IDENTIFICAÇÃO AUTOMÁTICA DE REGIÕES EXONS E INTRONS EM SEQUÊNCIAS DE DNA

PALMA, Wallace Pannace[1]
ZALEWSKI, Willian[2]

RESUMO
O estudo do sequenciamento de DNA possibilita o desenvolvimento de novos métodos para diagnosticar doenças e a formulação de novos medicamentos. Os genes codificadores de proteínas constituem a parte funcional do DNA, onde estão as instruções biológicas básicas para a expressão gênica (produção de proteínas), são compostos por segmentos introns, sequências não codificadoras, e exons, sequências que serão traduzidas para proteínas. Os avanços na automação do sequenciamento de DNA possibilitaram a criação de grandes quantidades de dados, gerando uma alta demanda por métodos de análise mais eficientes. Técnicas computacionais como aprendizado de máquina tem sido amplamente utilizadas, porém o desempenho geral destes métodos ainda não é considerado satisfatório. Neste trabalho utilizamos algoritmos de seleção de atributos para identificar os nucleotídeos que melhor identificam a subsequência com o intuito de melhorar o desempenho de técnicas de aprendizado de máquina para a identificação de regiões introns e exons.
Palavras-chaves: aprendizado de máquina, dna, classificação, sequências.

1 INTRODUÇÃO

Os avanços na tecnologia e na automação do sequenciamento de DNA tornaram possível a geração de uma grande quantidade de dados de sequências de DNA. Esse grande crescimento de dados gerou uma demanda significativa por métodos de análise mais eficientes, tais como a utilização de técnicas computacionais. A localização precisa das junções de introns e exons e a compreensão da estrutura do gene nestas grandes bases de dados é um dos principais temas de estudo na área de bioinformática, pois possibilita o desenvolvimento de novos métodos para diagnosticar doenças e a formulação de novos medicamentos.

Diversos métodos computacionais foram propostos na última década para a identificação de genes eucarióticos. Em especial, abordagens que utilizam técnicas de aprendizado de máquina vem sendo exploradas para identificação de padrões

[1] Discente do curso de Engenharia Física do – ILACVN – UNILA; voluntário IC-UNILA E-mail: wpd.palma.2016@aluno.unila.edu.br.
[2] Doutor – ILATIT – UNILA. Orientador de voluntário IC-UNILA; E-mail: willian.zalewski@unila.edu.br.

em sequências de DNA. Contudo, o desempenho geral desses algoritmos para a predição de genes ainda não é considerado satisfatório.

Nesse contexto, neste trabalho nosso objetivo consiste em aplicar algoritmos de seleção de atributos para melhorar o desempenho dos algoritmos de aprendizado de máquina. Desse modo, desenvolvemos e avaliamos modelos de predição utilizando os algoritmos de *machine learning* em combinação com a seleção de atributos para a identificação automática de pontos de junção exon-intron e intron-exon em genes.

2 FUNDAMENTAÇÃO TEÓRICA

Pontos de Junção: O genoma humano contém aproximadamente 23500 genes codificadores de nucleotídeos [3]. O processo de decodificação de proteínas a partir da informação contida nos genes é denominado expressão gênica. Os genes são compostos alternadamente de segmentos denominados introns e exons. Os exons são as regiões do gene que serão traduzidas para proteínas; os introns são as regiões não codificadoras. Na Figura 1 é apresentada uma representação esquemática do processo de decodificação de proteínas. A compreensão da expressão gênica só pode ser atingida a partir da correta identificação das áreas de junção nas sequências correspondentes aos genes e do tipo de junção (exon-intron ou intron-exon) [5].

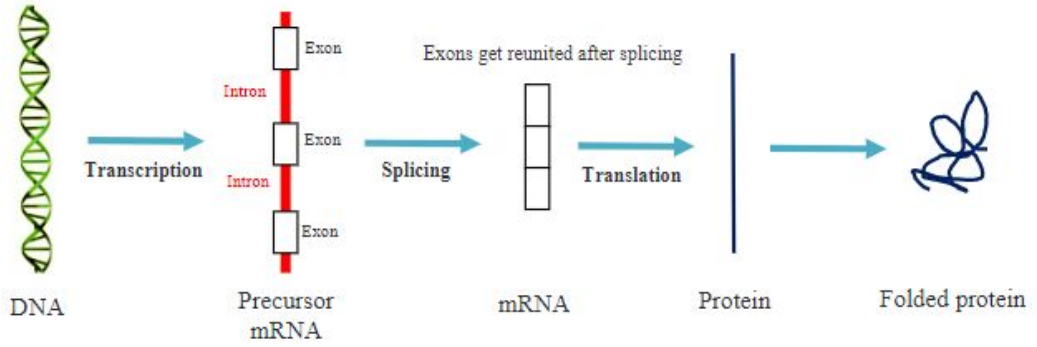

Figura 1: Síntese de RNA e tradução em proteína.

Aprendizado de Máquina: a aquisição de conhecimento de maneira automática através de algoritmos de aprendizado de máquina (AM) é baseada na inferência indutiva, na qual, os conhecimentos podem ser derivados de outros previamente conhecidos. Desse modo, induz-se uma hipótese a partir de um

conjunto de exemplos (amostra) que permitam caracterizar o domínio (população) que se deseja tratar. De acordo com o tipo de conceito utilizado para induzir uma determinada hipótese, os algoritmos de AM podem ser divididos em paradigmas: Paradigma simbólico: realizam o processo de aprendizagem de um conceito utilizando representações simbólicas por meio da análise de exemplos e contra-exemplos. Geralmente representadas na forma de expressão lógica, como a árvore de decisão (DT) ou rede semântica. Paradigma baseado em exemplos: guardam os exemplos e utilizam medidas de similaridade para identificar os os casos mais similares ao exemplo a ser analisado. Por exemplo o k-vizinhos mais próximos (KNN). Paradigma estatístico: utilizam modelos estatísticos para encontrar uma aproximação do conceito induzido. Dentre estes algoritmos, destacam-se o máquinas de vetores de suporte (SVM) e o aprendizado Bayesiano (GNB). Paradigma conexionista: utilizam construções matemáticas inspiradas em conexões neuronais do sistema nervoso humano. São exemplos deste as redes neuronais artificiais (MLP) [4].

Seleção de atributos: nas últimas décadas, a dimensionalidade dos dados envolvidos em tarefas com AM e *data mining* tem crescido exponencialmente. Dados com alta dimensionalidade tem apresentado desafios significativos aos métodos de aprendizado existentes. A seleção de atributos busca escolher um subconjunto reduzido de atributos relevantes a partir do conjunto original de acordo com certo critério de avaliação de relevância. Essa abordagem tem como objetivo melhorar o desempenho do aprendizado, reduzir o custo computacional e aumentar a interpretabilidade do modelo criado. Neste trabalho, utilizamos o modelo de filtro Correlation-based Feature Selection (CFS), que leva em conta a utilidade de atributos individuais para a predição da classe junto com o nível de intercorrelação entre eles. Também utilizamos o modelo incorporado da árvore de decisão que utiliza o critério GINI [2].

3 METODOLOGIA

O método proposto neste trabalho pode ser estruturado em três etapas principais: a) Transformação dos dados; b) Seleção dos atributos importantes e c) Classificação. Na primeira etapa (a), inicialmente as sequências de DNA foram

convertidas para números, como: A para 0, C para 1, G para 2, T para 3. Por exemplo, a sequência CCAGCT passaria a ser 110213. Na segunda etapa (b) foram aplicados os algoritmos de seleção de atributos GINI e CFS, com o intuito de identificar quais os segmentos são mais relevantes. Assim, uma segunda transformação foi aplicada sobre o conjunto de dados, nesse caso considerando somente os atributos selecionados como importantes. Finalmente, na terceira etapa (c) foram aplicados os algoritmos de aprendizado de máquina DT, KNN, SVM, MLP e GNB para a classificação das sequências de DNA. Cada algoritmo de classificação foi analisado sobre cada uma das transformações geradas na etapa (b) pelos algoritmos GINI e CFS. A avaliação do desempenho de cada combinação dos algoritmos utilizados foi realizada através da técnica k-partições *cross-validation,* considerando todos os exemplos da base de dados com k = 10.

Como mencionado, neste trabalho buscamos aplicar técnicas de aprendizado de máquina (AM) em sequências de DNA para a identificação de junções intron-exon (i-e) e exon-intron (e-i). Esta abordagem se deve a grande quantidades de dados e a grande eficiência destes métodos de AM, Para nossos testes, utilizamos a base de dados *Molecular Biology (Splice-junction Gene Sequences) Data-Set,* que contém 3190 subsequências de DNA com 60 nucleotídeos cada, classificadas de acordo com o tipo: i-e, e-i ou nenhuma (N). Nesse conjunto de dados, todos os exemplos foram retirados do Genbank 64.1, e são todos partes dos genes de primatas. Para manipulação dos dados e aplicação dos algoritmos utilizamos a linguagem Python por ser prática e comumente utilizada pela comunidade científica. Para os algoritmos de aprendizado de máquina utilizamos a biblioteca scikit-learn (http://scikit-learn.org). Os algoritmos de seleção de atributos utilizados foram GINI (biblioteca *scikit-learn*) e CFS (Correlation-based Feature Selection) proveniente da biblioteca *skfeature* (http://featureselection.asu.edu). Todos os algoritmos foram utilizados na configuração padrão.

4 RESULTADOS E DISCUSSÃO

Da seleção de atributos executada pelo CFS, os atributos selecionados foram os de índices [29, 31, 28, 30, 34, 27, 32, 33, 24, 22], ou seja passando a base de 60 nucleotídeos para 10 nucleotídeos a serem considerados pelos algoritmos em cada

amostra. Adicionalmente, como pode ser observado na Tabela 1, a precisão geral do DT foi a melhor nos três testes apresentados. Também, observa-se uma melhora expressiva para os algoritmos SVM e KNN.

	Dados brutos	CFS	GINI
DT	90% +/- 5%	91% +/- 4%	92% +/- 4%
SVM	85% +/- 4%	89% +/- 3%	90% +/- 3%
KNN (K=1)	63% +/- 5%	85% +/- 4%	87% +/- 4%
GNB	90% +/- 4%	89% +/- 5%	88% +/- 4%
MLP	65% +/- 2%	54% +/- 12%	76% +/- 10%

Tabela 1: Precisão do classificador em porcentagem de acerto.

5 CONCLUSÕES

Com base nos resultados apresentados neste trabalho observou-se que, a aplicação de técnicas de seleção de atributos para identificar os nucleotídeos mais importantes possibilitou a redução no tamanho do conjunto de dados, consequentemente a criação de modelos mais rápidos; e também a melhoria no desempenho dos algoritmos de AM. Adicionalmente, é interessante ressaltar que a seleção de atributos identificou nucleotídeos em sequência, o que pode indicar importância na sequencialidade da informação, o que está de acordo com o domínio, ou seja, viabilidade na utilização de técnicas que considerem este aspecto como por exemplo técnicas de séries temporais.

6 PRINCIPAIS REFERÊNCIAS BIBLIOGRÁFICAS

[1] DUA, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science

[2] HALL, Mark A. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. Hamilton, New Zealand, 2000.

[3] MARIAN AJ. Sequencing Your Genome: What Does It Mean? Methodist DeBakey Cardiovascular Journal. 2014;10(1):3-6.

[4] ZALEWSKI, Willian. Modelagem Simbólica de Padrões Morfológicos para a Classificação de Séries Temporais. Curitiba, PR, 2015.

[5] ZAW Zaw Htike, Shoon Lei Win, 2013, 'Classification of Eukaryotic Splice-junction Genetic Sequences Using Averaged One-dependence Estimators with Subsumption Resolution', Procedia Computer Science, vol. 23, pp. 36-43.