Análise de composição de conjunto de treinamento para avaliação de aprendizagem de máquina aplicada à predição de genes

Raíssa Silva[1], Kleber Padovani[1], Wendel Santos[1], Roberto Xavier[1], Ronnie Alves[1,2]

[1]Universidade Federal do Pará (UFPA)
Programa de Pós-Graduação em Ciência da Computação
Rua Augusto Corrêa, 1, Guamá – Belém – PA – Brazil

[2]Instituto Tecnológico Vale (ITV)
Rua Boaventura da Silva, 955, Nazaré – Belém – PA – Brazil

{r.lorenna, kleber.padovani, wendelrenann, rbxjunior}@gmail.com, ronnie.alves@itv.org

Resumo. *A metagenômica realiza o estudo de comunidades microbianas, conhecidas como metagenomas, descrevendo-as por meio de suas composições e das relações e atividades dos microrganismos que ali coabitam, permitindo assim um maior conhecimento a respeito dos fundamentos da vida e da ampla — e ainda pouco conhecida — diversidade microbiológica. Uma das formas de se realizar tal descrição é por meio da análise de informações de genes contidos em (meta)genomas, extraídas através do processo de identificação de genes em sequências de DNA denominado predição de genes. Este trabalho apresenta um estudo de caso que permite a análise do impacto da composição do conjunto de treinamento ao se utilizar aprendizagem de máquina na predição de genes codificadores de proteína.*

Abstract. *Metagenomics allows the study of microbial communities, known as metagenomes, describing them through their compositions and the relation and activities of the microorganisms that coexist there, thus allowing a deeper knowledge about the fundamentals of life and about the broad microbiological diversity, which is still poorly known. Such description can be achieved by the analysis of information from genes contained in (meta) genomes, extracted through the process of identifying genes in DNA sequences, called gene prediction. This work presents a study that allows the analysis of the impact of the training set composition when using machine learning in protein-coding genes prediction.*

1. Introdução

Predição gênica é um processo crucial na área da biologia computacional. Ela consiste na interpretação de sequências genômicas por meio de algoritmos computacionais. Um dos seus objetivos é a identificação de regiões dentro do genoma que podem ser codificadas em proteínas que irão atuar em processos regulatórios no organismo. Essa identificação nos informa a localização dentro do genoma e as características do gene codificante durante esta etapa de análise funcional. Devido a isso, a predição correta dos genes permite

uma boa acurácia na anotação e caracterização do grupo funcional do organismo e das relações entre seus genes[De Filippo et al. 2012].

Dentro da predição de genes, bem como nas outras áreas da genética molecular e bioinformática, utiliza-se com bastante frequência o acrônimo *ORF*, que representa o termo *Open Reading Frame* e é atribuído às sequências genômicas iniciadas e terminadas por determinadas trincas de nucleotídeos, conhecidas como códons. O códon ATG, também conhecida como *start codon*, determina o início de uma ORF em procariotos, enquanto os códons TAG, TGA e TAA, denominadas *stop codon*, encerram a sequência de nucleotídeos da ORF correspondente[Fassetti et al. 2017].

O conceito de ORF é importante dentro da predição de genes pois toda sequência de nucleotídeos correspondente a um gene — comumente referida por *CDS*, de *Co-Ding Sequence* — corresponde a uma ORF. Sendo assim, a detecção das ORFs é um passo importante na busca por genes nos genomas, inclusive em análises com sequências (meta)genômicas altamente fragmentadas. Contudo, a identificação de ORFs não é suficiente para a identificação de genes, uma vez que, embora todo gene corresponda a uma ORF, a recíproca não é verdadeira. Ao longo do genomas dos organismos, existem diversas ORFs que não correspondem a genes[Sieber et al. 2018].

A Figura 1 ilustra um trecho de um genoma fictício de um procarioto em que podemos observar dentro dele a existência de 1 CDS e 3 ORFs, denominadas como *ORF A*, *ORF B* e *ORF C*. Por meio da ilustração, pode-se observar 3 cenários relevantes relacionados com as ORFs: 1) a existência de ORFs que não correspondem a genes — cenário ilustrado pela ORF A; 2) a existência de ORF que corresponde a CDS — representado pela ORF B; e 3) a existência de ORF dentro de outra ORF, aqui denominado sub ORF — representado pela ORF C. Dessa forma, conclui-se que, embora uma ORF tenha potencial para ser uma CDS, é possível a existência de ORFs dentro das regiões do genoma que não correspondem a CDS — chamadas de regiões intergênicas —, bem como a existência de sub ORFs tanto nas regiões intergênicas quanto nas próprias CDS[Sieber et al. 2018].

Figura 1. Ilustração de trecho do genoma contendo 3 ORFs, denominadas como ORF A, B e C, sendo uma delas — a ORF B — também uma CDS. As trincas correspondentes a *start* e *stop codons* estão destacadas em negrito e identificadas, respectivamente, nas cores verde e vermelho, enquanto as regiões intergênicas estão destacadas em azul.

Uma variedade de métodos computacionais do AM têm sido utilizados para a identificação de gene [Rho et al. 2010] [Zhu et al. 2010],[Hoff 2009] [Noguchi et al. 2008]. Embora o estado da arte seja caracterizado principalmente pelos Modelos Ocultos de Markov, outros métodos robustos de aprendizagem, como as árvores aleatórias [Breiman 2001], ainda não foram aprofundadamente explorados e aplicados no problema de predição de genes.

Nesse contexto, o objetivo deste trabalho é apresentar um estudo de caso que pretende avaliar o impacto da composição do conjunto de treinamento no desempenho da

classificação automática de CDS utilizando árvores aleatórias. O primeiro aspecto observado é a utilização complementar de sub ORFs de CDS como instâncias positivas e o segundo aspecto refere-se ao uso de estratégia de balanceamento em casos de desbalanceamento de classes no conjunto de treinamento.

2. Materiais e Métodos

Na construção do conjunto de treino, foram utilizados 20 genomas finalizados de bactérias e, para a elaboração do conjunto de teste, similarmente, foram utilizados outros 5 genomas finalizados de bactérias. O genoma completo, as CDS e as tabelas de características das CDS constituem as anotações utilizadas para cada organismo de treino e de teste, as quais foram baixadas do NCBI [1]. Essas anotações, bem como os códigos utilizados, podem ser encontrados no material suplementar, disponível online em http://sourceforge.net/p/bsb18-genes.

Para a obtenção das instâncias positivas do conjunto de treino, foram utilizadas as CDS de cada organismo e duas de suas sub ORFs: a maior e a menor. Similarmente, as instâncias negativos foram obtidas por meio da extração das ORFs juntamente com a maior e a menor sub ORF das regiões intergênicas. As instâncias positivas do conjunto de teste correspondem apenas a CDS, sem sub ORFs, enquanto as instâncias negativas são extraídas utilizando a mesma metodologia adotada para a confecção do conjunto de treino.

Foi realizada a extração de características para cada ORF das instâncias positivas e negativas. De cada ORF (ou sub ORF), foram extraídos seis atributos, descritos em [Goés et al. 2014]: 1) percentual de bases de guanina e citosina (percentual GC) na ORF, 2) o percentual GC nas primeiras posições de todos os códons da ORF, 3) nas segundas posições dos códons, 4) nas terceiras posições dos códons, 5) o tamanho da ORF e 6) a variância de códon. Ao fim, cada instância, que corresponde a uma ORF, possui seis atributos e a sua classe.

O balanceamento das classes foi realizado somente no conjunto de treino e apenas para os casos em que o número de instâncias positivas era superior ao número de instâncias negativas. Para esse caso, foram adicionadas aleatoriamente sub ORFs extraídas de instâncias negativas, independentemente se seus tamanhos, ao conjunto de treinamento na tentativa de equilibrar as classes.

Foram construídos assim 4 conjuntos de treinamentos distintos, que foram posteriormente utilizados para treinar 4 modelos independentes de árvores aleatórias [Breiman 2001], com 150 árvores por modelo e utilizando pacote *Caret*, do *R*[Kuhn 2008]. Em dois modelos, denominados modelos *A* e *B*, foi utilizada a estratégia de balancemento de classes nos respectivos conjuntos de treinamento, enquanto nos outros 2 modelos, denominados *C* e *D*, não. Adicionalmente, foram utilizadas sub ORFs na produção de instâncias positivas dos conjuntos de treinamento dos modelos *A* e *C* , enquanto nos modelos *B* e *D* não.

O conjunto de teste foi construído seguindo a mesma estratégia adotada pelo modelo D, considerando assim apenas as CDS, sem suas respectivas sub ORFs, juntamente com as ORFs e suas duas ORFs negativas. Para cada sequência do conjunto de teste, as-

[1] http://www.ncbi.nlm.nih.gov/

sim como ocorrido no treinamento, são extraídos os atributos e utilizados seus respectivos rótulos (indicando se corresponde ou não a uma CDS) para avaliação futura de desempenho dos modelos. Como esperado, esses rótulos são removidos do conjunto de teste e as instâncias são submetidas à predição de cada modelo construído.

Para a análise da precisão dos modelos, foram adotadas três métricas: a *acurácia*, que indica o percentual de predições corretas; a *sensibilidade*, que mede o percentual de instâncias positivas corretamente classificados — indicando assim a capacidade do modelo de identificar um gene como tal; e a *especificidade*, que expressa o percentual de instâncias negativas corretamente classificadas — isto é, o percentual acerto do modelo ao analisar sequências que sabidamente não correspondem a genes. As mesmas sequências utilizadas nos testes dos modelos construídos foram também utilizadas para avaliação de desempenho da ferramenta *FragGeneScan*[Rho et al. 2010].

Adicionalmente, foram realizados experimentos considerando o tamanho das sequências do conjunto de teste. A partir da extração dos atributos, foram analisadas separadamente as instâncias consideradas pequenas, médias e grandes. Como sequências pequenas, foram consideradas as sequências de 60 pares de base (pb) a 300pb; como sequências médias, aquelas de 500pb a 800pb; e, como sequências grandes, foram selecionadas sequências de 1000pb a 1220pb.

3. Resultados

A Tabela 1 apresenta os resultados de acurácia, sensibilidade e especificidade obtidos pelos modelos *A*, *B*, *C* e *D* e pela ferramenta FragGeneScan nos experimentos que consideraram todas as instâncias do conjunto de treinamento, com os melhores resultados destacados em negrito. De modo similar, a Tabela 2 apresenta os valores de sensibilidade e especificidade para os modelos propostos que utilizaram árvores aleatórias, considerando separadamente as instâncias do conjunto de testes correspondentes a sequências pequenas, médias e grandes.

Tabela 1. Valores de acurácia, sensibilidade e especificidade para cada modelo utilizando todas as sequências do conjunto de testes

	Balanceamento	**Sub ORFs**	**Acurácia**	**Sens.**	**Espec.**
Modelo A	SIM	SIM	88,45%	94,29%	80,77%
Modelo B	SIM	NÃO	80,57%	98,11%	57,52%
Modelo C	NÃO	SIM	93,20%	93,79%	**92,43%**
Modelo D	NÃO	NÃO	**93,29%**	95,16%	90,83%
FragGeneScan	—	—	69,47%	**99,91%**	29,45%

Os testes de cada modelo construído e para o FragGeneScan também foram executados separadamente para cada organismo. Os resultados de acurácia, sensibilidade e especificidade desses testes também podem ser encontrados no material suplementar.

4. Conclusões

A partir da análise dos resultados apresentados, algumas conclusões são possíveis no que se refere ao uso da estratégia de balaceamento utilizada e o uso de sub ORFs de CDS. Ao considerarmos as taxas de acurácia dos classificadores no conjunto de testes completo, é

Tabela 2. Valores de sensibilidade e especificidade para sequências pequenas (de 60pb a 300pb), médias (de 500pb a 800pb) e grandes (de 1000pb a 1200pb)

	Pequenas		**Médias**		**Grandes**	
	Sens.	**Espec.**	**Sens.**	**Espec.**	**Sens.**	**Espec.**
Modelo A	70,07%	77,41%	96,77%	**87,10**%	98,89%	**86,97**%
Modelo B	**93,08**%	43,43%	**98,28**%	78,74%	**99,31**%	80,07%
Modelo C	63,55%	**96,80**%	96,70%	86,89%	98,62%	86,20%
Modelo D	71,68%	95,11%	97,21%	85,03%	99,10%	85,44%

possível observar que os melhores resultados foram obtidos pelos modelos treinados com conjuntos de treinamento que não utilizaram a estratégia de balanceamento de classes proposta (modelos C e D), que alcançaram taxas superiores a 93%. Essa conclusão permanece válida ao considerarmos as taxas de especificidade, que foram superiores a 90% somente para esses dois modelos.

Considerando que a estratégia de balanceamento utilizada favorece especialmente a classe negativa do conjunto de treinamento — ou seja, com ela, são fornecidas mais informações a respeito de ORFs que não correspondem a CDS —, sua utilização demonstrou-se inapropriada para os conjuntos de dados utilizados, uma vez que as taxas reduziram com a sua utilização, especialmente a especificidade, que retrata a taxa de acerto dentre as instâncias conhecidamente negativas. Entretanto, é válido notar que houve uma melhoria leve na sensibilidade, indicando que os modelos melhoraram suas capacidades de reconhecimento de exemplos positivos após o balanceamento.

Contudo, essas melhorias mostram-se discretas quando comparadas às reduções em acurácia e especificidade. A sensibilidade aumentou em 0,5% quando se utilizou o balanceamento no conjunto de treinamento que utilizou sub ORFs de CDS e 2,95% ao se utilizar balaceamento no conjunto de treinamento sem sub ORFs de CDS. Em contrapartida, as reduções na acurácia e especificidade ao se utilizar o balanceamento foram, respectivamente, de 4,75% e 11,66% para o conjunto de treinamento com sub ORFs e as reduções nos conjuntos sub ORFs foram, nesta ordem, de 12,72% e 33,31%.

De modo similar, o uso de sub ORFs também não se mostrou apropriado para a construção dos modelos propostos, pois, embora tenha melhorado a especificidade, ele reduziu a sensibilidade. Ao se analisar apenas a especificidade — que avalia a taxa de acerto das instâncias conhecidamente negativas —, nota-se que o uso das sub ORFs foi favorável, especialmente quando se utilizou o balanceamento de classes, com as taxas aumentando de 57,52% para 80,77%. Sem utilizar balanceamento, a especificidade também melhorou, porém, de modo mais discreto. Por outro lado, a sensibilidade reduziu ao se utilizar sub ORFs, tanto nos conjuntos com balanceamento quanto no conjunto sem balaceamento.

Outra conclusão relevante com base nos dados é a obtenção de melhores resultados ao se utilizar árvores aleatórias em comparação com os resultados obtidos pela ferramenta FragGeneScan, que utiliza Modelos Ocultos de Markov. Os modelos baseados em árvores aleatórias apresentaram taxas consideravelmente melhores de acurácia e especificidade. Embora o FragGeneScan tenha obtido melhores taxas de sensibilidade, a distância percentual entre o desempenho da ferramenta e os demais modelos é relati-

vamente pequena. Ao compararmos as taxas do FragGeneScan com o modelo C, por exemplo, que obteve a menor sensibilidade, temos uma diferença percentual de 6,12% em favor do FragGeneScan, contra as diferenças percentuais de 23,73% e 62,98% em favor do modelo C para acurácia e especificidade, respectivamente.

Analisando as taxas de especificidade e sensibilidade dos modelos com árvores aleatórias considerando diferentes tamanhos de sequências, observou-se também que a sensibilidade em todos os modelos melhora à medida que o tamanho das sequências aumentam. Já a especificidade apresentou comportamento similar, mas menos acentuado, quando se utilizou balanceamento. Porém, sem utilizar balanceamento, sequências maiores apresentaram menores taxas de especificidade que as sequências pequenas.

Com base no exposto, concluiu-se que, de forma geral, o uso de sub ORFs como instâncias positivas do conjunto de treinamento e o balanceamento de classes por meio de enriquecimento do conjunto de instâncias negativas interferiram negativamente no poder de generalização dos modelos baseados em árvores aleatórias na classificação supervisionada de genes. A exploração de outras formas de balanceamento, o estudo de características extraídas das sequências e a exploração de parâmetros das árvores aleatórias são propostos como trabalhos futuros de investigação nesta área.

Referências

[Breiman 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[De Filippo et al. 2012] De Filippo, C., Ramazzotti, M., Fontana, P., and Cavalieri, D. (2012). Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Briefings in bioinformatics*, 13(6):696–710.

[Fassetti et al. 2017] Fassetti, F., Giallombardo, C., Leone, O., Palopoli, L., Rombo, S. E., Ruffolo, P., and Saiardi, A. (2017). Automatic simulation of rna editing in plants for the identification of novel putative open reading frames. *PeerJ Preprints*, 5:e3362v1.

[Goés et al. 2014] Goés, F., Alves, R., Corrêa, L., Chaparro, C., and Thom, L. (2014). Towards an ensemble learning strategy for metagenomic gene prediction. In *Advances in Bioinformatics and Computational Biology*, pages 17–24. Springer International Publishing.

[Hoff 2009] Hoff, K. J. (2009). *Gene prediction in metagenomic sequencing reads*. PhD thesis, Georg August University Göttingen.

[Kuhn 2008] Kuhn, M. (2008). Building predictive models inRUsing thecaretPackage. *Journal of Statistical Software*, 28(5).

[Noguchi et al. 2008] Noguchi, H., Taniguchi, T., and Itoh, T. (2008). Metageneannotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA research*, 15(6):387–396.

[Rho et al. 2010] Rho, M., Tang, H., and Ye, Y. (2010). Fraggenescan: predicting genes in short and error-prone reads. *Nucleic acids research*, page gkq747.

[Sieber et al. 2018] Sieber, P., Platzer, M., and Schuster, S. (2018). The definition of open reading frame revisited. *Trends in Genetics*, 34(3):167–170.

[Zhu et al. 2010] Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic acids research*, 38(12):e132–e132.