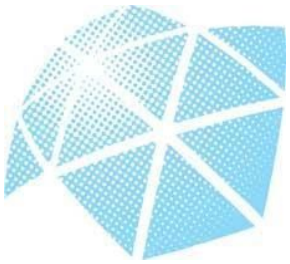
# RESSALVA

Atendendo solicitação do(a) autor(a), o texto completo desta tese será disponibilizado somente a partir de 03/03/2019.

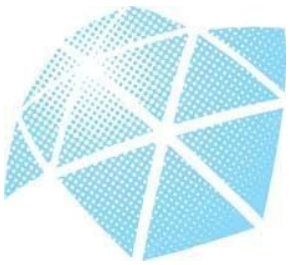


Universidade Estadual Paulista “Júlio de Mesquita Filho”
Instituto de Biociências de Botucatu
Programa de Pós-Graduação em Biotecnologia

# Predição de rotas metabólicas de enzimas utilizando aprendizado de máquina

**Rodrigo de Oliveira Almeida**

Botucatu - SP
2018



Universidade Estadual Paulista “Júlio de Mesquita Filho”
Instituto de Biociências de Botucatu
Programa de Pós-Graduação em Biotecnologia

# Predição de rotas metabólicas de enzimas utilizando aprendizado de máquina

**Doutorando: Rodrigo de Oliveira Almeida**
**Orientador: Dr. Guilherme Targino Valente**

Tese apresentada ao Programa de Pós-Graduação em Biotecnologia do Instituto de Biociências de Botucatu da Universidade Estadual Paulista “Júlio de Mesquita Filho”, para obtenção do título de doutor.

Botucatu - SP
2018

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Almeida, Rodrigo de Oliveira.
Predição de rotas metabólicas de enzimas utilizando aprendizado de máquina / Rodrigo de Oliveira Almeida. - Botucatu, 2018

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Guilherme Targino Valente
Capes: 90400003

1. Aprendizado do computador. 2. Bioinformática. 3. Enzimas. 4. Proteínas - Metabolismo.

Palavras-chave: Aprendizado de máquina; Bioinformática; Enzimas; Rotas metabólicas.

**"Quanto mais nos elevamos, menores parecemos aos olhos daqueles que não sabem voar".**

**Friedrich Wilhelm Nietzsche**

**Dedico este trabalho à minha família e amigos que, mesmo à distância, contribuíram para minha formação pessoal e profissional.**

## Agradecimentos

Ao Programa de Pós-Graduação em Biotecnologia da Universidade Estadual Paulista, pela oportunidade de atuar no curso de doutorado e suporte aos trabalhos.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio financeiro.

Ao meu orientador, Dr. Guilherme Targino Valente, pela amizade e confiança, além das contribuições importantes para o desenvolvimento do trabalho e para minha formação.

Ao Dr. Rafael Plana Simões e Dr. Ney Lemke, pela amizade e auxílio no trabalho.

Ao Dr. Henrik Stotz, da University of Hertfordshire – UK, que me recebeu em seu laboratório para a realização de meu estágio no exterior, além da confiança em meu trabalho.

Aos amigos do SBGL (Systems Biology and Genomic Laboratory – UNESP) e LGI (Laboratório de Genômica Integrativa - UNESP), pelo convívio e aprendizado durante estes anos.

Aos colegas de pós-graduação e professores da UNESP, pela oportunidade de participar e contribuir em suas pesquisas e trabalhos.

**Resumo**

Enzimas são uma classe de proteínas responsáveis por catalisar diversos tipos de reações químicas presentes em diferentes rotas metabólicas, sendo assim o principal foco de estudo nas áreas de engenharia metabólica e biologia sintética. Contudo, a anotação de enzimas e a identificação da rota metabólica em que atuam, são frequentemente baseados na similaridade de sequências previamente descritas. A falta e dificuldade de anotação das enzimas se devem pela diversidade funcional em sequências similares de famílias proteicas, sequências espécie-específicas e a dificuldade na definição de homologia em larga escala. De modo a auxiliar a superar tais problemas, o presente trabalho objetivou criar um classificador de rotas metabólicas de enzimas baseado inteiramente nas características da estrutura primária de enzimas e utilizando aprendizado de máquina. A ferramenta computacional criada (mAppLe - Metabolic Pathway Prediction of Enzymes) é composta por 11 preditores de rotas metabólicas de fungos, podendo assim auxiliar nas anotações dos bancos de dados e em trabalhos nas diferentes áreas de pesquisa, como biologia sintética e engenharia metabólica. As performances médias de predição foram de 94% de acurácia, 44% de taxa de falsa descoberta, 67% de F-*score*, 98% de sensitividade, 93% de especificidade e 0,69 para coeficiente de correlação de Matthews. Com base no desempenho dos preditores criados, constata-se que a ferramenta computacional criada pode ser aplicada com grande sucesso na predição de rotas metabólicas de enzimas de fungos, independente da similaridade das sequências.

Palavras-chave: Aprendizado de máquina, Enzimas, Rotas metabólicas.

**Abstract**

Enzymes are a class of proteins that are responsible for catalyzing chemical reactions in numerous metabolic pathways and are often "main targets" in metabolic engineering and synthetic biology. However, enzyme annotation and metabolic pathway identifications are often based on sequence similarities to previously well-described enzymes. Functional diversity in similar sequences of protein families, species-specificity, and difficult-to-define large-scale homologies results in difficulties and a lack of annotation. Here, we present the mAppLe (Metabolic Pathway Prediction of Enzymes), the first metabolic pathway classifier for enzymes based only on primary structure features and a machine learning approach, surpassing limitations imposed by sequence similarities. This tool is composed of 11 pathways predictors for fungi, that can help databank annotations and several type of researches like synthetic biology and metabolic engineering. Results show an average performance of 94% to accuracy, 44% false discovery rate, 67% F-*score*, 98% sensitivity, 93% specificity and 0.69 to Matthews coefficient correlation. Based on the performance of this predictors, the computational tool created (mAppLe) can be applied successfully to predict pathways of enzymes of the fungi, independent of sequence similarity.

key-words: Machine learning, Enzymes, Metabolic pathways.

## Sumário

# 1. INTRODUÇÃO

## 1.1. Enzimas

Enzimas são as proteínas mais notáveis e altamente especializadas, ponto central nos processos bioquímicos. Catalisam inúmeras reações intra e extracelulares, com alta velocidade e especificidade, degradando macromoléculas para precursores mais simples, e transformando e conservando energia (LODISH et al., 2004 ; NELSON e COX, 2011). Realizam as mais diversas reações bioquímicas, com pH e temperaturas específicas para o correto funcionamento. Suas concentrações e atividades podem ser reguladas, de forma a permitir suas ações dentro das oscilações do meio na qual se encontram. Essas regulações podem ser via inibição por *feedback*, regulação alostérica, fosforilação, compartimentalização, cofatores, entre outros.

O sítio ativo de uma enzima (local onde o substrato se liga para conversão em um produto) contém resíduos de aminoácidos que se ligam ao substrato e agem na substituição de grupos específicos, realizando assim a transformação química (NELSON e COX, 2011). A transformação e conservação da energia acontecem com uma série de reações interconectadas, formando longas rotas que permitem a sobrevivência, crescimento e reprodução celular (ALBERTS, 2015), gerando, contudo, uma extensa rede metabólica (ORTH et al., 2011).

Atualmente, são conhecidos diversos tipos de reações bioquímicas/metabólicas, genes que regulam cada tipo de enzima, e substratos e produtos referentes a uma enzima específica, possibilitando assim calcular o fluxo de metabólitos de uma determinada rota metabólica. Com este tipo de abordagem (análise de fluxo de metabólitos), pode-se reconstruir redes metabólicas e possibilitar a predição da taxa de crescimento de um organismo, ou até mesmo a taxa de produção de um metabólito específico de interesse (ORTH et al., 2011). Um determinado conjunto de enzimas que catalisam reações bioquímicas específicas em um organismo, transformando um composto inicial até ao composto final necessário, é chamado de rota metabólica (PLANES e BEASLEY, 2009).

Sendo assim, uma rota metabólica é uma parte da extensa e complexa rede metabólica (SCHREIBER, 2003).

A atividade biológica de uma enzima é tipicamente determinada por uma parte da cadeia polipeptídica conhecida como domínio (TIAN et al., 2004, NELSON e COX, 2011). Estes domínios são regiões com funções bem definidas e uma proteína pode possuir diferentes domínios em sua estrutura. Embora a função de um domínio seja conservado, ele pode ser alterado por mutações, deleções e inserções, podendo gerar um novo domínio e até mesmo uma nova função (BULJAN e BATEMAN, 2009). Sendo assim, os domínios definem a função a ser exercida pela enzima e seu local de atuação de uma rota bioquímica. Logo, a informação da função enzimática (*EC number*, discutido logo em seguida) das enzimas de um determinado genoma abre a possibilidade de reconstrução de rotas bioquímicas/metabólicas completas, sendo que um único *gap* em alguma rota pode indicar uma anotação equivocada ou um gene ainda não anotado (GINSBURG, 2009).

Extensos estudos ainda são realizados com enzimas na busca de uma classificação funcional (FREILICH et al., 2005), bem como suas participações em uma ou mais rotas metabólicas e conservação da sequência (PEREGRIN-ALVAREZ et al., 2003). Analisando a distribuição filogenética de enzimas de *Escherichia coli*, Peregrin-Alvarez et al. (2003) relatam que, embora as enzimas sejam amplamente distribuídas e altamente conservadas durante a evolução, sua participação nas rotas metabólicas podem variar significativamente.

Contudo, em 1956, o presidente da União Internacional de Bioquímica estabeleceu uma Comissão Internacional sobre Enzimas com o objetivo de resolver problemas relacionados à classificação e nomenclatura das mesmas. Foi então aprovado em 1961 um relatório com unidades, símbolos e nomenclatura para enzimas, no qual cada classe de enzima é subdividida e cada enzima contém um código único de quatro dígitos, chamado de "*Enzyme Commission number*" ou "*EC number*" (TIPTON e BOYCE, 2000).

A respeito do *EC number*, sendo as enzimas classificadas de acordo com o tipo de reação que realizam, o primeiro dígito (classe) define o tipo de reação geral catalisada, com valores de 1 a 6, constituindo as reações de oxidoredutase, transferases, hidrolases, liases,

isomerases e ligases, respectivamente. O segundo e terceiro dígitos indicam a subclasse e sub-subclasse, respectivamente; nessas subclasses e sub-subclasses, geralmente, encontra-se diversas especificações como grupos químicos de atuação da enzima e o produto a ser formado, respectivamente. O quarto dígito é um número identificador da enzima dentro de uma determinada sub-subclasse.

Atribuir um código *EC number* a uma enzima está longe de ser uma tarefa trivial, tanto computacionalmente quanto experimentalmente. Por vias computacionais, realiza-se uma análise de similaridade das sequências de aminoácidos da enzima desconhecida contra um banco de dados. Ao obter uma alta taxa de similaridade com alguma sequência deste banco de dados, as anotações da função enzimática (*EC number*) desta sequência serão transferidas para a nova enzima analisada (sendo este procedimento discutido mais à frente, na seção 1.3). Por vias experimentais (moroso e muito mais complexo, porém com maior exatidão), a enzima deve ser purificada e uma série de análises bioquímicas devem ser realizadas (determinação do pH e da temperatura de maior atividade, necessidade de cofatores, velocidade da reação, entre outros) a fim de definir o tipo de substrato (ou substratos) que tal enzima consegue degradar além de determinar o produto formado. Vários bancos de dados como KEGG (KANEHISA et al., 2002), BRENDA (SCHOMBURG et al., 2002), ExplorEnz (McDONALD et al., 2009), Uniprot (The UniProt Consortium) e EcoCyc (KARP et al., 2000) fornecem informações sobre rota metabólica e a reação enzimática nos processos celulares. Tais informações são primordiais para desenvolvimento de novos produtos biotecnológicos e para pesquisas mais detalhadas, como análise de fluxo de metabólitos, engenharia metabólica e biologia sintética.

Ressalta-se que uma boa anotação do genoma de uma espécie é de extrema importância para toda a comunidade científica, pois permite realizar e fundamentar pesquisas na área computacional (possibilitando a melhoria de sistemas de predição), assim como pesquisas com foco em novos produtos biotecnológicos (engenharia metabólica e biologia sintética). Enzimas bem anotadas formam um grupo funcional ideal para estudos de mudanças fenotípicas e divergência/redundância funcional nas espécies (FREILICH et al., 2005).

## 6. CONCLUSÃO

As performances médias de predição (sobre o conjunto de validação) dos modelos que compõe a ferramenta mAppLe foram de 94% de acurácia, 44% de taxa de falsa descoberta, 67% de F-*score*, 98% de sensitividade, 93% de especificidade e 0,69 para coeficiente de correlação de Matthews. A taxa de acerto de classificação correta da ferramenta mAppLe foi de 76,4%. Com uma abordagem completamente diferenciada, esta ferramenta poderá superar os problemas encontrados por outros programas (por se basearem na similaridade de sequências). Entretanto, futuras melhorias ainda serão feitas no mAppLe, de modo a ampliar as rotas metabólicas a serem preditas, aumentar a performance de predição e inferir a função enzimática das amostras analisadas.

## 7. REFERÊNCIAS


ALBERTS, B., JOHNSON, A., LEWIS, J., MORGAN, D., RAFF, M., ROBERTS, K., WALTER, P. Cell chemistry and bioenergetics. In: Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., Walter, P. **Molecular biology of the cell**. New York: Garland Science, Taylor and Francis group, 6th edition, 2015, p. 51-73.

ANGERMUELLER, C., PÄRNAMAA, T., PARTS, L., STEGLE, O. Deep learning for computational biology. **Molecular Systems Biology**, v. 12, p. 878, 2016.

ARAKAKI, A.K., HUANG, Y., SKOLNICK, J. EFICAz2: enzyme function inference by a combined approach enhanced by machine learning. **BMC bioinformatics**, 10:107, 2009.

AZUAJE, F. Computational models for predicting drug responses in cancer research. **Briefings in Bioinformatics**, v. 18, 820-829, 2017.

BULJAN, M., BATEMAN, A. The evolution of protein domain families. **Biochemical Society Transactions**, v. 37, p. 751-755, 2009.

CLAUDEL-RENARD, C., CHEVALET, C., FARAUT, T., KAHN, D. Enzyme-specific profiles for genome annotation: PRIAM. **Nucleic Acids Research**, 31, p. 6633-6639, 2003.

DOBSON, P.D., DOIG, A.J. Predicting Enzyme Class From Protein Structure Without Alignments. **Journal of Molecular Biology**, v. 345, p. 187-199, 2005.

FABRIS, F., MAGALHÃES, J.P.F., ALEX, A. A review of supervised machine learning applied to ageing research. **Biogerontology**, v. 18, p. 171-188, 2017.

FILHO, M.A.C.P.P. Metagenômica: princípios e aplicações. In: Faleiro, F.G., Andrade, S.R.M., Reis Júnior, F.B. Biotecnologia: estado da arte e aplicações na agropecuária. Planaltina-DF: Embrapa Cerrados, 2011, p. 174-193.

FREILICH, S., SPRIGGS, R.V., GEORGE, R.A., AL-LAZIKANI, B., SWINDELLS, M., THORNTON, J.M. The complement of enzymatic sets in different species. **Journal of Molecular Biology**, v. 349, p. 745-763, 2005.

FRIEDBERG, I. Automated protein function prediction - The genomic challenge. **Briefings in Bioinformatics**, v. 7, p. 225-242, 2006.

GINSBURG, H. Caveat emptor: limitations of the automated reconstruction of metabolic pathways in Plasmodium. **Trends in Parasitology**, v. 25, p. 37-43, 2009.

GLASNER, M.E., GERLT, J.A., BABBITT, P.C. Evolution of enzyme superfamilies. **Current Opinion in Chemical Biology**, v. 10, p. 492-497, 2006.

GUTERL, J.K., GARBE, D., CARSTEN, J., STEFFLER, F., SOMMER, B., REISSE, S., PHILIPP, A., HAACK, M., RÜHMANN, B., KOLTERMANN, A., KETTLING, U., BRÜCK, T., SIEBER, V. Cell-free metabolic engineering: Production of chemicals by minimized reaction cascades. **ChemSusChem**, v. 5, p. 2165-2172, 2012.

GUYON, I., ELISSEEFF, A. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, v.3, p.1157-1182, 2003.

HECKMANN, D., SCHLÜTER, U., WEBER, A.P.M. Machine Learning Techniques for PredictingCrop Photosynthetic Capacity from Leaf Reflectance Spectra. **Molecular Plant**, v. 10, p. 878-890, 2017.

JIANG, Y., ORON, T.R., CLARK, W.T., BANKAPUR, A.R., D'ANDREA, D., LEPORE, R., FUNK, CHRISTOPHER, S., KAHANDA, I., VERSPOOR, K.M., BEN-HUR, A., et al.. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. **Genome biology**, 17:184, 2016.

KANDOI, G., ACENCIO, M.L., LEMKE, N. Prediction of druggable proteins using machine learning and systems biology: A mini-review. **Frontiers in Physiology**, v. 6, p. 366, 2015.

KANEHISA, M., GOTO, S., KAWASHIMA, S., NAKAYA, A. The KEGG databases at GenomeNet. **Nucleic Acids Research**, v. 30, p. 42-46, 2002.

KARIMPOUR-FARD, A., EPPERSON, L.E., HUNTER, L.E. A survey of computational tools for downstream analysis of proteomic and other omic datasets. **Human Genomics**, v. 9 (1), p. 28, 2015.

KARP, P.D., RILEY, M., SAIER, M., PAULSEN, I.T., PALEY, S.M.; PELLEGRINI-TOOLE, A. The EcoCyc and MetaCyc databases. **Nucleic Acids Research**, v. 28, p. 56-59, 2000.

KORMAN, T.P., Opgenorth, P.H., Bowie, J.U. A synthetic biochemistry platform for cell free production of monoterpenes from glucose. **Nature Communications**, v. 8, p. 15526, 2017.

LARRAÑAGA, P., CALVO, B., SANTANA, R., BIELZA, C., GALDIANO, J., INZA, I., LOZANO, J.A., ARMAÑANZAS, R., SANTAFÉ, G., PÉREZ, A., ROBLES, V. Machine learning in bioinformatics. **Briefings in Bioinformatics**, v. 7, p. 86-112, 2006.

LIBBRECHT, M.W., NOBLE, W.S. Machine learning applications in genetics and genomics. **Nature Reviews Genetics**, v. 16, p. 321-332, 2015.

LODISH, H., BERK, A., MATSUDAIRA, P., KAISER, C.A., KRIEGER, M., SCOTT , M.P., ZIPURSKY, S.L., DARNELL, J. Protein structure and function. In: Lodish, H., Berk, A., Matsudaira, P., Kaiser, C.A., Krieger, M., Scott , M.P., Zipursky, S.L., Darnell, J., Molecular Cell Biology. W.H. Freeman and Company, 5th edition, 2004,p. 59-99.

MALHIS, N., WONG, E.T.C., NASSAR, R., GSPONER, J. Computational Identification of MoRFs in Protein Sequences Using Hierarchical Application of Bayes Rule. **PlosOne**, v. 10, p. 1-15, 2015.

McDONALD, A.G., BOYCE, S. AND TIPTON, K.F. ExplorEnz: the primary source of the IUBMB enzyme list. **Nucleic Acids Research**, v. 37, p. 593-597, 2009.

NELSON, D.L.; COX, M.M. Enzimas. In: NELSON, D.L.; COX, M.M. Lehninger Principles of Biochemistry. New York: W. H. Freeman and Company, 5th edition, 2011. p. 183-234

NIELSEN, J., FUSSENEGGER, M., KEASLING, J., LEE, S.Y., LIAO, J.C., PRATHER, K., PALSSON, B. Engineering synergy in biotechnology. **Nature Chemical Biology**. v. 10, p. 319-322, 2014.

ORTH, J.D., THIELE, I., PALSSON, B.Ø. What is flux balance analysis? **Nature Biotechnology**. v. 28, p. 245-248, 2010.

PEREGRIN-ALVAREZ, J., MANUEL TSOKA, S., OUZOUNIS, C.A. The phylogenetic extent of metabolic enzymes and pathways. **Genome Research**, v. 13, p. 422-427, 2003.

PIREDDU, L., SZAFRON, D., LU, P., GREINER, R. The Path-A metabolic pathway prediction web server. **Nucleic Acids Research**, v.34, p.714-719, 2006.

PLANES, F.J.,BEASLEY, E. Path finding approaches and metabolic pathways. **Discrete Applied Mathematics**. v.157, p. 2244-2256, 2009.

POPTSOVA, M.S., GOGARTEN, J.P. Using comparative genome analysis to identify problems in annotated microbial genomes. **Microbiology**, v.156, p.1909-1917, 2010.

QUESTER, S., SCHOMBURG, D. EnzymeDetector: an integrated enzyme function prediction tool and database. **BMC Bioinformatics**, 12:376, 2011.

ROST, B. Enzyme function less conserved than anticipated. **Journal of Molecular Biology**, v. 318, p. 595-608, 2002.

SAEYS, Y., INZA, I., LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. **Bioinformatics**, v. 23, p. 2507-2517, 2007.

SCHOMBURG, I., CHANG, A. SCHOMBURG, D. BRENDA, enzyme data and metabolic information. Nucleic Acids Research, v. 30, p. 47-49, 2002.

SCHREIBER, F. Visual comparison of metabolic pathways. **Journal of Visual Languages and Computing**. v. 14, p.327-340, 2003.

STEPHANOPOULOS, G., KEASLING, J., GONZALEZ, R. Metabolic Engineering and Synthetic Biology in Strain Development. **ACS Synthetic Biology**, v. 1, p. 491-492, 2012.

TANIGUCHI, H., OKANO, K., HONDA, K. Modules for *in vitro* metabolic engineering: Pathway assembly for bio-based production of value-added chemicals. Synthetic and Systems **Biotechnology**, v. 2, p. 65-74, 2017.

THE UNIPROT CONSORTIUM. UniProt: the universal protein knowledgebase, **Nucleic Acids Research**, v. 45, p. 158-169, 2017.

TIAN, W., ARAKAKI, A.K., SKOLNICK, J. EFICAz: A comprehensive approach for accurate genome-scale enzyme function inference. **Nucleic Acids Research**. v. 32, p. 6226-6239, 2004.

TIPTON, K., BOYCE, S. History of the enzyme nomenclature system. **Bioinformatics**, v. 16, p. 34-40, 2000.

VARSHAVSKY, R., GOTTLIEB, A., LINIAL, M., HORN, D. Novel unsupervised feature filtering of biological data. **Bioinformatics**, v. 22, p. 507-513, 2006.

WU, G., YAN, Q., JONES, J.A., TANG, Y.J., FONG, S.S., KOFFAS, M.A.G. Metabolic Burden: Cornerstones in Synthetic Biology and Metabolic Engineering Applications. **Trends in Biotechnology**, v. 34, p. 652-664, 2016.

YANG, J., YAN, R., ROY, A., XU, D., POISSON, J., ZHANG, Y. The I-TASSER Suite: protein structure and function prediction. **Nature Methods**, v. 12, p. 7-8, 2014.

ZHANG, L., TAN, J., HAN, D., ZHU, H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. **Drug Discovery Today**, 2017, In press. https://doi.org/10.1016/j.drudis.2017.08.010

ZHANG, Y.H.P., SUN, J., ZHONG, J.J. Biofuel production by *in vitro* synthetic enzymatic pathway biotransformation. **Current Opinion in Biotechnology**, v. 21, p. 663-669, 2010.