CHEMBIOINFO-07: EJIBCE & Chem-Bioinfo. Congress Coimbra, Portugal-München, Germany-Ch. Hill, USA, 2021

Big Data Database Information Fusion Problem in AI-guided Drug Discovery Full Product Life Cycle Analysis

Viviana F. Quevedo-Tumailli[a, b] *Bernabé Ortega-Tenezaca,*[a,b] *and Humberto González-Díaz* [c,d,e,*]

[a] *RNASA-IMEDIR, Department of Computer Science, Faculty of Informatics, University of A Coruña (UDC), 15071, A Coruña, Spain.*
[b] *Universidad Estatal Amazónica UEA, Puyo, Pastaza, Ecuador.*
[c] *Department of Organic and Inorganic Chemistry, University of Basque Country UPV/EHU, 48940 Leioa, Spain.*
[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-07-11215

Abstract. Artificial Intelligence/Machine Learning (AI/ML) guided drug discovery is an interesting strategy to reduce costs in Drug discovery, Vaccine design, Nanoparticle-drug delivery systems assembly, Biomarkers validation, etc. Until now, most efforts of applying AI/ML techniques in this area focus on one of the phases of product development and not in the full Product Life Cycle (PLC). PLC approaches are of the major importance towards a rational design and sustainability of the final production process. In Pharmaceutical Industry context PLC approach have also multiple specific phases ranging from chemical synthesis/isolation of molecular entities to preclinical studies and preliminary exploratory clinical studies (phase 0) to clinical studies (phase I, II, III) and pharmaco-epidemiology and post-marketing studies (phase IV) in real population. Consequently, integral Product Life Cycle (PLC) should incorporate analysis of all or at least various of these phases. However, relevant information for different phases of the PLC, most of the time, may be disperse on different databases. On this situation also emerge multiple cases of contradictory, incomplete, highly variable, sparse, over/under represented, large volume sub-sets of information. In addition, the information available has multiple labels or assay boundary conditions. Some of these conditions are continuous variables like dose, temperature, time of assay, multiple values pharmacological parameters (Ki, IC50, MIC, etc.). Nevertheless, many of these conditions are non-ordinated numeric labels.

We can identify denote these conditions as cj. We are talking, for instance, of c0 = label of property measured (Ki, IC50, MIC, etc.), c1 = name of target protein, c2 = cell line, c3 = tissue, c4 = organism of assay, c5 = shape of nanoparticle, c6 = type of clinical assay, c7 = gender of patients, etc. In addition, many of these variables may be co-linear, co-dependent, or nested somehow among forming complex networks of interrelationships. For instance, we can measured the same set of parameters c0 to different drug for a subset of target proteins c1 expressed some of them in different tissues c3 of multiple organisms of assay c4, etc. This can be represented as a complex network of interconnections of these labels. Yet another point, usually these conditions can be managed as ontologies associated to an ontology dictionary cj = c0, c1, c2, c3, ... cn of deep n. Each one of these ontologies may have many levels or terms. For instance, organisms c4 may be multiple, eg.; human, mouse, rat, rabbit, etc. One last point, many of the instances of the dataset (not only the input variables) are complex systems (formed by sub-systems) with a network-like internal structure. We can see here structure as all the parts of the sub-system, the labels of these parts, the properties of weights of these parts, and the interconnection or links between these parts. This is for instance the case of drugs, proteins, metabolic networks, brain, etc. They all can be seen as sub-systems represented as molecular graphs of interconnected atoms, or protein structure network of interconnected aminoacids, metabolic network of interconnected reactions, etc. These graphs/networks may be constructed at different levels. For instance, the protein may be a network of atoms or a network of aminoacids, the brain may be seen as a network of neurons or a network of cortex regions. Also a population of patients in a sexual disease transmission network or flu epidemic break may be represented as a network of personal contacts or a network of towns. Due to the high amount and complexity of the information to be analyzed in a full/partial PLC analysis in this area this can be seen as a genuine Big Data problem. One approach to this problem may be the use of AI/ML as we mentioned at the beginning.

However, the use of these methods from a PLC point of view implies the use of Information Fusion (IF) techniques to pre-process all the information from different sources and put all the pieces together in a single dataset susceptible of analysis by AI/ML method. In this context, we have proposed Information Fusion, Perturbation Theory, and Machine Learning (IFPTML) method for PLC analysis in Pharmaceutical industry. IFPTML (IF + PT + ML) have three phases. The first phase carry out the IF of all the previous information. The second phase calculate PT operators able to numerically codify and compact all information treated in IF phase related to labels, ontology, network-like structures, etc. Last the ML phase develops the ML model and implement it in a user-friendly software.

References

1. Quevedo-Tumailli V, Ortega-Tenezaca B, González-Díaz H. IFPTML Mapping of Drug Graphs with Protein and Chromosome Structural Networks vs. Pre-Clinical Assay Information for Discovery of Antimalarial Compounds. Int J Mol Sci. 2021 Dec 2;22(23):13066.
2. Santana R, Zuluaga R, Gañán P, Arrasate S, Onieva E, Montemore MM, González-Díaz H. PTML Model for Selection of Nanoparticles, Anticancer Drugs, and Vitamins in the Design of Drug-Vitamin Nanoparticle Release Systems for Cancer Cotherapy. Mol Pharm. 2020 Jul 6;17(7):2612-2627. doi: 10.1021/acs.molpharmaceut.0c00308. Epub 2020 Jun 8.
3. Ortega-Tenezaca B, González-Díaz H. IFPTML mapping of nanoparticle antibacterial activity vs. pathogen metabolic networks.Nanoscale. 2021 Jan 21;13(2):1318-1330. doi: 10.1039/d0nr07588d.
4. Westkämper, E (2000). "Live Cycle Management and Assessment: Approaches and Visions Towards Sustainable Manufacturing". CIRP Annals. 49 (2): 501–522. doi:10.1016/s0007-8506(07)63453-2.