Universidade de São Paulo
Instituto de Matemática e Estatística
MAC0499 - Trabalho de Formatura Supervisionado

Eduardo do Nascimento Evaristo
Supervisor: João Carlos Setubal

Introdução e Motivação

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.

A ideia então é, através do AM, predizer fenótipos de novas unidades taxonômicas a partir de seus genomas. Fenótipos podem ser entendidos como as características de um organismo. Para este tudo temos 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).
- Fixação de nitrogênio, que diz respeito à capacidade de um organismo converter nitrogênio gasoso em amônia ($N_2 \rightarrow NH_3$) ou outro composto nitrogenado. Exemplos de bactérias fixadoras de nitrogênio são as rizóbias, quando fixadas nas raízes de leguminosas, e as cianobactérias.

Outro fator interessante de se explorar é descobrir quais *features* são relevantes para a predição dos fenótipos, presença/ausência de determinados genes, porcentagem guanina-citosina (*GC* %) e *codon usage* são exemplos de possíveis *features.*

Objetivos

- Aprofundar conhecimentos tanto em aprendizado de máquina quanto em genômica e bioinformática
- Estudar dois fenótipos bacterianos de interesse: temperatura de crescimento e capacidade fixar nitrogênio
- Criar conjuntos de genomas de treinamento, teste, e validação para esses dois fenótipos
- Escolher um modelo de aprendizado de máquina que seja adequado para os fenótipos de interesse (sendo que poderemos ter um modelo para temperaturas e outro para fixação de nitrogênio)
- Utilizar esses modelos para realização de testes
- Caso os testes tenham bons resultados, aplicar os modelos em MAGs

Metodologia

Os genomas para treinamento serão obtidos do NCBI (National Center for Biotechnology Information), uma plataforma confiável para obtenção de dados na área de bioinformática. A divisão entre conjuntos de teste, treinamento e validação será feita com o método *k-fold cross validation*, dependendo do número de amostras obtidas.
A lista de organismos termofílicos e não termofílicos será obtida de um artigo que examinou as temperaturas de crescimento ótimas de genomas microbianos no NCBI e disponibilizou uma tabela com seus nomes[6].
A lista de organismos que fixam/não fixam nitrogênio será obtida através de um filtro no *web-service* GOLD (Genomes online database).

Pretende-se utilizar a biblioteca *scikit-learn* para questões relacionadas ao aprendizado de máquina, tanto para os algoritmos de SVM e redes neurais como para divisão dos conjuntos em treinamento, teste, validação e para o cálculo das métricas de avaliação. As métricas de avaliação utilizadas serão acurácia, especificidade e sensibilidade:

$$\text{Acurácia} = TP + TN / (TP + FP + TN + FN)$$
$$\text{Especificidade} = TN / (TN + FP)$$
$$\text{Sensibilidade} = TP / (TP + FN)$$

Onde TP, TN, FP, FN são o número de organismos identificados corretamente como possuidores do fenótipo, o número de organismos identificados corretamente como não-possuidores do fenótipo, o número de organismos que não possuem o fenótipo mas classificados como possuidores e o número de organismos que possuem o fenótipo mas classificados como não-possuidores respectivamente.

Ao final, quando o reconhecimento da ferramenta estiver satisfatório pretendemos aplicá-la para o reconhecimento de uma coleção de MAGs que temos no laboratório do professor João Carlos Setubal. A coleção conta com metagenômicos de:
Zoológico de SP: compostagem (60), lago S. Francisco (51), fezes de bugios (55) e Esponjas do Sistema Recifal Amazônico (115), os números entre parênteses são a quantidade de MAGs.

Planejamento e Cronograma

Seguem as etapas de planejamento que serão seguidas para o desenvolvimento do projeto:

1 - Estudos relacionados ao tema, tanto relacionados à microbiologia, como DNA, metagenômica e dogma central [1][2][3], quanto relacionados ao aprendizado de máquina, como artigos da área e ferramentas já desenvolvidas [3][4][5] .
2 - Seleção de dados de treinamento para o projeto. Para começar foi feita a escolha do fenótipo *temperature range*. Para isso será necessário obter a temperatura ótima de crescimento de alguns organismos através da database GOLD e depois baixar seus genomas de uma base de dados confiável, no caso NCBI.
3 - Treinamento e implementação de algoritmos de aprendizado de máquina. Serão utilizados redes neurais e SVM, a escolha foi feita baseada em resultados de artigos na área e por familiaridade com os mesmos.
4 - Aplicar as tarefas 2 e 3 ao segundo fenótipo escolhido, capacidade de fixar nitrogênio.
5 - Assim que os métodos de predição estiverem apresentando resultados satisfatórios, aplicá-los em um conjunto de MAGs.
6 - Analisar os resultados obtidos e conclusões.
7 - Redigir a monografia e produzir o pôster.

	Abril	Maio	Jun	Jul	Agos	Set	Out	Nov
1	X	X	X		X			
2					X			
3					X	X	X	
4						X	X	
5						X	X	
6						X	X	X
7							X	X

Referências

1. Metagenomics: Application of Genomics to Uncultured Microorganisms. Jo Handelsman.Microbiology and Molecular Biology Reviews Dec 2004, 68 (4) 669-685; **DOI:** 10.1128/MMBR.68.4.669-685.2004

2. Castrignano, Silvana Beres, & Nagasse-Sugahara, Teresa Keico. (2015). Abordagem metagenômica e causalidade em virologia. *Revista de Saúde Pública*, *49*, 21. Epub April 10, 2015. https://doi.org/10.1590/S0034-8910.2015049005475
3. Cock PA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B and de Hoon MJL (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25, 1422-1423

4. From genomes to phenotypes: Traitar, the microbial trait analyzer. Aaron Weimann, Kyra Mooren, Jeremy Frank, Phillip B Pope, Andreas Bremges, Alice C McHardy. *mSystem* (2016) doi:10.1101/043315
5. Feldbauer, R., Schulz, F., Horn, M., & Rattei, T. (2015). Prediction of microbial phenotypes based on comparative genomics. *BMC bioinformatics*, *16 Suppl 14*(Suppl 14), S1. https://doi.org/10.1186/1471-2105-16-S14-S1
6. **Lin, H. e Chen, W.** Prediction of thermophilic proteins using feature selection technique. *Journal of Microbiological Methods.* 2011, Vol. 84, pp. 60-70.