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

Relatório Final de Atividades

Extração de características a partir de redes complexas: Um estudo de caso na análise de sequências genômicas

Vinculado ao projeto
Uma abordagem baseada em redes complexas para análise genômica

Bruno Mendes Moro Conque
Bolsista CNPq
Tecnologia em análise e desenvolvimento de sistemas da informação
Prof. Dr. Fabrício Martins Lopes

Área do Conhecimento: Sistemas de computação
CAMPUS CORNÉLIO PROCÓPIO, 2014

BRUNO MENDES MORO CONQUE
FABRÍCIO MARTINS LOPES

RELATÓRIO FINAL DE ATIVIDADES PIBIC

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

CORNÉLIO PROCÓPIO, 2014

SUMÁRIO

1 INTRODUÇÃO 2
2 REVISÃO BIBLIOGRÁFICA 3
2.1 GENÉTICA 3
2.2 REDES COMPLEXAS 3
2.2.1 MEDIDAS 4
2.3 MINERAÇÃO DE DADOS 6
2.3.1 VALIDAÇÃO CRUZADA 7
3 MATERIAIS E MÉTODOS 8
3.1 BANCO DE DADOS DBTSS 8
3.2 GENOME BROWSER 8
3.3 METODOLOGIA 8
3.3.1 REDES COMPLEXAS 8
4 RESULTADOS 11
4.1 REDES COMPLEXAS 11
4.1.1 EXPERIMENTOS 11
5 CONCLUSÕES 15
REFERÊNCIAS 16

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 principalanalisar o organismo em uma forma holística.

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, caracterizar 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).

Frente a essa perspectiva, esse trabalho tem o intuito de utilizar das redes complexas como um método para extração de características de sequências genômicas, propondo um estudo no sentido de que se os resultados forem coerentes e satisfatórios, vir a ser mais um método para a caracterização de regiões encontradas no DNA de diferentes organismos.

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. (ALTMAN; JIMENEZ, 2012) 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 REDES COMPLEXAS

Considerado uma extensão da teoria dos grafos, as redes complexas é definida como um grafo que mostra uma estrutura irregular dos vértices (nós) ligados por arestas (COSTA et al.,

2007). A teoria de redes complexas estende o formalismo da teoria dos grafos por adição de medidas e métodos baseados nas propriedades reais de um sistema (COSTA et al., 2007). Esta teoria apresenta aplicações multidisciplinares, abrangendo várias ciências, tais como biologia, ciência da computação, física, matemática, sociologia entre outras. Deste modo, diversos sistemas do mundo real pode ser representado por meio de redes complexas, como a ligação entre os aeroportos (GUIMERA; AMARAL, 2004; GUIMERÀ et al., 2005), a Internet (ALBERT; JEONG; BARABÁSI, 1999), redes sociais (WASSERMAN, 1994), redes neurais (COSTA; SPORNS, 2005; RUBINOV; SPORNS, 2010) e sistemas biológicos (LOPES et al., 2014; LOPES; JR; COSTA, 2011; JEONG et al., 2000).

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.

2.2.1 MEDIDAS

Dentre as medidas possíveis de se extrair de uma rede complexa, 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:

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

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:

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

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.

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).

2.3 MINERAÇÃO DE DADOS

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 (HALL et al., 2009), 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).

2.3.1 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 é apresentado a metodologia utilizada para realizar a caracterização das sequências genômicas.

3.3.1 REDES COMPLEXAS

Rotinas em JAVA foram implementadas 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 1: 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.

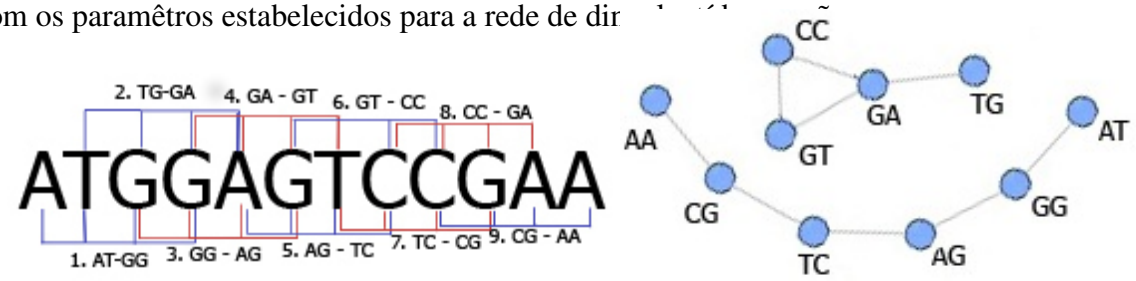
Sendo assim, dado uma arquivo fasta cada sequência identificada dará origem a 6 redes complexas não direcionais. O número de redes estipulado (6) é de acordo com os parâmetros

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 que é o numero de caracteres que representará os vértices e P que 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.

Para uma melhor ilustração, um exemplo de uma rede de dinucletideos com parâmetros TP = 2 e P 1 é demonstrado na Figura **??** e Figura 2. Considerando a sequência 'ATGGAGTCCGAA', a Figura 1 demonstra como são formados os vértices baseado nos parâmetros estabelecidos. A Figura 2 demonstra a rede resultante das ligações encontradas.

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

Considerando a seguinte sequência 'ATGGAGTCCGAA', as ligações encontradas de acordo com os paramêtros estabelecidos para a rede de dir



(a) (b)

Figura 1: A aplicação da metodologia proposta para uma rede dinucleotídeos considerando a sequência ATGGAGTCCGAA com parâmetros $p = 1$ e $WS = 2$. (a) é como as arestas e vértices são definidos e (b) representa a rede resultante.

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

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)) e 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).

4 RESULTADOS

Nesse capítulo são apresentado os resultado obtidos na classificação utilizando as características extraídas com a metodologia proposta.

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: *CDS*, *Intergenic* e *Hspromoter*.

4.1 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.

4.1.1 EXPERIMENTOS

Individual

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

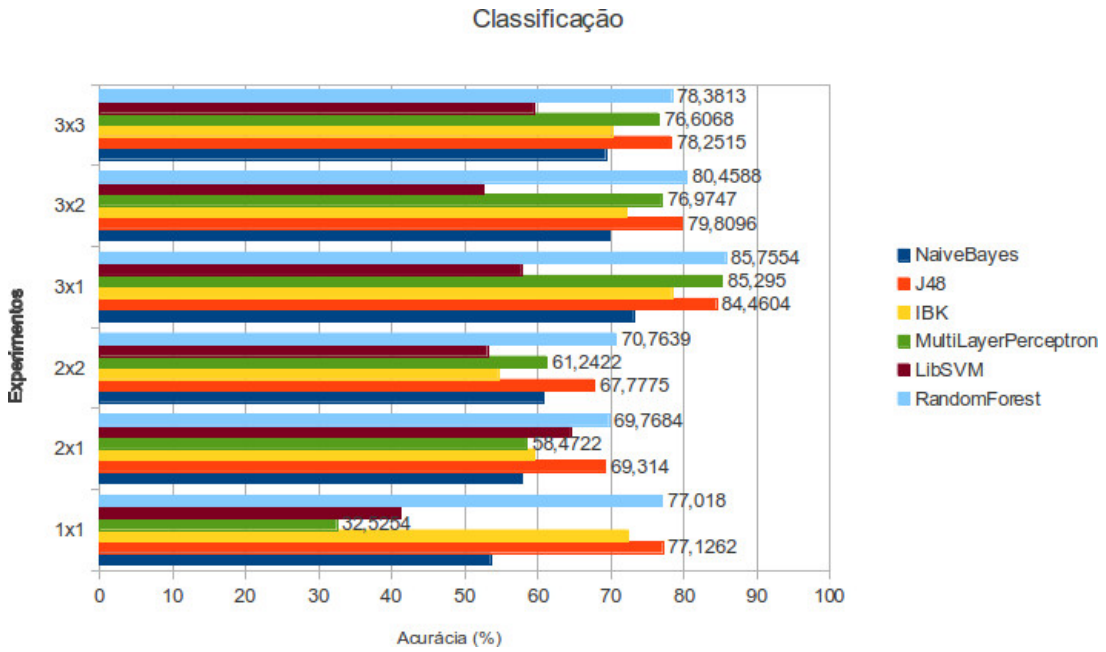


Figura 2: Acurácia da Classificação - Redes Complexas - Individual

Observa-se na Figura 2 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.

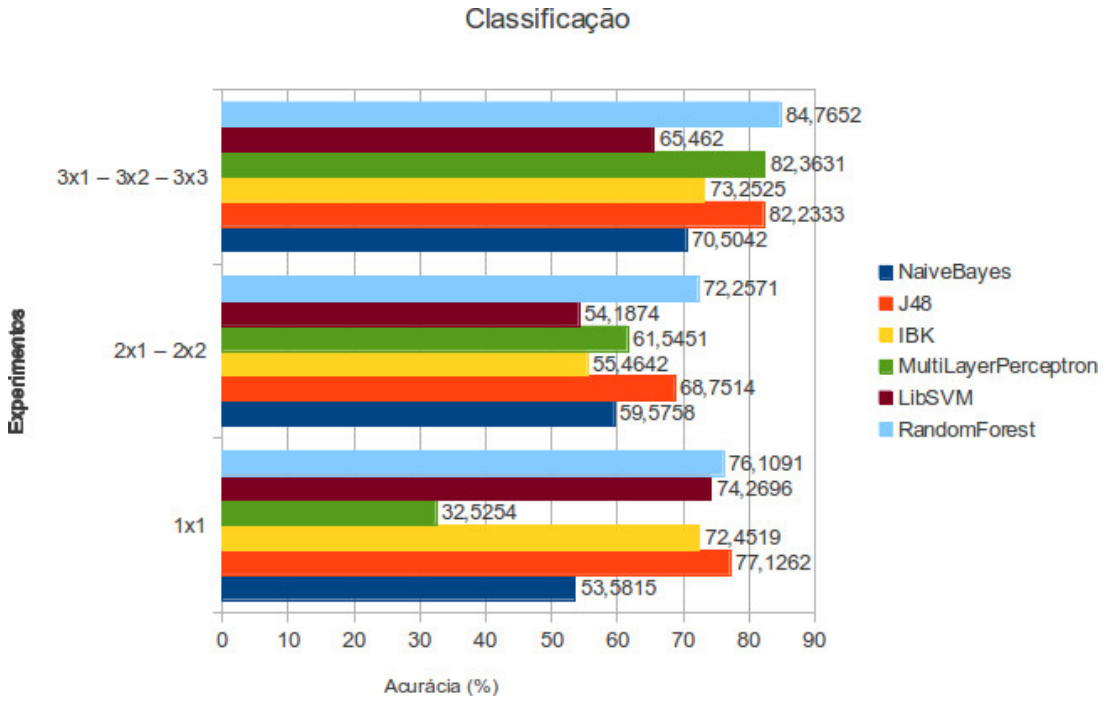


Figura 3: Acurácia da Classificação - Redes Complexas - Palavras Iguais

A Figura 3 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 4.

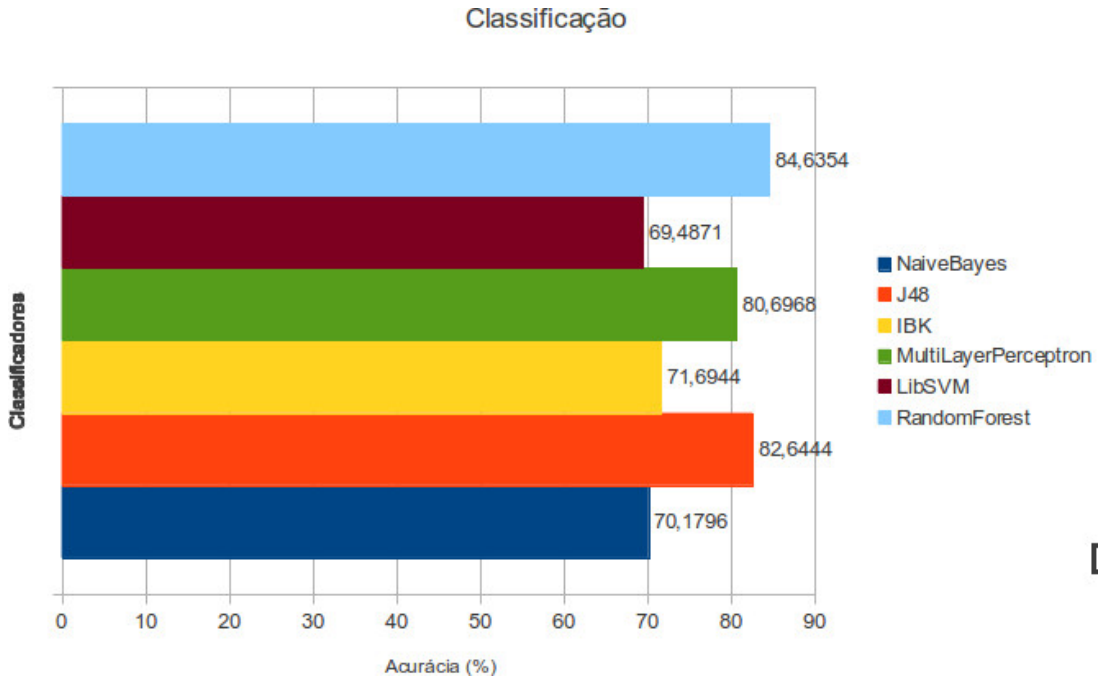


Figura 4: Acurácia da Classificação - Redes Complexas - Todos Juntos

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.

A caracterização das sequências genômicas esta 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.

5 CONCLUSÕES

O trabalho como um todo teve seu objetivo alcançado. Um extrator de características de sequências genômicas foi implementado e avaliado.

A classificação com as medidas extraídas pela metodologia aqui proposta obteve 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 2.

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

Além dos resultados obtidos, as redes complexas proporciona ainda a possibilidade de novos experimentos devido a flexibilidade do algoritmo para entradas maiores em relação ao tamanho da palavra e do passo, criando como consequência um número maior de redes, retratando uma aferição maior das ligações entre os nucleotídeos dentro das sequências podendo assim melhorar a caracterização das mesmas. 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 aplicação para identificação de espécies afim de construir arvorés filogenéticas.

REFERÊNCIAS

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

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

ALTMAN, E.; JIMENEZ, T. Ns simulator for beginners. **Synthesis Lectures on Communication Networks**, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–184, 2012.

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.

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

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

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.

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.

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.

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>.

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. 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.

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.

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

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

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.

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.

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.

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.