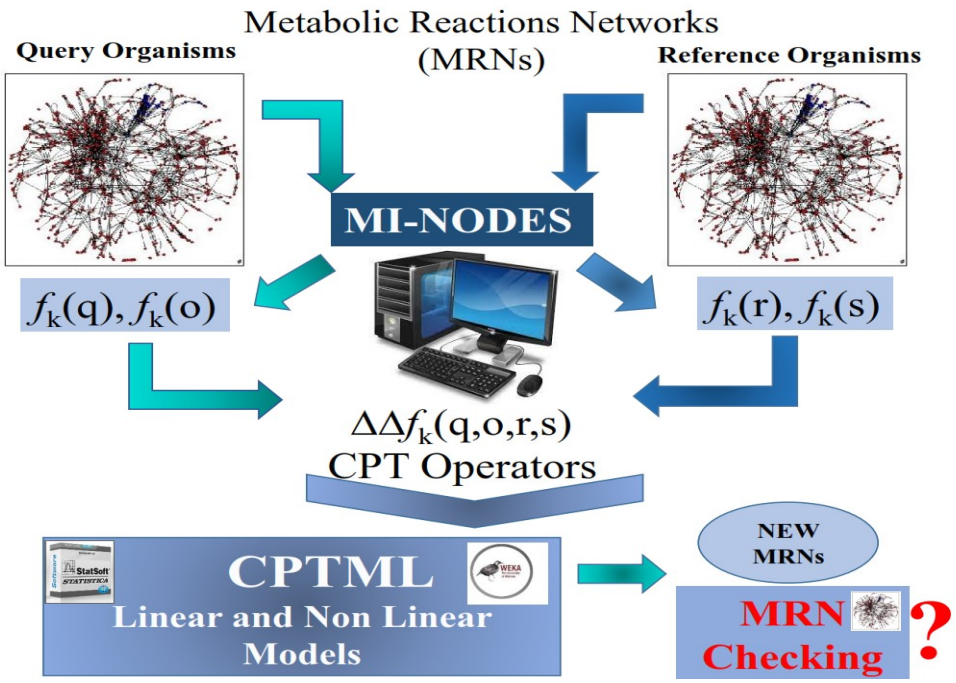
Combinatorial Perturbation-Theory Machine Learning (CPTML) Models for Curation of Metabolic Reaction Networks

Karel Diéguez-Santana,[a] Gerardo M. Casañola-Martin,[b,c] James R. Green [b], Bakhtiyor Rasulev [c], and Humberto González-Díaz [a,d,e,]*

[a] Department of Organic and Inorganic Chemistry, University of Basque Country UPV/EHU, 48940 Leioa, Spain.
[b] Department of Systems and Computer Engineering, Carleton University, K1S 5B6, Ottawa, ON, Canada.
[c] Department of Coatings and Polymeric Materials, North Dakota State University, Fargo, ND, 58102, USA
[d] BIOFISIKA: Basque Center for Biophysics, University of Basque Country UPV/EHU, 48940 Leioa, Spain.
[e] IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Biscay, Spain.

Doi: 10.3390/mol2net-07-11195

Graphical Abstract



Abstract. Metabolic Reaction Networks (MRNs) are complex networks produced by thousands of chemical reactions or transformations (links) of metabolites (nodes) in a live organism. An essential goal of chemical biology is to test the connectivity (structure) of these complex MRNs models presented for new microorganisms with promising features. In theory, we can undertake hands-on testing (Manual Curation). However, due to the large number of possible combinations of node pairs, this is a difficult operation (possible metabolic reactions). We combined Perturbation Theory, and Machine Learning approaches in this study to find a CPTML model for MRNs>40 organisms compiled by Barabasis' group. First, we used a novel type of node index termed Markov linear indices fk to quantify the local structure of a very large collection of nodes in each MRN. Next, for over 150 000 MRN query and reference node combinations, we computed CPT operators. Finally, we fed these CPT operators into several ML algorithms. The CPTML linear model obtained using the LDA algorithm is capable of distinguishing nodes (metabolites) with correct reaction assignment from nodes with incorrect reaction assignment with accuracy, specificity, and sensitivity values ranging from 85 to 100 % in both the training and external validation data series. Meanwhile, the top three non-linear models with more than 97.5 % accuracy were found to be PTML models based on Bayesian networks, J48-Decision Tree, and Random Forest algorithms. The new work sets the door for the investigation of MRNs from various organisms using PTML models. Finally, the new CPTML could be a useful tool for determining the structure of MRNs in new species in biotechnology.

The main bibliographic sources used in this paper are listed below **[1-10].**

References

1. Bornholdt, S.; Schuster, H.G. *Handbook of Graphs and Complex Networks: From the Genome to the Internet*; WILEY-VCH GmbH & CO. KGa.: Wheinheim, 2003.

2. Duardo-Sanchez, A.; Gonzalez-Diaz, H.; Pazos, A. MIANN Models of Networks of Biochemical Reactions, Ecosystems, and US Supreme Court with Balaban-Markov Indices. *Current Bioinformatics* **2015**, *10*, 658-671, doi:10.2174/1574893610666151008012752.
3. Duardo-Sanchez, A.; Gonzalez-Diaz, H.; Pazos, A. MI-NODES Multiscale Models of Metabolic Reactions, Brain Connectome, Ecological, Epidemic, World Trade, and Legal-Social Networks. *Current Bioinformatics* **2015**, *10*, 692-713.
4. Gonzalez-Diaz, H.; Arrasate, S.; Gomez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J.M. General Theory for Multiple Input-Output Perturbations in Complex Molecular Systems. 1. Linear QSPR Electronegativity Models in Physical, Organic, and Medicinal Chemistry. *Current Topics in Medicinal Chemistry* **2013**, *13*, 1713-1741.
5. Gonzalez-Diaz, H.; Duardo-Sanchez, A.; Ubeira, F.M.; Prado-Prado, F.; Perez-Montoto, L.G.; Concu, R.; Podda, G.; Shen, B. Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers. *Curr Drug Metab* **2010**, *11*, 379-406.
6. Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z.N.; Barabasi, A.L. The large-scale organization of metabolic networks. *Nature* **2000**, *407*, 651-654.
7. Witten, H.I.; Frank, E. *Data Mining: Practical machine learning tools and techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, USA, 2005.
8. Diéguez-Santana, K.; Casañola-Martin, G.M.; Green, J.R.; Rasulev, B.; González-Díaz, H. Predicting Metabolic Reaction Networks with Perturbation-Theory Machine Learning (PTML) Models. *Current Topics in Medicinal Chemistry* **2021**, *21*, 819-827, doi:10.2174/1568026621666210331161144.
9. Diéguez-Santana, K.; Rivera-Borroto, O.M.; Puris, A.; Pham-The, H.; Le-Thi-Thu, H.; Rasulev, B.; Casañola-Martin, G.M. Beyond Model Interpretability using LDA and Decision Trees for α-Amylase and α-Glucosidase Inhibitor Classification Studies. *Chemical Biology & Drug Design* **2019**, doi:10.1111/cbdd.13518.
10. Diéguez-Santana, K.; González-Díaz, H. Towards machine learning discovery of dual antibacterial drug–nanoparticle systems. *Nanoscale* **2021**, doi:10.1039/d1nr04178a.