Universidade de São Paulo
Instituto de Matemática e Estatística
Bacharelado em Ciência da Computação

Técnicas de aprendizado de máquina para predição de fenótipos em bactérias

Eduardo do Nascimento Evaristo

Monografia Final

mac 499 — Trabalho de Formatura Supervisionado

Orientador: Prof. Dr. João Carlos Setubal
Coorientador: Me. Bruno Koshin Vázquez Iha

São Paulo
16 de Março de 2021

Técnicas de aprendizado de máquina para predição de fenótipos em bactérias

Eduardo do Nascimento Evaristo

Esta é a versão original da monografia elaborada pelo candidato Eduardo do Nascimento Evaristo, tal como submetida à Comissão Julgadora.

Agradecimentos

Agradeço primeiramente à minha família, que sempre me apoiou durante toda minha jornada até aqui. Agradeço também aos amigos que fiz durante este período de faculdade e que levarei por toda minha vida, por todos os momentos vividos e aproveitados junto a eles. Agradeço imensamente ao meu orientador, coorientador e demais pessoas que me ajudaram a produzir este trabalho e que me trouxe grande aprendizado.

Resumo

Eduardo do Nascimento Evaristo. **Técnicas de aprendizado de máquina para predição de fenótipos em bactérias**. Monografia (Bacharelado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2021.

A metagenômica é o estudo do material genético obtido a partir de amostras ambientais, conhecido como metagenoma. Os MAGs, metagenome-assembled genomes, são genomas montados de amostras metagenômicas que podem representar novas unidades taxonômicas. O fato de serem novas unidades dificulta as compreensões de suas capacidades metabólicas, e portanto justificam o uso de aprendizado de máquina para facilitar tal compreensão. Fenótipos podem ser entendidos como as características de um organismo. Para este estudo tivemos dois fenótipos de interesse:

- Temperatura de crescimento, que caracteriza as temperaturas ótimas, mínimas e máximas de crescimento de um organismo. Exemplos desse fenótipo são seres psicrófilos (15°C 20°C), mesófilos (20°C 45°C) e termófilos (50°C 122°C).
- Tipo de metabolismo do organismo. Por exemplo fixação de nitrogênio, que diz respeito à capacidade de um organismo converter nitrogênio gasoso em amônia ($N2 \rightarrow NH3$) ou outro composto nitrogenado. Exemplos de bactérias fixadoras de nitrogênio são as rizóbias, quando fixadas nas raízes de leguminosas, e as cianobactérias. Outros exemplos de metabolismo são redutores de sulfato e redutores de nitrato.

Este trabalho teve como objetivo predizer as capacidades fenotípicas acima através de técnicas e algoritmos de aprendizado de máquina, a partir de sequências codificadoras de proteínas identificadas dos organismos utilizados. Foram consideradas para predição técnicas como redes neurais, *support vector machines (SVM)* e regressão logística.

Palavras-chave: Machine-learning. Predição de fenótipos. Grupos Ortólogos

Abstract

Eduardo do Nascimento Evaristo. **Título do trabalho**. Capstone Project Report (Bachelor). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2021.

Metagenomics is the study of genetic material obtained from environmental samples, known as metagenoma. MAGs, metagenome-assembled genomes, are genomes assembled from metagenomic samples that may represent new taxonomic units. The fact that they are new units makes it difficult to understand their metabolic capabilities, and therefore justify the use of machine learning to facilitate such an understanding. Phenotypes can be understood as the characteristics of an organism. For this study we had two phenotypes of interest:

- Growth temperature, which characterizes the optimal, minimum and maximum growth temperatures of an organism. Examples of this phenotype are psychrophilic beings (15 ° C 20 ° C), mesophiles (20 ° C 45 ° C) and thermophiles (50 ° C 122 ° C).
- Nitrogen fixation, which refers to the ability of an organism to convert gaseous nitrogen into ammonia ($N2 \rightarrow NH3$) or another nitrogenous compound. Examples of nitrogen-fixing bacteria are rhizobia, when attached to the roots of legumes, and cyanobacteria.

This work aimed to predict the phenotypic capabilities above through machine learning techniques and algorithms, from coding sequences of proteins identified from the organisms used. Techniques such as neural networks, support vector machines (SVM) and logistic regression were considered for prediction.

Keywords: Machine-learning. Phenotype prediction. Ortholog Groups

Lista de Abreviaturas

DNA	Ácido desoxirribonucleico
RNA	Ácido ribonucleico
ABNT	Associação Brasileira de Normas Técnicas
URL	Localizador Uniforme de Recursos (*Uniform Resource Locator*)
COGs	Cluster of Orthologous Groups
GOLD	Genome OnLine Database
NCBI	National Center for Biotechnology Information
HMM	Hidden Markov Model
BLAST	Basic Local Alignment Search Tool
IME	Instituto de Matemática e Estatística
USP	Universidade de São Paulo

Lista de Figuras

Sumário

Capítulo 1

Introdução

1.1 Contextualização

A bioinformática é o estudo da aplicação de técnicas da informática, matemática e estatística para análise de dados, geração e gerenciamento de informação na biologia. A bioinformática auxilia na obtenção e comparação de dados genéticos, na compreensão da árvore evolutiva, no sequenciamento e anotação de genomas entre muitas outras coisas.

1.2 Justificativa

A metagenômica é o estudo do material genético obtido a partir de amostras ambientais, conhecido como metagenoma. Os MAGs, *metagenome-assembled genomes*, são genomas montados de amostras metagenômicas que podem representar novas unidades taxonômicas. O fato de serem novas unidades dificulta as compreensões de suas capacidades metabólicas, uma vez que não é possível relacioná-las com outras unidades taxonômicas e portanto justificam o uso de técnicas de aprendizado de máquina para facilitar tal compreensão.

1.3 Motivações

Decidi trabalhar com esse tema devido à sua natureza, pois sempre tive interesse em aplicar o conhecimento obtido na computação em outras áreas e tive o prazer de trabalhar um pouco com a bioinformática. Nos últimos anos do curso me interessei pela Inteligência Artificial e fazer este trabalho de conclusão de curso foi uma oportunidade para me aprofundar na área e descobrir um pouco mais sobre bioinformática.

1.4 Objetivos

Este trabalho teve como objetivo testar métodos para predição dos fenótipos temperatura de crescimento e tipo de metabolismo em bactérias a partir de seu genoma, especificamente analisando suas sequências codificadoras de proteínas com técnicas de aprendizado de máquina.

Capítulo 2

Conceitos

Para um entendimento completo deste trabalho, abaixo serão explicados alguns conceitos importantes tanto da computação quanto da biologia e bioinformática.

2.1 Biologia e Bioinformática

2.1.1 Genética

DNA e RNA

O ácido desoxirribonucleico, conhecido como DNA (*deoxyribonucleic acid*), é um composto orgânico no qual encontra-se as informações necessárias para o desenvolvimento, funcionamento e transmissão de características hereditárias do ser vivo. Em seres eucariotos o DNA se encontra dentro do núcleo e em seres procariotos o DNA se encontra disperso no citoplasma. Já o ácido ribonucleico, conhecido como RNA (*ribonucleic acid*), é responsável codificação e descodificação durante o processo de tradução de proteínas.

Tanto o DNA quanto o RNA são compostos por uma cadeia de nucleotídeos ligados em combinações específicas. Um nucleotídeo é formado por um grupo fosfato, um açúcar (desoxirribose no caso do DNA e ribose no RNA) e uma base nitrogenada, que pode ser uma das quatro: adenina (A), citosina (C), guanina (G) ou timina (T), para o DNA, ou uracila (U), para o RNA.

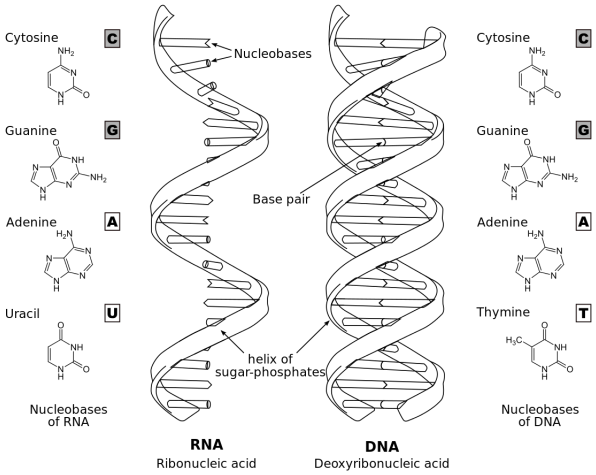

Figura 2.1: *Estrutura de RNA e DNA DNA e RNA*

Genes e Cromossomos

Um gene pode ser definido como "a unidade física e funcional fundamental da hereditariedade. Um gene é uma sequência ordenada de nucleotídeos localizada em uma posição particular em um cromossomo particular que codifica um produto funcional específico (isto é, uma proteína ou molécula de RNA)."[2].

Um cromossomo é uma estrutura compacta que contém informação genética , isto é, vários genes e sequências de nucleotídeos com funções específicas nas células dos seres vivos, ele é formado por uma única molécula de DNA.[3] A ordem dos nucleotídeos de uma molécula de DNA pode ser determinada através dos métodos de sequenciamento de DNA.

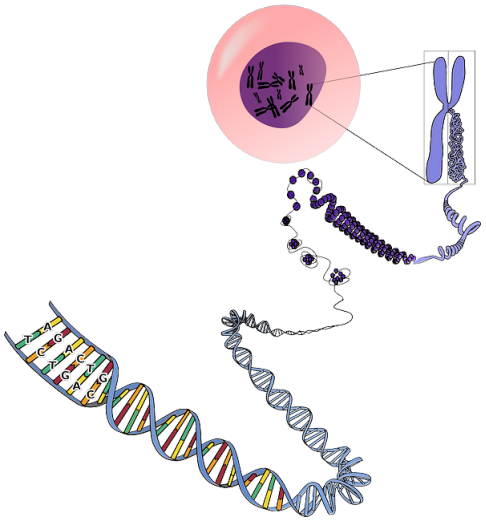

Figura 2.2: *Cromossomo*

Homologia

A homologia é o compartilhamento de uma estrutura biológica entre organismos diferentes que possuem origem comum, pois foi herdada de um ancestral em comum. Os genes de organismos distintos que foram herdados a partir do mesmo ancestral são ditos genes homólogos. Podemos citar três tipos de homologia entre genes: ortólogos, parálogos e xenólogos, nosso interesse aqui é na ortologia.

Genes que divergiram pelo processo de especiação e a história do gene reflete a história da espécie são chamados de ortólogos. Genes ortólogos tendem a conservar função.[4].

A identificação de ortólogos constitui uma das tarefas fundamentais da biologia molecular e evolucionária,[5]diversos bancos de dados que fornecem relações ortólogas foram desenvolvidos ao longo dos anos e têm sido utilizados para diversas funções. Entre elas para a comparação de organismos por categorias funcionais.[6]

2.2 Computação

2.2.1 Aprendizado de máquina

No aprendizado de máquina supervisionado, algoritmos são utilizados para gerar modelos preditivos a partir da observação de um conjunto de dados rotulados, isto é, dados já classificados, este conjunto é mais conhecido como conjunto de treinamento. Nesse caso os rótulos do conjunto de treinamento utilizado são referentes ao fenótipo, como "alta temperatura", "baixa temperatura"etc.. Dessa forma, um algoritmo de classificação buscará produzir um modelo preditivo generalizador, baseado nas informações contidas no conjunto de treinamento, que seja capaz de classificar dados cujo rótulo é desconhecido. Quando os valores a serem preditos são limitados e discreto, temos um problema de classificação, quando estes valores são contínuos, temos um problema de regressão. [7]. Para a predição dos fenótipos neste projeto foram utilizadas redes neurais e regressão logística, que serão explicadas abaixo.

2.2.2 Redes Neurais

"Redes Neurais Artificiais são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência."[8]. Uma rede neural é composta por várias unidades de processamento, que recebem valores de entrada e realizam uma operação sobre eles, produzindo então uma saída que pode servir de entrada para outra unidade de processamento em uma camada adiante.

Normalmente cada valor de entrada é multiplicado por um peso e somado ao produto seguinte e ao final é aplicada uma função de ativação.

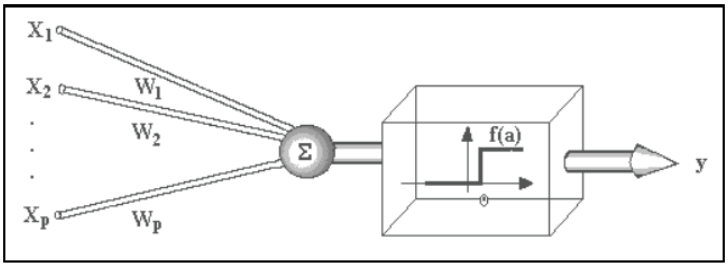

Figura 2.3: *Unidade de processamento de uma rede neural* Redes Neurais

Na maioria dos modelos esses pesos são atualizados de acordo com uma função custo a cada iteração, isto é, a cada exemplo fornecido para a rede.

A arquitetura de uma rede neural é normalmente organizada em camadas, que podem ser divididas em camada de entrada, camadas intermediárias e camada de saída.

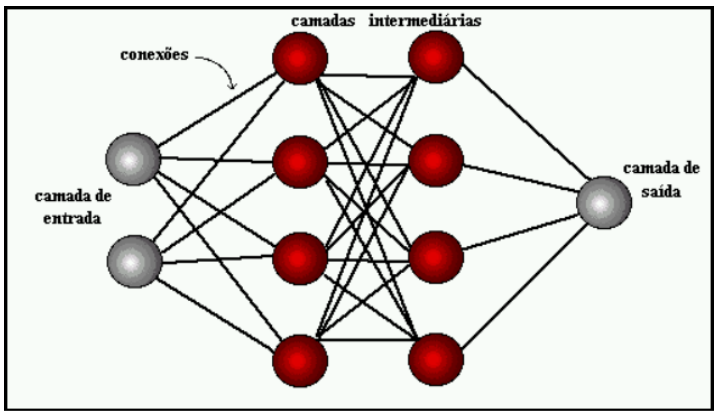

Figura 2.4: *Arquitetura de uma rede neural genérica* Redes Neurais

2.2.3 Regressão Logística

A regressão logística é um algoritmo de classificação utilizado para predizer um conjunto de classes discretas, é um método de análise preditiva baseada em probabilidade. A lei da função logística é descrita por:

$$p(i) = \frac{1}{1 + e^{(}b_0 + b_1 x_1, i + ... + b_n x_n, i)} \tag{2.1}$$

Como podemos ver função de probabilidade nada mais é que uma combinação linear da entrada $x_i, ..., x_n$ e $b_0, ..., b_n$ são seus pesos. A função custo utilizada para atualização do vetor de pesos no caso da regressão logística multinomial é:

$$L_{\log}(y_i, p_i) = -(y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \tag{2.2}$$

Onde y_i é o rótulo do dado i e p_i é a predição dada pela regressão logística do dado i.

Capítulo 3

Implementação e Desenvolvimento

Neste capítulo será descrito todo o processo que foi feito para obtenção dos resultados, tendo como partida dois fenótipos de interesse: temperatura de crescimento e fixação de nitrogênio.

3.1 Lista de organismos

Para iniciar todo o processo foi necessário obter uma lista de organismos que possuem os fenótipos de interesse. Para isso, foi realizada uma busca no banco de dados online GOLD (Genomes Online Database)[9] utilizando os filtros adequados para selecionar os organismos corretos.

Para a temperatura de crescimento, a busca resultou em:

- 80 organismos hipertermófilos;
- 624 organismos termófilos;
- 33 organismos termotolerantes;
- 765 organismos mesófilos;
- 196 organismos psicrotolerantes;
- 201 organismos psicrófilos.

Totalizando 1899 bactérias. Para o tipo de metabolismo a busca resultou em:

- 117 organismos redutores de sulfato;
- 304 organismos fixadores de nitrogênio;
- 117 organismos redutores de nitrato;
- 96 organismos metanogênicos.

Totalizando 634 bactérias referentes ao fenótipo tipo de metabolismo.

3.2 *Download* de sequências

Com a lista de organismos em mãos era necessário então baixar as sequências de aminoácidos referentes à proteínas identificadas em regiões codificantes do DNA de um gene, contidos em um arquivo de formato .faa (FASTA amino acid).

```
>lcl|NZ_AUMW01000001.1_prot_1 [locus_tag=H569_RS18790] [protein=IS110 family transposase] [pseudo=true] [frame=3] [location=<1..200] [gbkey=CDS]
RGANRAAVAVAHSILTIVYHIIKRKQPYIELGPTFYEERKRETVIKQSIKKLESLGYKVTVELIA
>lcl|NZ_AUMW01000001.1_prot_WP_027363251.1_2 [locus_tag=H569_RS0100010] [protein=UPF0280 family protein] [protein_id=WP_027363251.1] [location=complement(349..1092)] [gbkey=CDS]
MTDYVNRTYRLLHNQKDLSFFRVVVKETDLDIGVPGRFVTPALIEGVRKEVLTIRSQIEGYISENELFLTTLEPFLVSKE
APELVRIMAEAGNAAGIGPMSAVAGAIAHHIGSYLARRSSEVIVENGGDIYLRSGRLRKVGIFAGPSPFTNKLAIEIEPH
QTPLGICTSSGTVGHSLSFGCADAVVILAPSTALADAVATATGNIVQSEADLQAAVDFAMGIKSVTGAMVIKNDKLAAAG
NIRLAPV
```

Figura 3.1: *Exemplo de um arquivo .faa*

Para isso, foi realizado um *script* que lê a lista de organismos obtida em 3.1, procura seu identificador e realiza o *download* do arquivo .faa no banco de dados do NCBI (National Center for Biotechnology Information)[10], disponível *online*.

Deve-se ressaltar que nem todos os organismos obtidos em 3.1 estavam possuíam genomas disponíveis no banco de dados do NCBI, uma série de fatores podem ser responsáveis pela sua ausência no banco [1]1, como baixa qualidade da sequência, baixa contagem de genes etc. Portanto ao final desta etapa tínhamos (a contagem de organismos de cada tipo segue em parênteses):

- Hipertermófilos (18);
- Termófilos (310);
- Termotolerantes (22);
- Mesófilos (269);
- Psicrotolerantes (131) ;
- Psicrófilos (270).

Totalizando 1020 bactérias referentes à temperatura de crescimento. E (a contagem de organismos de cada tipo segue em parênteses):

- Fixadoras de nitrogênio (90);
- Redutoras de nitrato (85);
- Redutoras de sultato (92);
- Metanogênicas (53).

Totalizando 320 bactérias referentes ao metabolismo.

Pudemos observar que a distribuição de organismos por grupo na temperatura de crescimento era desproporcional e poderia dificultar ou mesmo impossibilitar a classificação de grupos pouco representados, como hipertermófilos e termotolerantes. A distribuição de organismos por grupo referentes ao metabolismo foi uniforme o suficiente para seguir

inalterada. Foi decidido então, para facilitar o treinamento dos dados e obter melhores resultados, reduzir o número de grupos referentes à temperatura de crescimento para três da seguinte forma:

- Alta temperatura (350): Hipertermófilos, Termófilos e Termotolerantes;
- Mesófilos (269);
- Baixa temperatura (401): Psicrotolerantes e Psicrófilos.

3.3 Grupos ortólogos

Esta etapa consistiu em extrair uma medida de proximidade entre os organismos. Isso foi feito através da busca por grupos ortólogos entre as proteínas, para isso foi necessário utilizar o programa HMMER[12], mais especificamente seu comando *hmmsearch* em conjunto com o banco de dados eggNOG[13] que contém 206782 grupos ortólogos baseados em genomas procarióticos para busca e comparação com as proteínas.

A ideia deste comando é realizar uma busca de cada família de proteínas representada no arquivo *.hmm* no arquivo de sequências fornecido (*.faa*), produzindo como saída os melhores resultados, isto é, uma tabela de valores indicativos do pertencimento de uma sequência à uma família.

Essa parte provou-se a mais custosa por motivos computacionais, de armazenamento e tempo: o comando *hmmsearch* utilizado junto ao banco de dados eggNOG pode durar até 9 horas (dependendo do tamanho do arquivo de sequências), na maioria das vezes a saída era produzida dentre 2 a 4 horas. Para isso foi necessário utilizar as máquinas do Laboratório de Bioinformática do Instituto de Química da USP, permitindo rodar 15 arquivos de sequências por vez.

```
Query:       2Z7M1.faa.final_tree.fa  [M=295]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence                              Description
    ------- ------ -----    ------- ------ -----   ---- --  --------                              -----------
  ------ inclusion threshold ------
      0.023   10.9   0.0      0.077    9.2   0.0    1.7  2  lcl|NC_014831.1_prot_WP_013494655.1_225  [gene=mgtE] [locus_tag=TMAR_RS01
```

Figura 3.2: *Exemplo de um trecho de um arquivo .out produzido por hmmsearch*

3.4 Transformação dos dados

Para utilizar a saída produzida pelo processo, foi necessário realizar uma extração das informações relevantes contidas no arquivo, que consistia em obter o *E-value* de melhor *bit score* de cada uma das famílias de proteínas.

"Um *bit score* é uma pontuação de razão logit (base dois) comparando a probabilidade do perfil HMM com a probabilidade de uma hipótese nula (um modelo de sequência aleatória distribuída de forma idêntica e independente, como no BLAST). Um *E-value* é o número de resultados esperados por atingir este *bit score* ou maior por "acaso", ou seja, se

a busca tivesse sido feita em um banco de dados de tamanho idêntico composto apenas de sequências não homólogas aleatórias."(Finn, 2011, tradução nossa)

Em seguida foram aplicadas duas funções a cada *E-value* obtido para utilizar na seção de treinamento dos dados. A primeira função f consistia numa transformação simples da seguinte forma:

$$f(Eval) = \begin{cases} 1, & \text{para } Eval \leq 1.10^{-20} \\ 0.5, & \text{para } 1.10^{-20} \leq Eval \leq 0.001 \\ 0, & \text{para } Eval > 0.001 \end{cases} \tag{3.1}$$

A segunda função g foi definida como:

$$g(Eval) = -\log Eval \tag{3.2}$$

```
COG3708.faa.final_tree.fa       0
COG3709.faa.final_tree.fa       1.0
COG3710.faa.final_tree.fa       1.0
```

Figura 3.3: *Exemplo de um trecho de um arquivo aplicando f*

```
2Z7WX.faa.final_tree.fa 74.14266750356873
2Z7WY.faa.final_tree.fa 1.455931955649724
2Z7WZ.faa.final_tree.fa 1.7958800173440752
```

Figura 3.4: *Exemplo de um trecho de um arquivo aplicando g*

Ao final disso tudo, temos um arquivo simples onde cada linha representa um identificador de um grupo ortólogo cadastrado na base eggNOG e seu valor correspondente.

3.5 Aprendizado de máquina

Com os arquivos em mãos, iniciou-se a parte final do processo. A plataforma escolhida para desenvolver e aplicar as técnicas de computação foi o Google Colab, utilizando a linguagem de programação Python. Foram utilizadas também as bibliotecas scikit-learn para métricas de avaliação, regressão logística e pré-processamento de dados e keras para construção das redes neurais.

Inicialmente foi cogitado utilizar SVM, mas em resultados preliminares não apresentou números satisfatórios, por este motivo a técnica foi descartada.

Capítulo 4

Resultados

4.1 Temperatura de Crescimento

Devido a transformação da variável categórica temperatura de crescimento, foram determinados os seguintes *labels* para classificação: 0 = Alta temperatura; 1 = Média temperatura; 2 = Baixa temperatura.

4.1.1 Redes Neurais

A melhor combinação de parâmetros para uma rede neural de três camadas apresentou os seguintes resultados:
Métricas:

	precision	recall	f1-score	support
0	0.78	0.86	0.82	57
1	0.64	0.67	0.65	55
2	0.82	0.74	0.77	91
accuracy			0.75	203

Figura 4.1: *Métricas precisão, recall, f1-score e support (quantidade)*

Matriz de confusão:

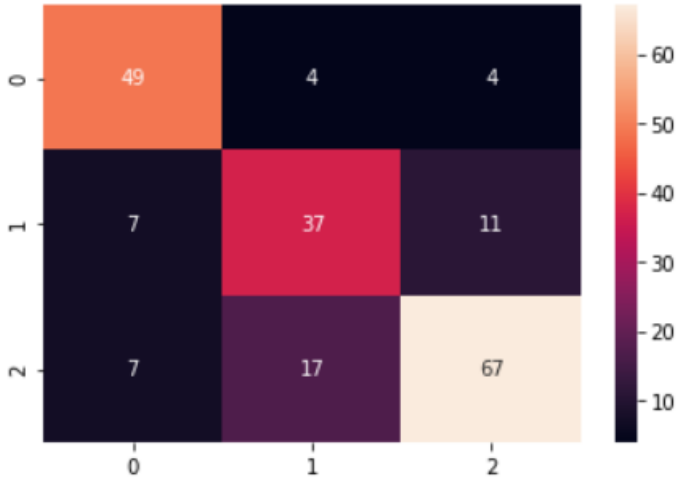

Figura 4.2: *Matriz de confusão*

4.1.2 Regressão Logística

Métricas:

	precision	recall	f1-score	support
0	0.76	0.95	0.84	57
1	0.71	0.67	0.69	55
2	0.84	0.74	0.78	91
accuracy			0.78	203

Figura 4.3: *Métricas precisão, recall, f1-score e support (quantidade)*

Matriz de confusão:

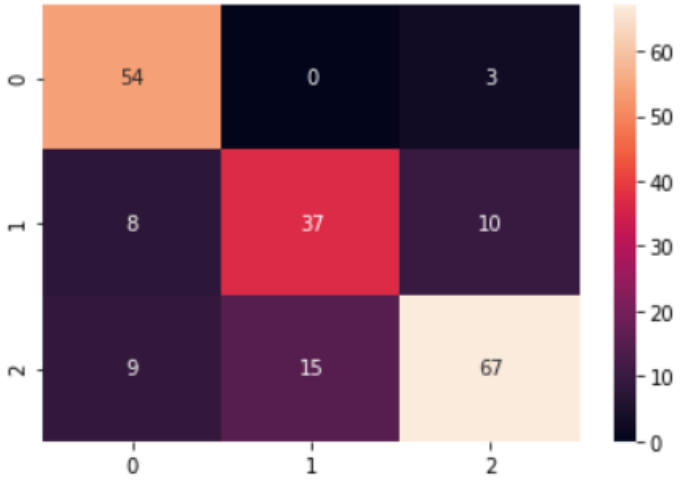

Figura 4.4: *Matriz de confusão*

4.2 Tipo de Metabolismo

Devido a transformação da variável categórica temperatura de crescimento, foram determinados os seguintes *labels* para classificação: 0 = Metanogênicos 1= Redutores de nitrato; 2 = Fixadores de Nitrogênio; 3 = Redutores de sulfato.

4.2.1 Redes Neurais

A melhor combinação de parâmetros para uma rede neural de três camadas apresentou os seguintes resultados:
Métricas:

	precision	recall	f1-score	support
0	0.33	0.75	0.46	4
1	0.69	0.53	0.60	17
2	0.91	0.95	0.93	22
3	0.95	0.86	0.90	21
accuracy			0.80	64

Figura 4.5: *Métricas precisão, recall, f1-score e support (quantidade)*

Matriz de confusão:

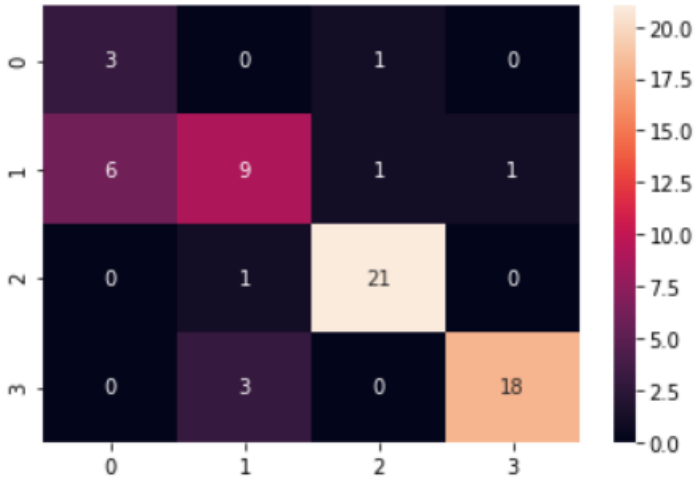

Figura 4.6: *Matriz de confusão*

4.2.2 Regressão Logística

Métricas:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	4
1	0.93	0.82	0.87	17
2	0.96	1.00	0.98	22
3	0.91	0.95	0.93	21
accuracy			0.94	64

Figura 4.7: *Métricas precisão, recall, f1-score e support (quantidade)*

Matriz de confusão:

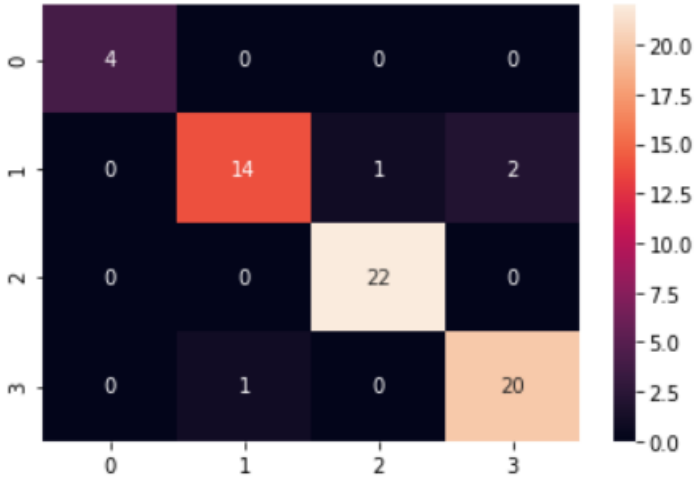

Figura 4.8: *Matriz de confusão*

4.3 Análise

Pudemos observar que para ambos os fenótipos o método de aprendizado por regressão logística apresentou resultados mais satisfatórios: 78% de acurácia em temperatura de crescimento e 94% de acurácia em tipo de metabolismo. Podemos notar também que o tipo de metabolismo apresentou resultados superiores à temperatura de crescimento em ambos os métodos de predição.

Capítulo 5

Conclusão e Considerações Finais

Levando-se em conta que este trabalho teve como objetivo testar métodos para predição dos fenótipos, este trabalho cumpriu seu objetivo. Todavia, existe espaço para aperfeiçoamento e desdobramentos.

A quantidade de organismos disponíveis para treinamento após obter a lista pelo GOLD, utilizando os filtros adequados, e ter realizado os downloads pelo NCBI foi consideravelmente baixa (ver 3.2), especialmente para o fenótipo tipo de metabolismo. Esse fator deve ser considerado e limita conclusões referentes aos desempenho das técnicas empregadas, já que o conjunto de dados para teste também é muito pequeno.

Outro fator a ser considerado foi a dimensionalidade dos dados. Mesmo aplicando pré-processamento nos dados, o conjunto ainda apresentou alta dimensionalidade (ver 3.5). Este é um fator que afeta o treinamento, uma vez que possui uma maior quantidade de combinações de valores de entrada.

Possíveis sugestões para aperfeiçoamentos e desdobramentos são: elaborar uma lista de organismos utilizando múltiplas bases de dados, buscando aumentar o número de espécies; utilizar algum outro banco de dados de grupos ortólogos; utilizar outras ferramentas além do HMMER; utilizar outras técnicas de aprendizado de máquina.

Considerando agora as motivações e objetivos pessoais, posso afirmar que este trabalho excedeu-se. No final, foram utilizados não só técnicas e conceitos da inteligência artificial, que era o esperado, como também da computação paralela e *bash script* que se mostraram essenciais. Este trabalho foi uma grande fonte de aprendizado.

Bibliografia

[1] Hugenholtz, P et al. “Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity.” Journal of bacteriology vol. 180,18 (1998): 4765-74. doi:10.1128/JB.180.18.4765-4774.1998

[2] João C. Setubal et al. (eds.), Comparative Genomics: Methods and Protocols, Methods in Molecular Biology, vol. 1704.

[3] Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, Christian von Mering, Peer Bork, eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses, Nucleic Acids Research, Volume 47, Issue D1, 08 Jan 2019, pgs D309–D314. Disponível em: https://doi.org/10.1093/nar/gky1085. Acesso em: 13 Mar 2021.

[4] Galperin MY, Kristensen DM, Makarova KS, Wolf YI, Koonin EV. Microbial genome analysis: the COG approach. Brief Bioinform. 2019 Jul 19;20(4):1063-1070. Disponível em: https://pubmed.ncbi.nlm.nih.gov/28968633/. Acesso em: 13 Mar. 2021.

[5] JOAQUIM, Leyla Mariane; EL-HANI, Charbel Niño. A genética em transformação: crise e revisão do conceito de gene. Sci. stud., São Paulo , v. 8, n. 1, p. 93-128, mar. 2010 . Disponível em http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1678-31662010000100005&lng=pt&nrm=iso. Acesso em 26 fev. 2021.

[6] Karp, Gerald (2008). Cell and Molecular Biology. Concepts and Experiments (em inglês) 5ª ed. New Jersey: John Wiley. p. 390-395.

[7] Padilha, V. A. e Carvalho, A. C. P. L. F., Mineração de Dados em Python, cap. 5. Disponível em: https://edisciplinas.usp.br/pluginfile.php/4125431/mod_resource/content/2/mineracaodadosbiologicos-parte5.pdf. Acesso em: 27 fev. 2021.

[8] Carvalho, A. C. P. L. F., Redes Neurais Artificiais. Disponível em: https://sites.icmc.usp.br/andre/research/neural/. Acesso em: 27 fev. 2021.

[9] Supratim Mukherjee, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Jagadish Chandrabose Sundaramurthi, Janey Lee, Mahathi Kandimalla, I-Min A Chen, Nikos C Kyrpides and T B K Reddy. Genomes OnLine Database (GOLD) v.8: overview and updates. Nucl. Acids Res. (2020) doi: doi.org/10.1093/nar/gkaa983

[10] National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information;. Disponível em: https://www.ncbi.nlm.nih.gov/

[11] Assembly Anomalies and Other Reasons a Genome Assembly may be Excluded from RefSeq, National Center for Biotechnology Information;. Disponível em: https://www.ncbi.nlm.nih.gov/assembly/help/anomnotrefseq/

[12] HMMER: Biological sequence analysis using profile hidden Markov models. 3.1. Sean R. Eddy and the HMMER development team, 2020. http://hmmer.org/. Acesso em: 10 Dez. 2020.

[13] eggNOG. Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, Christian von Mering, Peer Bork Nucleic Acids Res. 2019 Jan 8. Disponível em: https://academic.oup.com/nar/article/47/D1/D309/5173662. Acesso em: 21 Out. 2020.

[14] Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W29-37. doi: 10.1093/nar/gkr367. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125773/. Acesso em: 10 Dez. 2020.

[15] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. Disponível em: https://scikit-learn.org/. Acesso: em 12 Jan. 2021.