentagem dos casos onde o modelo criado diz que deve ser positivo, mas seu valor valor real é negativo. Quanto mais próximo de 0 melhor a classificação.

Verdadeiro Positivo (***True Positive***): é a porcentagem de casos verdadeiros dentre todos os casos verdadeiros. Quanto mais próximo de 1 melhor a classificação.

Matriz de Confusão: (***Confusion Matrix***): é uma matriz quadrada que indica as classificações corretas e erradas. A classe que está sendo representada aparece na linha. As classificações aparecem nas colunas. A diagonal da matriz indica às classificações corretas.

2.7 MINERAÇÃO DE DADOS

A mineração de Dados (*Data Mining*) pode ser definida como um processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos em um conjunto de dados (FAYYAD et al., 1996). É uma abordagem multidisciplinar que incorpora técnicas utilizadas em diferentes áreas como Inteligência Artificial, Aprendizado de Máquina, Base de Dados e Estatística (REZENDE et al., 2003).

Um forte exemplo que virou matéria na mídia no mundo todo foi um relato publicado pela empresa americana *Walmart* onde a mesma informou que através do estudo em suas bases de dados encontrou uma forte relação entre cerveja e fraldas, sendo esta obtida através da análise do histórico de vendas onde os dois itens foram identificados juntos várias vezes em diferentes compras (DENNIS; MARSLAND; COCKETT, 2001).

Uma das principais tarefas da mineração de dados é a predição, sendo esta, dividida em dois problemas centrais: a classificação e a regressão (WEISS, 1998).

A classificação é uma função que mapeia determinados dados de entrado e um número limitado de categorias. Nela cada amostra pertence a uma classe, entre um conjunto predefinido de classes. Tais amostras consistem de um conjunto de atributos (vetor de características). O algoritmo de classificação por sua vez, tem como objetivo encontrar um relacionamento entre os atributos passados e a classe.

Assim, o processo de classificação consiste em obter um modelo baseado em um determinado conjunto de dados para predizer a classe de um exemplo novo e desconhecido.

Reconhecer padrões dentro de um conjunto de dados é uma das três etapas encontradas atualmente dentro da mineração de dados. Estes padrões servem como chave para realizar a classificação de diferentes grupos contidos em um mesmo conjunto. Como no caso deste trabalho, onde através das medidas extraídas nas metodologias é realizada a classificação das regiões representadas pelos conjuntos de dados escolhidos, sendo os atributos utilizados pelo classificador para separar as classes envolvidas, os padrões encontrados.

A regressão é similar a classificação se comparado seus conceitos. Sua principal diferença é que a regressão como resultado gera o valor de uma variável dependente desconhecida do valor de outras variáveis desconhecidas. Como por exemplo o preço de uma casa (variável dependente), que é resultado de muitas variáveis independentes, tais como: a metragem quadrada da casa, tamanho do terreno, revestimento interno, bairro em que a casa é situada, possíveis reformas, etc. O modelo da regressão nesse caso é criado com base nos preços de outras casas semelhantes a mesma cujo quer encontrar o preço.

Ambas as tarefas citadas da predição foram exemplificadas de maneira superficial mas o suficiente para começar os estudos e utilização de ferramentas que lidam com essas tarefas encontradas na mineração de dados. É o caso da ferramenta WEKA já descrito anteriormente, onde o mesmo foi utilizado para realizar a classificação neste trabalho utilizando os seguintes métodos classificadores: *Instance-based learning algorithms* (IBK) (AHA; KIBLER; ALBERT, 1991), J48 (QUINLAN, 1993), *MultiLayer Perceptron* (RUSSELL et al., 1995), *Naive Bayes* (NB) (JOHN; LANGLEY, 1995), *RandomForest* (BREIMAN, 2001) e Support Vector Machines (SVM) (ABE, 2010).

2.7.1 VALIDAÇÃO CRUZADA

O método de Validação Cruzada consiste em dividir o conjunto de amostras em n subconjuntos (*folds*) (KOHAVI et al., 1995). Após dividido, uma será a base de testes para a validação do modelo e os $n-1$ restantes serão utilizados para o treinamento. O processo de validação cruzada é repetido n vezes, de forma que cada subconjunto seja usada uma vez como base de testes para a validação do modelo.

No final do processo, é calculado a média dos resultados obtidos na classificação de cada subconjunto afim de obter o desempenho médio do classificador nos n testes. O propósito de repetir o processo n vezes, é aumentar a confiabilidade da estimativa da precisão do classificador.

3 MATERIAIS E MÉTODOS

3.1 BANCO DE DADOS DBTSS

O DBTSS é um banco de dados que contém posições exatas dos locais de início de transcrição (tss), determinado com a técnica denominada tss-seq nos genomas de várias espécies (YAMASHITA et al., 2012). Para este trabalho, foi utilizado um conjunto de dados composto por 1500 sequências.

O conjunto de dados se encontra disponível em `ftp://ftp.hgc.jp/pub/hgc/db/dbtss/Yamashita_NAR/`.

3.2 GENOME BROWSER

As sequências codificantes de proteína foram extraídas a partir de um conjunto de 1.600 genes RefSeq selecionados aleatoriamente a partir do genoma humano (hg18). Para obter as sequências que codificam proteínas, a região intrônica de cada gene foi removida e os segmentos de codificação foram concatenados. Usando o mesmo genoma foram também selecionados aleatoriamente 1.520 regiões não-codificantes. Obteve-se a anotação de cada gene e o genoma hg18 do banco de dados do navegador UCSC Genome (KENT et al., 2002).

3.3 METODOLOGIA

Nesta seção são apresentados os dois métodos utilizados para realizar a caracterização das sequências genômicas.

Para extração das medidas das duas metodologias foram implementados algoritmos em JAVA, aos quais se encontra em mais detalhes nas subseções abaixo.

3.3.1 MÉTODO 1: TEORIA DA INFORMAÇÃO

Para esta metodologia, as rotinas implementadas calculam a ocorrência de nutleotídeos, dinucleotídeos e trinucleotídeos dentro da sequência.

Dado um arquivo fasta é realizado o processamento do mesmo onde para cada sequência identificada é criado um diretório que irá conter a princípio arquivos de texto com a frequência de nucleotídeos, dinucleotídeos e trinucleotídeos da mesma.

Após a criação dos diretórios de todas as sequências, é realizado uma varredura dos mesmos para se obter as medidas que irão formar o vetor de característica da sequência onde é essas medidas encontradas serão salvas em novos arquivos de texto e representadas por histogramas e gráficos de linha. Por fim, uma nova rotina é executada para gerar um arquivo arff para ser classificado no WEKA.

Cada sequência irá gerar 3 vetores (nucleotídeos, dinucleotídeos e trinucleotídeos) de características constituído pelos seguintes valores: histograma (número de ocorrência), entropia do histograma, soma de entropia, e valor máximo da maximização da entropia. Os resultados obtidos na classificação serão apresentados logo mais abaixo na seção 4.

Abaixo temos um exemplo de uma sequência CDS e alguns de seus arquivos gerados.

AAMP
ATGGAGTCCGAATCGGAAAGCGGGGCTGCTGCTGACACCCCCCCACTGGAGACC
CTATGGAGTCCGAATCGGAAAGCGGGGCTGCTGCTGACACCCCCCCACTGGAGA
CCCTAAGCTTCCATGGTGATGAAGAGATTATCGAGGTGGTAGAACTTGATCCCGG
TCCGCCGGACCCAGATGACCTGGCCCAGGAGATGGAAGATGTGGACTTTGAGGA
GAAGAGGAGGAAGAGGGCAACGAAGAGGGCTGGGTTCTAGAACCCCAGGAAG*

Tabela 1: Histograma de Nucleotídeos

Valores do histograma de nucleotideos da figura 9			
A - 276.0	C - 335.0	T - 262.0	G - 429.0

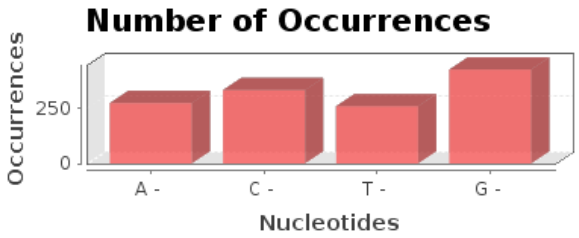


Figura 9: Histograma de nucleotídeos da sequência AAMP.

Tabela 2: Histograma de dinucleotídeos

Valores do histograma de dinucleotídeos da figura 10			
AA - 51.0	AC - 72.0	AT - 47.0	AG - 106.0
CA - 84.0	CC - 115.	CT - 97.0	CG - 39.0
TA - 26.0	TC - 54.0	TT - 43.0	TG - 137.0
GA - 113.0	GC - 94.0	GT - 75.0	GG - 146.0

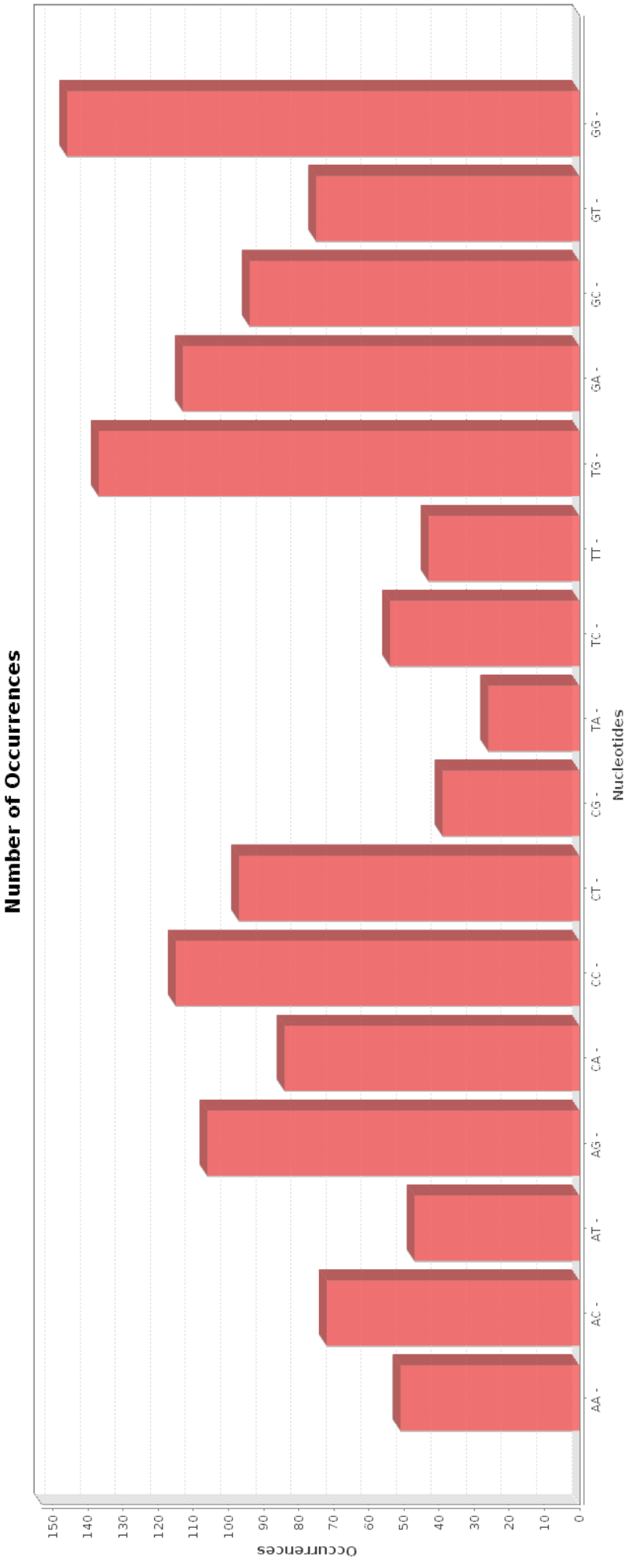


Figura 10: Histograma de dinucleotídeos da sequência AAMP.

Tabela 3: Histograma de trinucleotídeos

Valores do histograma de trinucleotídeos da figura 11			
AAA - 10.0	AAC - 8.0	AAT - 3.0	AAG - 30.0
ACA - 9.0	ACC - 33.0	ACT - 23.0	ACG - 7.0
ATA - 5.0	ATC - 13.0	ATT - 4.0	ATG - 25.0
AGA - 28.0	AGC - 26.0	AGT - 18.0	AGG - 34.0
CAA - 14.0	CAC - 22.0	CAT - 14.0	CAG - 34.0
CCA - 34.0	CCC - 28.0	CCT - 38.0	CCG - 15.0
CTA - 11.0	CTC - 17.0	CTT - 18.0	CTG - 50.0
CGA - 10.0	CGC - 5.0	CGT - 6.0	CGG - 18.0
TAA - 4.0	TAC - 7.0	TAT - 8.0	TAG - 7.0
TCA - 12.0	TCC - 21.0	TCT - 12.0	TCG - 9.0
TTA - 4.0	TTC - 6.0	TTT - 14.0	TTG - 19.0
TGA - 32.0	TGC - 23.0	TGT - 25.0	TGG - 57.0
GAA - 23.0	GAC - 35.0	GAT - 21.0	GAG - 34.0
GCA - 29.0	GCC - 33.0	GCT - 24.0	GCG - 8.0
GTA - 6.0	GTC - 18.0	GTT - 7.0	GTG - 43.0
GGA - 43.0	GGC - 40.0	GGT - 25.0	GGG - 37.0

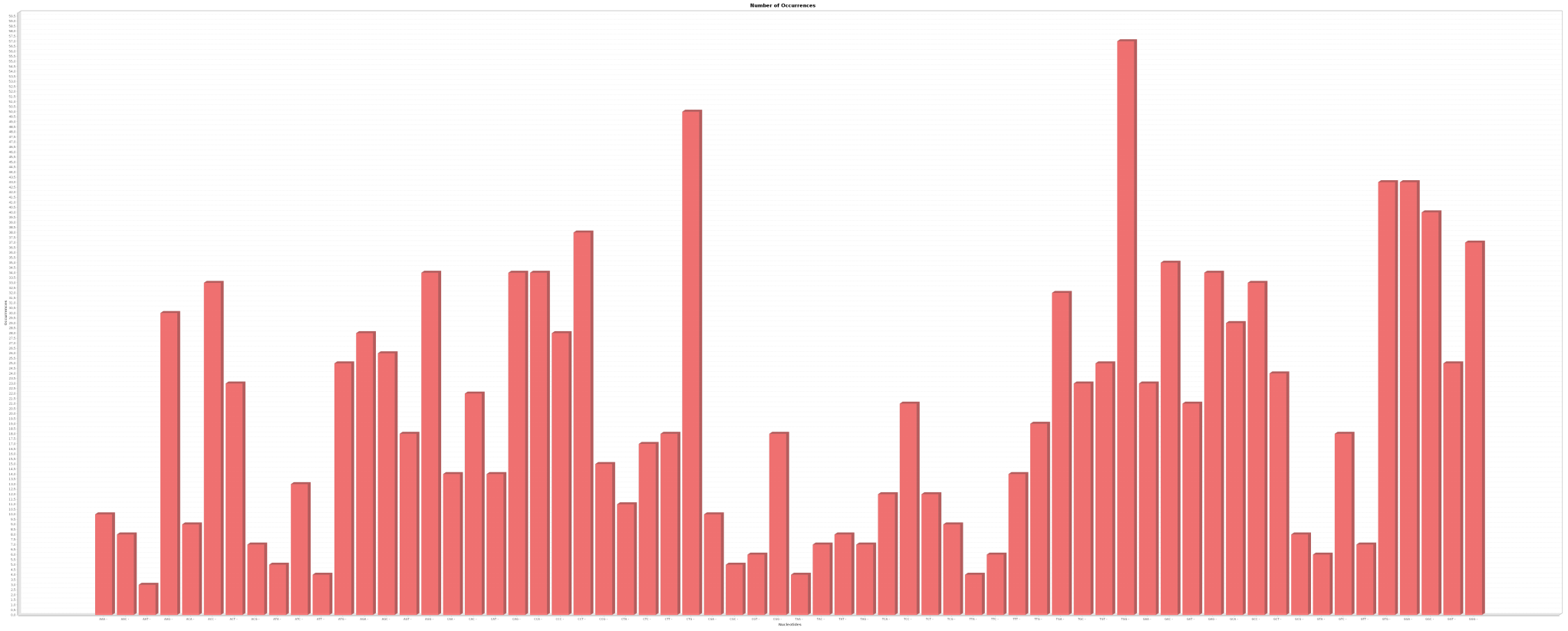


Figura 11: Histograma de trinucleotídeos

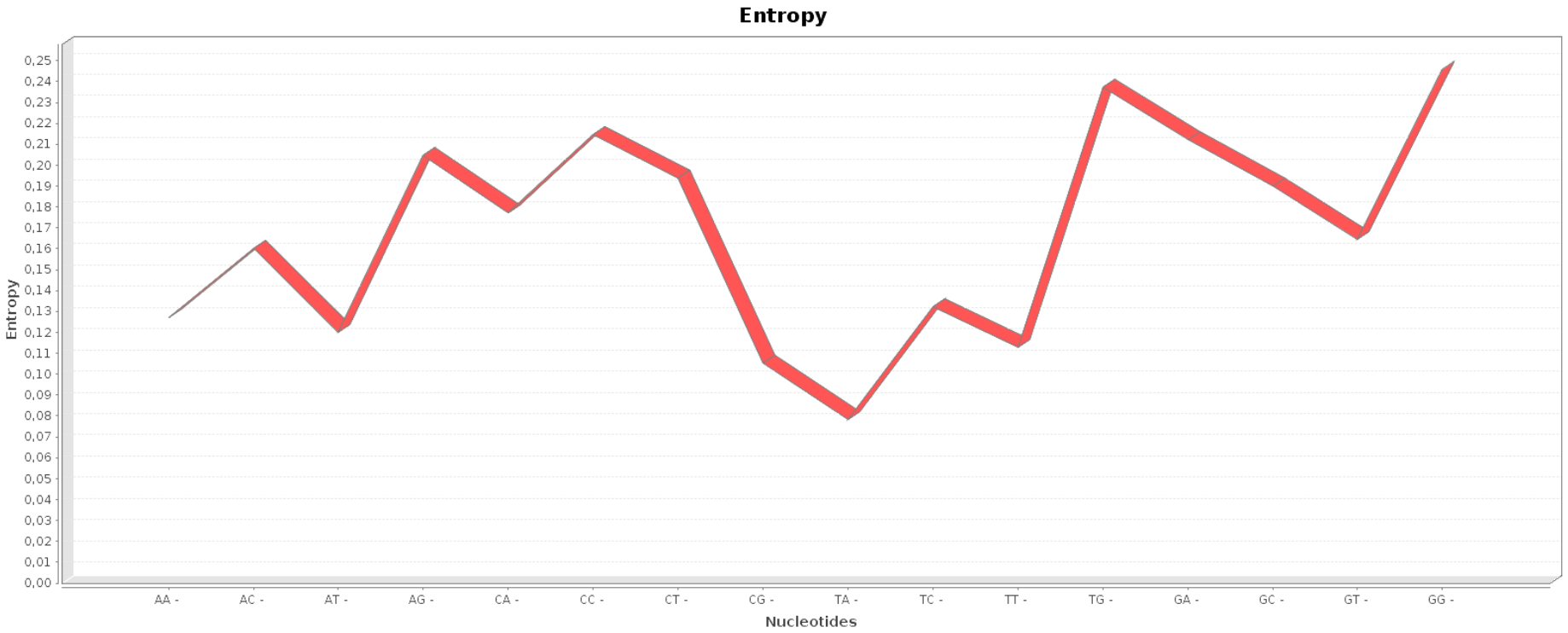


Figura 12: Entropia de trinucleotídeos

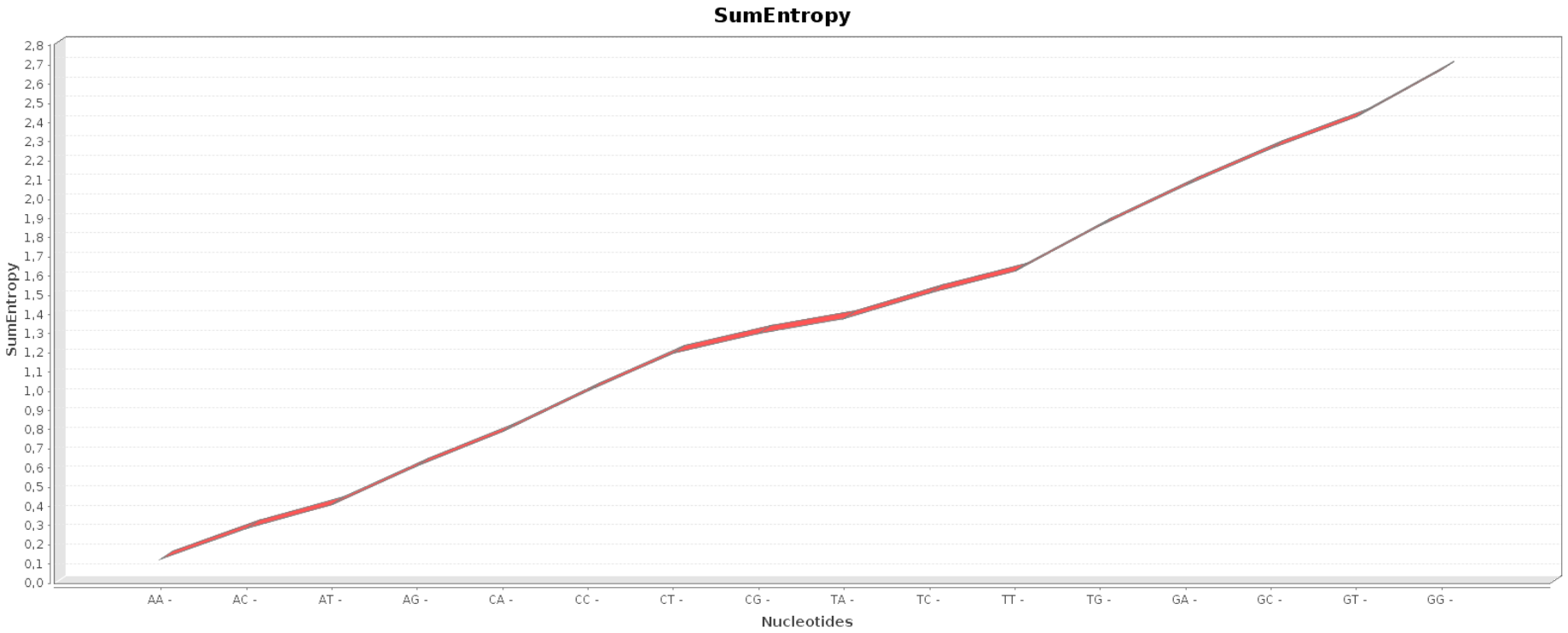


Figura 13: Soma da entropia de nucleotídeos

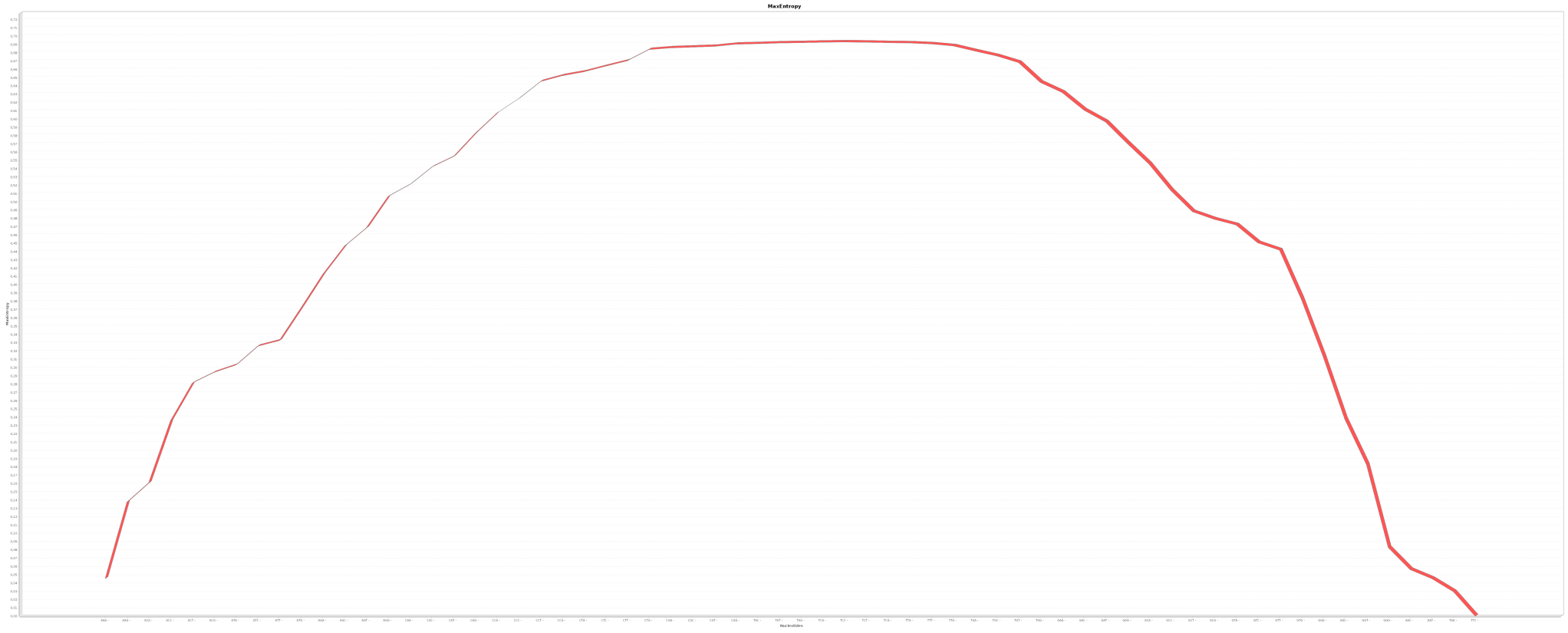


Figura 14: Maximização da entropia de trinucleotídeos

3.3.2 MÉTODO 2: REDES COMPLEXAS

No caso da segunda metodologia, as rotinas foram utilizados para retratar as redes complexas a partir das sequências genômicas em arquivos de texto.

Como o tamanho das sequências podem ser muito grande, o recurso de memória principal pode não ser suficiente para suportar a mesma. Para este problema uma logística de divisão da sequência foi implementada de acordo com o número de threads da máquina, onde para cada thread, um pedaço é atribuído para ser processado. Essa solução utiliza o acesso randômico no arquivo realizando a leitura em bytes ao invés de sequencial lendo os caracteres propriamente dito.

Assim, para cada thread é atribuído um index que será o ponteiro indicando a partir de qual byte do arquivo deve-se começar a ler, fazendo com que cada tarefa em paralelo realize o acesso de pedaços distintos e não sequenciais dentro do arquivo, de modo que os bytes correspondentes ao pedaço corrente sendo processado após seu término é descartado, poupando memória.

Os arquivos gerados se encontram em um formato específico que será interpretados pelo pacote *igraph* (CSARDI; NEPUSZ, 2006) da ferramenta *R* (TEAM et al., 2005), utilizado para extrair suas respectivas medidas.

Entre os formatos disponíveis que são interpretados pelo *R* como uma rede complexas, os testados foram os formatos *ncol* e *edgelist*

O formato *ncol* representa uma rede no arquivo de maneira simples. As ligações são escritas uma em cada linha em um arquivo em branco, onde um vértice é separado do outro por um espaço simples. Caso haja uma nova ligação entre vértices de uma conexão já existente, basta repetir a ligação como mostrado abaixo.

Tabela 4: Exemplo de um arquivo do formato *ncol*

A C
A C
C T
C G
A G

Com o formato *edgelist* não foi possível identificar como é sua representação em arquivos de texto das redes complexas. O que ficou claro é que ele não suporta letras como sendo os vértices e no caso da tentativa de representar a rede no formato semelhante ao *ncol*, o arquivo é aceito, porém o numéro de vértices que será gerado na rede criada pelo *R* é equivocado, sendo ele correspondente ao maior valor do identificador encontrado dentre os vértices existentes. Ou

seja, se for uma rede apenas com 3 vértices e um deles tem o identificador “1450”, a rede criada terá 1450 vértices. Por fim o formato utilizado foi o *ncol* devido ao sucesso da representação correta das redes após alguns testes.

Sendo assim, dado uma arquivo fasta cada sequência identificada dará origem a 6 redes complexas não direcionais. O número de redes estipulado (6) é de acordo com os parâmetros considerados para a metodologia. Considerando que os vértices da rede serão representado pela ocorrência de encontro dos nucleotideos dentro da sequência, temos os casos de nucleotídeos, dinucleotídeos e trinucleotídeos e os seguintes parâmetros:

TP: é o numero de caracteres que representará os vértices

P: é o número de casas que será andado na sequência a partir da palavra (vértice) corrente para obtenção de uma nova palavra e constituição da ligação entre os vértices.

Exemplo:

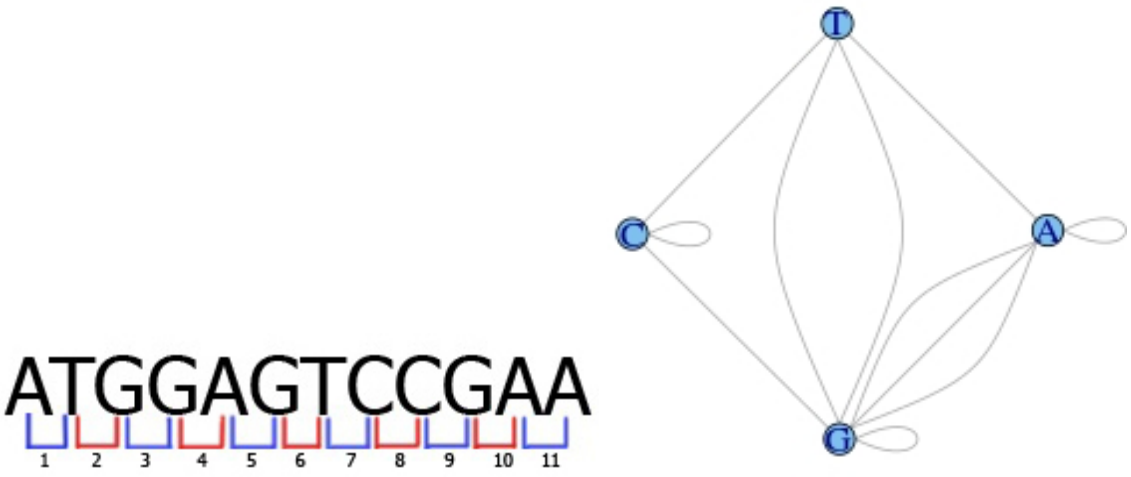
Rede de Nucleotídeos.

Paramêtros: TP = 1 e P = 1.

A ligação entre os vértices será estabelecida entre uma palavra de 1 caractere com a próxima palavra a sua frente considerando o tamanho do passo em questão, no caso 1 caractere.

Considerando a seguinte sequência: ATGGAGTCCGAA

As ligações encontradas de acordo com os parâmetros estabelecidos para a rede de nucleotídeos será:



(a) (b)

Figura 15: Rede de nucleotídeos considerando a sequência ATGGAGTCCGAA com os parâmetros P = 1 e TP = 1. (a) como os nós são definidos e (b) a rede resultante .

Tabela 5: Ligações da rede de Nucleotídeos

Ligações da rede de nucleotideos da Figura 15	
Ligação	Peso da aresta
A - A	1
A - T	1
A - G	3
T - C	1
T - G	2
C - C	1
C - G	1
G - G	1

Lembrando que por ser uma rede não direcional, a ligação entre os vértices será a mesma ignorando a origem e destino dos vértices.

Um caso mais específico seria o de uma rede de dinucleotídeos. A mesma é mostrada logo abaixo:

Redes de dinucleotídeos.

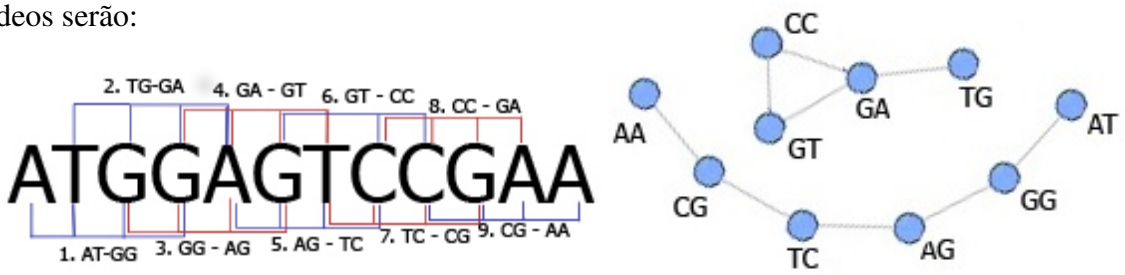
Paramêtros: TP = 2 e P = 1.

A ligação entre os vértices será estabelecida entre uma palavra de 2 caracteres com a próxima palavra a sua frente considerando o tamanho do passo em questão, no caso 1 caractere.

Considerando a seguinte sequência: ATGGAGTCCGAA

As ligações encontradas de acordo com os paramêtros estabelecidos para a rede de dinucle-

otídeos serão:



(a) (b)

Figura 16: Rede de dinucleotídeos considerando a sequência ATGGAGTCCGAA com os parâmetros P = 1 e TP = 2. (a) como os nós são definidos e (b) a rede resultante .

Tabela 6: Ligações da rede de Dinucleotídeos

Ligações da rede de dinucleotídeos da Figura 16(b)	
Ligação	Peso da aresta
AT - GG	1
TG - GA	1
GG - AG	1
GA - GT	1
AG - TC	1
GT - CC	1
TC - CG	1
CC - GA	1
CG - AA	1

Para a validação das medidas extraídas das redes foi utilizado um modelo pequeno (*Toy Model*), como demonstrado abaixo.

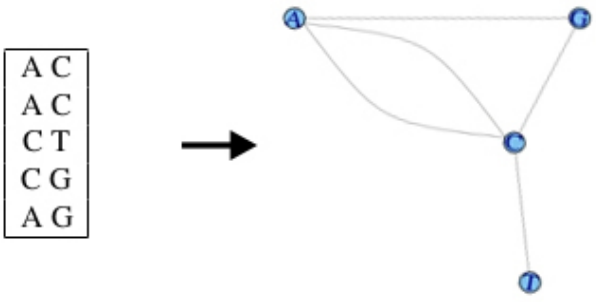


Figura 17: Rede ToyModel.

Fonte: Autoria Própria.

Considerando a Figura 18, teremos os seguintes valores para as medidas extraídas:

Tabela 7: Medidas Toy Model

Caminho mínimo médio	1.333333
Coeficiente de Cluster	0.6
Coeficiente de Cluster (average)	0.7777778
Centralidade	0.1666667
Motifs 3	NA NA 2 1
Motifs 4	NA NA NA NA 0 NA 0 1 0 0 0
Número de comunidades	4
Desvio Padrão	0.4303315
Minimo	0.3333333
Máximo	1.333333

Para obtenção das medidas foram utilizadas as seguintes funções oferecidas pelo pacote *igraph*:

Caminho Minimo Médio: *average.path.length(graph, directed=FALSE, unconnected=FALSE)*

Coeficiente de Cluster: *transitivity(graph, type=c("undirected"), vids=NULL, weights=NULL, isolates=c("NaN", "zero")))* e *transitivity(graph, type=c("average"), vids=NULL, weights=NULL, isolates=c("NaN", "zero")))*

Na documentação do pacote *igraph* não foi encontrado a descrição do efeito do paramêtro *average* setado em *type*, sendo o mesmo o único a apresentar resultado diferente em relação aos demais possíveis valores. Deste modo, as duas medidas foram consideradas no vetor de características extraído de cada rede.

Centralidade: *betweenness(graph, v=V(graph), directed = FALSE, weights = NULL, nobigint = TRUE, normalized = TRUE)*

Dentre os tipos de centralidade descritos na revisão literária neste trabalho, de acordo com a documentação do *igraph*, a função *betweenness* calcula a centralidade de intermediação.

Motifs: *graph.motifs(graph, size=3)*

Dentre os tipos de motifs que podem ser encontrados dentro de uma rede, os mais considerados são os de tamanho 3 e 4, sendo estes tamanhos os únicos suportados pela função. De acordo com a documentação, a função *graph.motifs* busca em um grafo os motifs de um determinado tamanho e retorna um vetor numérico contendo o número de diferentes motifs. A ordem dos motifs é definida por sua classe de isomorfismo.

Para obtenção dos motifs de tamanho 4 foi utilizada a mesma função, tendo apenas o parâmetro *size* alterado para 4.

Número de Comunidades: *walktrap.community(graph, weights = NULL, steps = 4, mer-*

ges = TRUE, modularity = TRUE, membership = TRUE)

A função *walktrap.community* retorna um vetor com as comunidades detectadas dentro da rede. O tamanho do vetor é equivalente ao número de comunidades encontradas, sendo este obtido pela função *length(vector)*.

Desvio Padrão: *sd(vector, na.rm = FALSE)*

A função *sd* retorna o desvio padrão calculado com base nos valores contidos no vetor passado por parâmetro. Para obtenção do desvio padrão o vetor considerado foi o de ocorrências das ligações dos vértices, sendo este retornado pela função *degree(graph, v=V(graph), normalized=TRUE)*, onde o paramêtro *normalized* impõe a normalização dos valores (divisão da ligação de cada vértice pelo número total de vértices subtraído de uma unidade).

Máximo: *which.max(vector)*

A função *which.max* retorna o índice da posição onde se encontra o maior valor encontrado dentro do vetor passado por paramêtro. Para obtenção do valor máximo, o vetor considerado foi o de ocorrências das ligações dos vértices, sendo este retornado pela função *degree(graph, v=V(graph), normalized=TRUE)*.

Mínimo: *which.min(vector)*

A função *which.min* retorna o índice da posição onde se encontra o menor valor encontrado dentro do vetor passado por parâmetro. Para obtenção do valor mínimo, o vetor considerado também foi o de ocorrências das ligações dos vértices, sendo este retornado pela função *degree(graph, v=V(graph), normalized=TRUE)*.

No geral, para cada sequência são geradas as redes a partir das seguintes combinações de parâmetros:

- 1 rede de nucleotídeos (TP = 1; P = 1)
- 2 redes de dinucleotídeos ((TP = 2; P = 1) e (TP = 2; P = 2))
- 3 redes de trinucleotídeos ((TP = 3; P = 1), (TP = 3; P = 2) e (TP = 3; P = 3))

O vetor de características dessa metodologia será constituído pelas medidas extraídas das redes, as quais são elas as mesmas citadas anteriormente (Caminho mínimo médio, Coeficiente de Cluster, Centralidade, *Motifs*, Número de comunidades, desvio padrão, mínimo e máximo do grau dos vértices).

3.3.3 CONFIGURAÇÃO DO AMBIENTE DE EXECUÇÃO

Para execução dos experimentos foi utilizado um hardware com a seguinte especificação:

- Modelo do Processador: Intel(R) Core(TM) i7-3820 CPU @ 3.60GHz
- Quantidade de Núcleos de Processamento: 8
- Memória *Cache*: 10240Kb
- Memória Principal: 16Gb
- Sistema Operacional: Distribuição Linux Ubuntu 12.10

4 RESULTADOS

Nesse capítulo são apresentado os resultado obtidos na classificação utilizando as características extraídas com as metodologias propostas.

Para validação dos resultados foi aplicada a técnica de validação cruzada com o valor de $K = 10$ (Ten-Fold Cross-Validation). Os classificadores utilizados para as duas metodologias apresentadas foram os já citados anteriormente: *Naive Bayes*, IBK, *MultiLayer Perceptron*, SVM, J48 e *RandomForest*.

Para a maioria dos classificadores os parâmetros de execução foram os valores padrões. No caso do SVM, o parâmetro *KernelType* foi alterado para linear ao invés de radial, isso devido a precisão imposta na análise das características, onde o linear é mais flexível e menos minucioso na classificação das medidas, ou seja, se o conjunto de características forem ótimas descritoras de sua classe, então o radial é mais recomendado.

Todos atributos (medidas) foram considerados na tarefa classificação por todos os classificadores.

Os dados como já descrito anteriormente, foram extraídos das bases da dados descritas no capítulo 3, sendo os dados separados em 3 classes de diferentes regiões encontradas no DNA humano: *CDS*, *Intergenic* e *Hspromoter*.

4.1 HISTOGRAMAS

Nesta metodologia como dito anteriormente, cada sequência gera 3 vetores de características, aos quais são equivalentes ao número de experimentos realizados individualmente com as medidas extraídas das sequências genômicas, como mostrado na Figura 19:

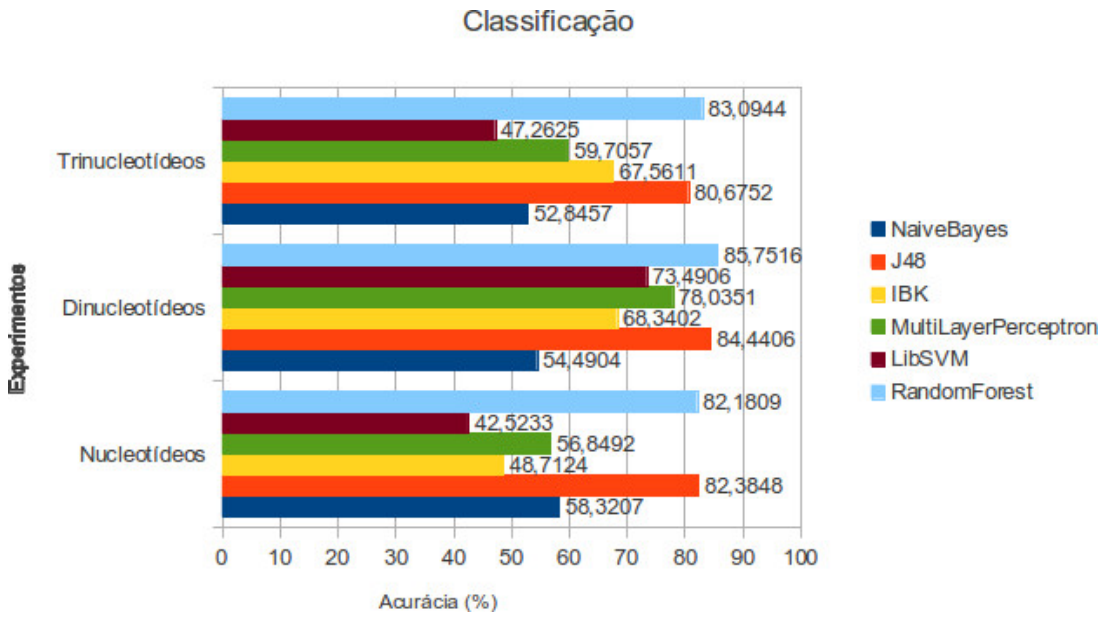


Figura 18: Acurácia da Classificação - Histogramas

A Figura 19 demonstra o percentual total de acerto na classificação para cada classificador nos três experimentos. Olhando a figura, fica claro que dentre os classificadores utilizados os que obtiveram melhores resultados foram o *RandomForest* e *J48* consecutivamente.

Os resultados como um todo foram bons nesta metodologia. Dentre os três experimentos realizados, as características que melhor obtiveram resultados descrevendo suas classes foram as obtidas com dinucleotídeos onde a acurácia chegou a quase 86% com o *RandomForest.*

4.2 REDES COMPLEXAS

Devido a variedade de medidas extraídas com cada rede gerada pela relação de parâmetros propostos para a metodologia, alguns experimentos foram realizados afim de verificar a reação dos mesmos frente a classificação, como mostrado abaixo.

4.2.1 EXPERIMENTOS

Individual

Nesse experimento a classificação foi realizada uma para cada rede gerada.

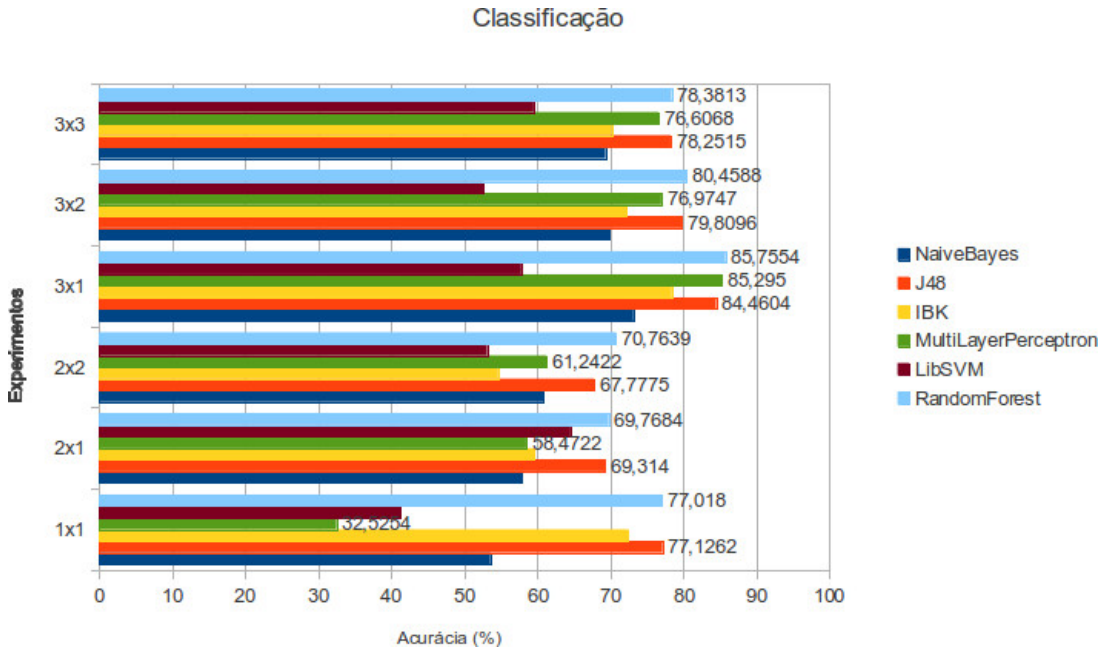


Figura 19: Acurácia da Classificação - Redes Complexas - Individual

Observa-se na Figura 20 que os melhores classificadores foram novamente *RandomForest* e *J48*, onde o *RandomForest* no experimento realizado com as medidas extraídas pela rede de trinucleotídeos com tamanho da palavra igual a 3 e o passo igual 1, obteve quase 86% de acerto, superando por pouco o melhor resultado obtido na metodologia anterior.

Palavras iguais

Neste experimento, foram unificados em um único vetor de características as medidas ao qual possuem a rede formada pelo mesmo tamanho da palavra.

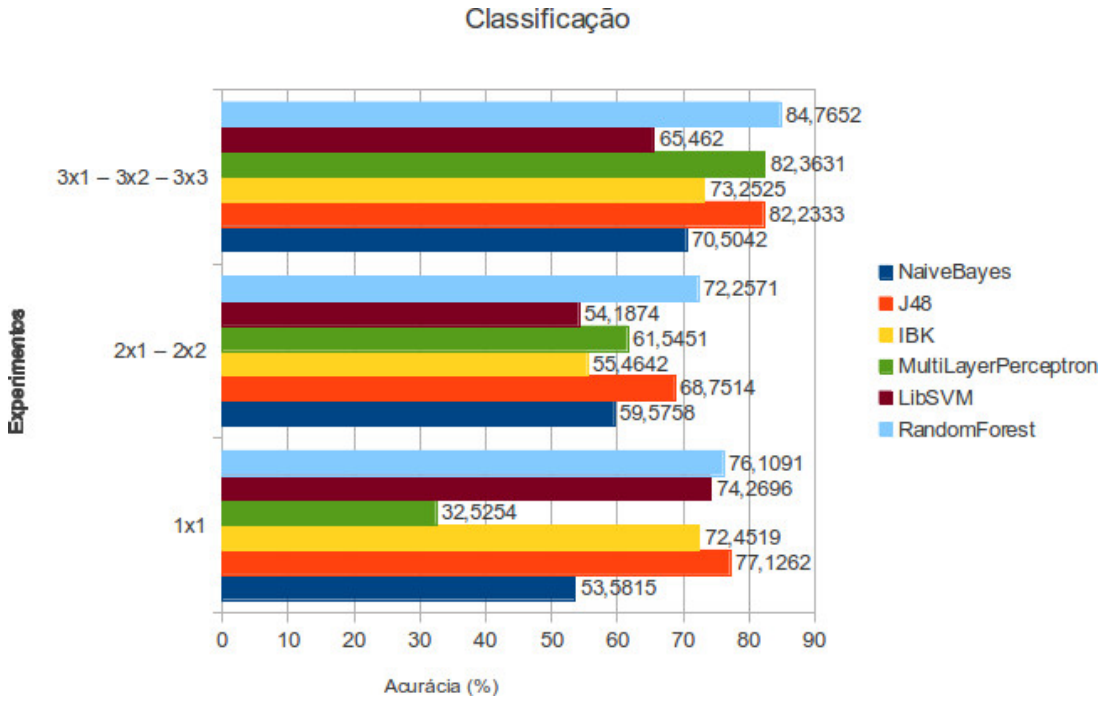


Figura 20: Acurácia da Classificação - Redes Complexas - Palavras Iguais

A Figura 21 demonstra que o melhor resultado obtido foi no caso de trinucleotídeos pelo classificador *RandomForest*, ficando bem proxímo novamente do *j48* e do *MultiLayerPerceptron*.

Todos juntos

Para esse experimento todas as características extraídas de todas as redes foram concatenadas em um único vetor, onde o melhor resultado mais uma vez foi obtido com o classificador *RandomForest*, tendo novamente o *J48* e *MultiLayerPerceptron* logo em seguida, como mostrado na Figura 22.

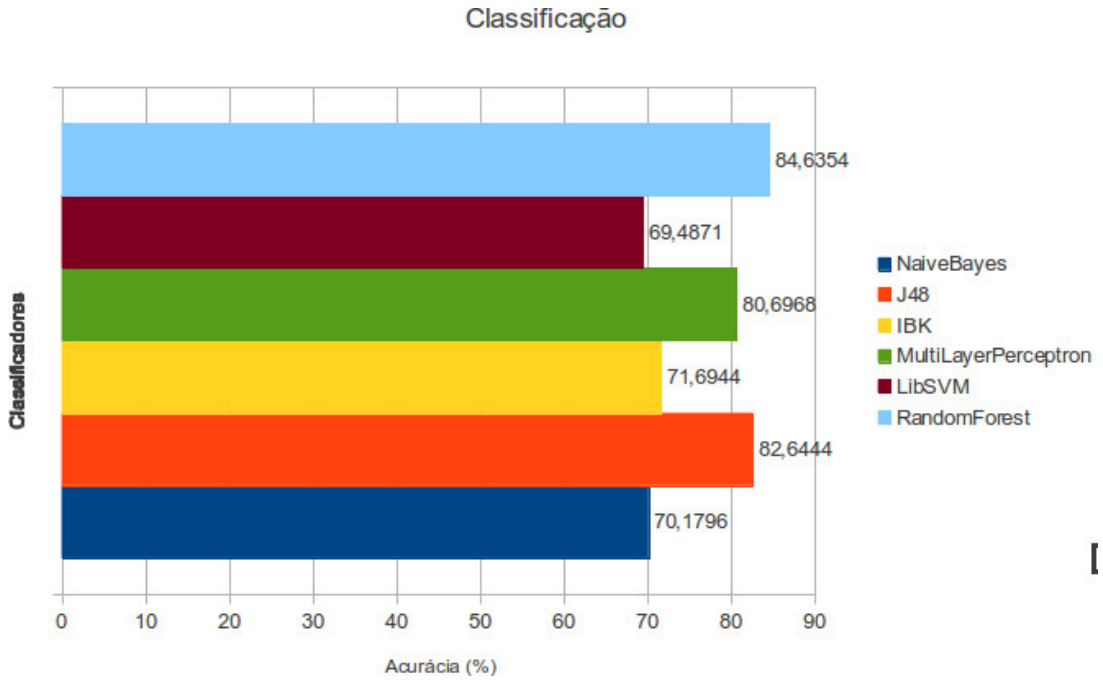


Figura 21: Acurácia da Classificação - Redes Complexas - Todos Juntos

Alguns trabalhos semelhantes que utilizaram da classificação e do método *cross-validation* para tratar outras abordagens mas no mesmo contexto de reconhecimento de padrões em sequências genômicas podem ser encontrados. Como no caso de (SUN; FAN; LI, 2003), onde o mesmo propõe um método para predizer regiões de *exons* (CDS) e *introns* (*Intergenic*), fazendo uso do *Support Vector Machines* (SVM) com valor de $k = 3$ na validação cruzada, chegando a alcançar 92,68% e 93,80% de acurácia em sequências de primatas e roedores. Um outro exemplo que utiliza esta mesma linha de pensamento é o trabalho realizado por (KASHIWABARA et al., 2007), onde os autores fazem uso da validação cruzada com valor de $k = 10$ na classificação para obtenção dos resultados de sua metodologia para o problema proposto.

De um modo geral, os resultados entre as duas metodologias foram semelhantes e bons. Porém as redes complexas obteve melhores resultados na maioria dos experimentos. A caracterização das sequências em ambas as metodologias está diretamente relacionada a frequência da ocorrência dos nucleotídeos dentro das mesmas. Deste modo, regiões intergênicas cujo apresentam uma maior repetição de nucleotídeos iguais seguidos, pode acabar gerando uma particularidade no sentido de vir a ser um padrão identificado dentro de sequências desse gênero, dificultando a inferência correta das outras classes, já que por ser majoritária desse padrão, se uma amostra de outra classe possuí característica semelhante ela será inferida como um falso positivo da classe intergênica.

5 CRONOGRAMA REALIZADO

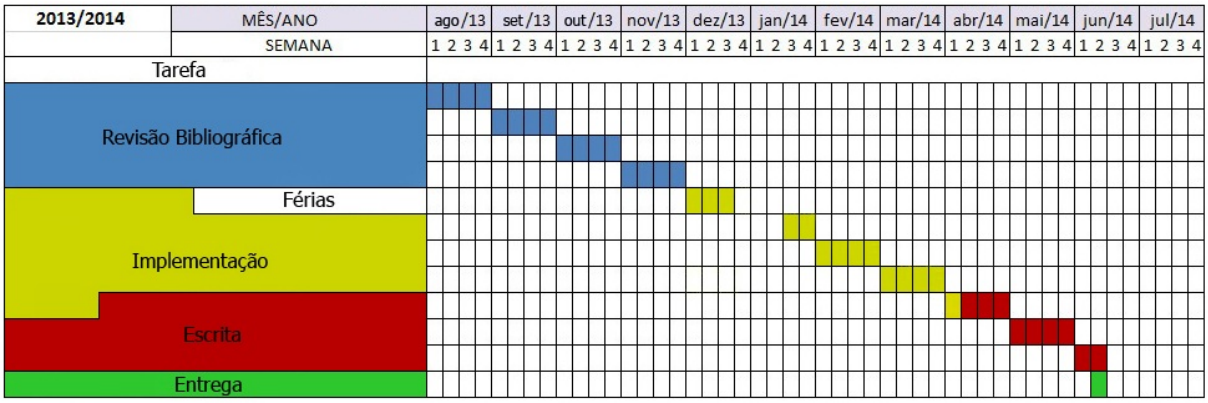


Figura 22: Cronograma de Atividades.

Fonte: Autoria Própria.

6 CONSIDERAÇÕES

O trabalho como um todo teve seus objetivos alcançados. Dois extratores de características de sequências genômicas foram implementados e avaliados.

As classificações de ambas as metodologias obtiveram bons resultados, principalmente com os classificadores *RandomForest*, *J48* e *MultiLayerPerceptron*, dando destaque ao *RandomForest*, onde o mesmo apresentou melhor resultado em quase todos experimentos, alcançando com maior acurácia o valor de 85,7554 % na classificação, como mostrado na Figura 20.

Os resultados obtidos indicam que através dessa abordagem de extração de características é possível alcançar bons níveis de classificação considerando a simplicidade dos métodos uma vez que são utilizadas somente as sequências genômicas sem nenhum outro conhecimento acerca delas.

A caracterização das sequências genômicas nas duas metodologias estão diretamente relacionada com a frequência de ocorrência dos nucleotídeos dentro das sequências. Assim, regiões intergênicas que por ter uma maior repetição concentrada de nucleotídeos iguais, possibilita a identificação dessas sequências tornando este um padrão da mesma, induzindo a inferência correta das outras classes.

Se tratando da flexibilidade das metodologias propostas, as redes complexas fornecem a possibilidade da aplicação de valores maiores nos parâmetros TP e P, que como resultado teria um número maior de redes e características extraídas, o que pode melhorar a caracterização da sequência. Além disso a metodologia proposta pode ser aplicada para a classificação de outras classes de sequências genômicas, tais como: miRNA, transposição, genes codificantes de proteínas, genes de RNA, sequências regulatórias e muitas outras.

Afim de avaliar a unificação das duas metodologias, um novo experimento foi realizado. As medidas das duas metodologias foram unificadas em um único vetor de característica e utilizadas na classificação, onde mais uma vez o classificador RandomForest apresentou melhor resultado, alcançando 92% de acurácia. Devido ao resultado obtido, um pôster foi submetido e aprovado no Brazilian Symposium on Bioinformatics (BSB) que irá correr em Belo Horizonte

no mês de Outubro deste ano. Além disso, um artigo científico resultante desse trabalho foi redigido e submetido ao periódico BMC Bioinformatics, fator de impacto 2.67 e estrato A1 no qualis da área de ciência da computação, ao qual está sob revisão.

REFERÊNCIAS

ABE, S. **Support vector machines for pattern classification**. [S.l.]: Springer, 2010.

ADAMIC, L. A.; HUBERMAN, B. A. Power-law distribution of the world wide web. **Science**, American Association for the Advancement of Science, v. 287, n. 5461, p. 2115–2115, 2000.

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. **Machine learning**, Springer, v. 6, n. 1, p. 37–66, 1991.

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. **Reviews of modern physics**, APS, v. 74, n. 1, p. 47, 2002.

ALBERT, R.; JEONG, H.; BARABÁSI, A.-L. Internet: Diameter of the world-wide web. **Nature**, Nature Publishing Group, v. 401, n. 6749, p. 130–131, 1999.

BACKES, A. R.; CASANOVA, D.; BRUNO, O. M. A complex network-based approach for texture analysis. In: **Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications**. [S.l.]: Springer, 2010. p. 354–361.

BAEZA-YATES, R.; RIBEIRO-NETO, B. et al. **Modern information retrieval**. [S.l.]: ACM press New York, 1999.

BARABÁSI, A. **Linked: how everything is connected to everything else and what it means for business, science, and everyday life**. Plume, 2003. (Plume book). ISBN 9780452284395. Disponível em: <http://books.google.com.br/books?id=rydKGwfs3UAC>.

BARABÁSI, A.-L. Linked: How everything is connected to everything else and what it means. **Plume Editors**, 2002.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999.

BARABÁSI, A.-L.; ALBERT, R.; JEONG, H. Scale-free characteristics of random networks: the topology of the world-wide web. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 281, n. 1, p. 69–77, 2000.

BARRAT, A. et al. The architecture of complex weighted networks. **Proceedings of the National Academy of Sciences of the United States of America**, National Acad Sciences, v. 101, n. 11, p. 3747–3752, 2004.

BISHOP, C. M. **Neural networks for pattern recognition**. [S.l.]: Oxford university press, 1995.

BOCCALETTI, S. et al. Complex networks: Structure and dynamics. **Physics reports**, Elsevier, v. 424, n. 4, p. 175–308, 2006.

BOLLOBÁS, B. **Modern graph theory**. [S.l.]: Springer, 1998.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BUTTE, A. J.; KOHANE, I. S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In: **Pac Symp Biocomput**. [S.l.: s.n.], 2000. v. 5, p. 418–429.

COSTA, L. d. F.; RODRIGUES, F. A.; TRAVIESO, G. Protein domain connectivity and essentiality. **Applied physics letters**, AIP Publishing, v. 89, n. 17, p. 174101, 2006.

COSTA, L. d. F. et al. Characterization of complex networks: A survey of measurements. **Advances in Physics**, Taylor & Francis, v. 56, n. 1, p. 167–242, 2007.

COSTA, L. d. F.; SPORNS, O. Hierarchical features of large-scale cortical connectivity. **The European Physical Journal B-Condensed Matter and Complex Systems**, Springer, v. 48, n. 4, p. 567–573, 2005.

COVENEY, P. V.; HIGHFIELD, R. The arrow of time: A voyage through science to solve time's greatest mystery. **New York: Fawcett Columbine, 1991. 1st American ed.**, v. 1, 1991.

CSARDI, G.; NEPUSZ, T. The igraph software package for complex network research. **InterJournal, Complex Systems**, v. 1695, n. 5, 2006.

DANON, L. et al. Comparing community structure identification. **Journal of Statistical Mechanics: Theory and Experiment**, IOP Publishing, v. 2005, n. 09, p. P09008, 2005.

DENNIS, C.; MARSLAND, D.; COCKETT, T. Data mining for shopping centres–customer knowledge-management framework. **Journal of Knowledge Management**, MCB UP Ltd, v. 5, n. 4, p. 368–374, 2001.

DIAMBRA, L.; COSTA, L. d. F. Complex networks approach to gene expression driven phenotype imaging. **Bioinformatics**, Oxford Univ Press, v. 21, n. 20, p. 3846–3851, 2005.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. [S.l.]: John Wiley & Sons, 2012.

ERDŐS, P.; RÉNYI, A. On random graphs. **Publicationes Mathematicae Debrecen**, v. 6, p. 290–297, 1959.

FAYYAD, U. M. et al. Knowledge discovery and data mining: Towards a unifying framework. In: **KDD**. [S.l.: s.n.], 1996. v. 96, p. 82–88.

FREEMAN, L. C. A set of measures of centrality based on betweenness. **Sociometry**, JSTOR, p. 35–41, 1977.

GRIFFITHS, A. **Introdução à genética**. Guanabara Koogan, 2008. ISBN 9788527714976. Disponível em: <http://books.google.com.br/books?id=c0vjPgAACAAJ>.

GUIMERA, R.; AMARAL, L. A. N. Modeling the world-wide airport network. **The European Physical Journal B-Condensed Matter and Complex Systems**, Springer, v. 38, n. 2, p. 381–385, 2004.

GUIMERÀ, R. et al. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 102, n. 22, p. 7794–7799, 2005.

HALL, M. et al. The weka data mining software: An update. **SIGKDD Explor. Newsl.**, ACM, New York, NY, USA, v. 11, n. 1, p. 10–18, nov. 2009. ISSN 1931-0145. Disponível em: <http://doi.acm.org/10.1145/1656274.1656278>.

JAYNES, E. T. Information theory and statistical mechanics. **Physical review**, APS, v. 106, n. 4, p. 620, 1957.

JEONG, H. et al. Lethality and centrality in protein networks. **Nature**, Nature Publishing Group, v. 411, n. 6833, p. 41–42, 2001.

JEONG, H. et al. The large-scale organization of metabolic networks. **Nature**, Nature Publishing Group, v. 407, n. 6804, p. 651–654, 2000.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: MORGAN KAUFMANN PUBLISHERS INC. **Proceedings of the Eleventh conference on Uncertainty in artificial intelligence**. [S.l.], 1995. p. 338–345.

KASHIWABARA, A. et al. Splice site prediction using stochastic regular grammars. **Genetics and Molecular Research**, v. 6, p. 105–115, 2007.

KENT, W. J. et al. The human genome browser at ucsc. **Genome research**, Cold Spring Harbor Lab, v. 12, n. 6, p. 996–1006, 2002.

KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: **IJCAI**. [S.l.: s.n.], 1995. v. 14, n. 2, p. 1137–1145.

LOPES, F. M. **Um modelo perceptivo de limiarização de imagens digitais**. Dissertação (Mestrado) — Universidade Federal do Paraná, UFPR, 2003.

LOPES, F. M. et al. A feature selection technique for inference of graphs from their known topological properties: Revealing scale-free gene regulatory networks. **Information Sciences**, Elsevier, v. 272, p. 1–15, 2014.

LOPES, F. M.; JR, R. M. C.; COSTA, L. D. F. Gene expression complex networks: synthesis, identification, and analysis. **Journal of Computational Biology**, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 18, n. 10, p. 1353–1367, 2011.

LOPES, F. M.; MARTINS, D. C.; CESAR, R. M. Feature selection environment for genomic applications. **BMC bioinformatics**, BioMed Central Ltd, v. 9, n. 1, p. 451, 2008.

LOPES, F. M.; OLIVEIRA, E. A. de; CESAR, R. M. Inference of gene regulatory networks from time series by tsallis entropy. **BMC systems biology**, BioMed Central Ltd, v. 5, n. 1, p. 61, 2011.

LOPES, F. M. et al. Entropic biological score: a cell cycle investigation for grns inference. **Gene**, Elsevier, v. 541, n. 2, p. 129–137, 2014.

MARKEL, S.; LEON, D. **Sequence Analysis in a Nutshell: A Guide to Tools: A Guide to Common Tools and Databases**. O'Reilly Media, Incorporated, 2003. (In a Nutshell Series). ISBN 9780596004941. Disponível em: <http://books.google.com.br/books?id=GEr_cFsB62MC>.

MILGRAM, S. The small world problem. **Psychology today**, New York, v. 2, n. 1, p. 60–67, 1967.

MILO, R. et al. Network motifs: simple building blocks of complex networks. **Science**, American Association for the Advancement of Science, v. 298, n. 5594, p. 824–827, 2002.

NEWMAN, M. E. The structure and function of complex networks. **SIAM review**, SIAM, v. 45, n. 2, p. 167–256, 2003.

PROSDOCIMI, F. et al. Bioinformática: manual do usuário. 2012.

QUINLAN, J. R. **C4. 5: programs for machine learning**. [S.l.]: Morgan kaufmann, 1993.

REZENDE, S. O. et al. Mineração de dados. **Sistemas inteligentes: fundamentos e aplicações**, v. 1, p. 307–335, 2003.

RUBINOV, M.; SPORNS, O. Complex network measures of brain connectivity: uses and interpretations. **Neuroimage**, Elsevier, v. 52, n. 3, p. 1059–1069, 2010.

RUSSELL, S. J. et al. **Artificial intelligence: a modern approach**. [S.l.]: Prentice hall Englewood Cliffs, 1995.

SHANNON, C. E. A mathematical theory of communication. **ACM SIGMOBILE Mobile Computing and Communications Review**, ACM, v. 5, n. 1, p. 3–55, 2001.

SHEN-ORR, S. S. et al. Network motifs in the transcriptional regulation network of escherichia coli. **Nature genetics**, Nature Publishing Group, v. 31, n. 1, p. 64–68, 2002.

STROGATZ, S. H. Exploring complex networks. **Nature**, Nature Publishing Group, v. 410, n. 6825, p. 268–276, 2001.

SUN, Y.-F.; FAN, X.-D.; LI, Y.-D. Identifying splicing sites in eukaryotic rna: support vector machine approach. **Computers in biology and medicine**, Elsevier, v. 33, n. 1, p. 17–29, 2003.

TEAM, R. C. et al. R: A language and environment for statistical computing. **R foundation for Statistical Computing**, sn, 2005.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**. Elsevier Science, 2008. ISBN 9780080949123. Disponível em: <http://books.google.com.br/books?id=QgD-3Tcj8DkC>.

VOGELSTEIN, B.; LANE, D.; LEVINE, A. J. Surfing the p53 network. **Nature**, Nature Publishing Group, v. 408, n. 6810, p. 307–310, 2000.

WASSERMAN, S. **Social network analysis: Methods and applications**. [S.l.]: Cambridge university press, 1994.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. **nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998.

WEISS, S. M. **Predictive data mining: a practical guide**. [S.l.]: Morgan Kaufmann, 1998.

YAMASHITA, R. et al. Dbtss: Database of transcriptional start sites progress report in 2012. **Nucleic acids research**, Oxford Univ Press, v. 40, n. D1, p. D150–D154, 2012.