International Conference of Science, Engineering & Environmental Technology (ICONSEET), 2(2): 9-15, 2017
ISSN 0794-9650

www.repcomseet.com

Application of machine learning in prediction of bioactivity of molecular compounds: A review

Olutomilayo Olayemi Petinrin[1]* and Kemi Olatunbosun[2]
[1] Department of Information Systems, Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Malaysia.
[2] Department of Computer Science, Federal Polytechnic, Ede, Nigeria
Email:Olutomilayo.petinrin@gmail.com

Abstract: The use and implementation of machine learning has been majorly seen in chemoinformatics, drug discovery and development and especially bioactive molecule prediction. It has been shown that due to the inability of simple tools to handle the majorly increasing amount of data, machine learning comes to the rescue with its ability to handle high dimensional data, either with homogeneous or heterogeneous molecular structure. This study therefore reviews the current use, application, and principles of machine learning in chemoinformatics and drug discovery. It is discovered that the several machine learning methods or classifiers perform differently under the influence of several conditions. Also, no classifier can be said to claim superiority over another since they all perform differently depending on the dataset involved, and the classification process involved. The pharmaceutical industry will benefit more if better classifiers with better performance are implemented and this can be achieved with combinations of classifiers.

Keywords: Bioactivity, Chemoinformatics, Machine learning methods, Fingerprints, Herbal medicines, Pharmacology.

1. Introduction

Pharmacologically active molecules are known as bioactive molecules. These bioactive molecules are gotten either through the synthesis of chemical compounds or from natural sources. Previous attempts at discovering bioactive molecules in chemical compounds uses the traditional searching method, but due to the recent progress in computational chemistry, chemoinformatics and various aspects of science, the way by which bioactive molecules are discovered and classified has been modified and improved to aid better prediction and discovery in a large database of compounds. Due to these various developments in the discovery of bioactive molecules and also combinatorial chemistry, the topography of the modern day scientific discovery of bioactive molecules have been transformed, and this thereby makes it readily available for consumers use (Kang et al. 2013).

The application of informatics science to diverse scientific fields has resulted in the invention of several other disciplines such as Chemoinformatics, Bioinformatics, Neuroinformatics, Geoinformatics, Laboratory informatics, Social informatics, and Health informatics as the scientific field to which it was applied might be. The full optimization of information science and technology (IT) has become an integral solution to the drug discovery process, including and also a solution to various chemical-related problems. Chemoinformatics as a discipline, therefore combines various information resources in order to extract information from data and thereby derive knowledge from the extracted information. This derived information thus aids in making faster and better decisions in drug lead organization and identification area (Aktar et al. 2008).

The capability to use biological descriptors to illustrate small molecules has significantly developed to include different sets of data from biochemical assays to clinical adverse events, cellular phenotypes and protein expression and/or gene. In spite of large quantities of bioactivity data availability in public and corporate spheres, integration of these data and easy access to them will be significant for their application in molecular signatures. Ideally, for the data to be effectively utilized, the integration and identification of appropriate experimental data from diverse sources must need minimum or no manual involvement requiring assay data and metadata to be presented in a computer-readable format (Wassermann et al. 2015).

2. Chemoinformatics

Activities carried out via computer or performed on computer is termed "*In silico*". One of the *in silico* technologies used in researches relating to Chemistry is chemoinformatics. Chemoinformatics is a utilization of informatics methods to proffer solution to the chemical problems related with structural identification, synthesis design and molecular design (Gasteiger & Engel 2003). It is interdisciplinary with application of Computer science, Chemistry, and Mathematics. In principle, three methods are included in Chemoinformatics: logic-based, data-based, and principle-based. They are applied at different levels and different stage in chemical study. Most activities which have been usually practice in the agrochemical industries and drug for years are represented by chemoinformatics. The concurrent growth in computer technology has offered an array of computational tools to empower a scientist to transform from data to information to knowledge. These tools not only comprise of procedures for experimental data analysis, but also calculated properties of molecules generation (Hann & Green 1999). The concept of knowledge information is depicted in Fig. 1.

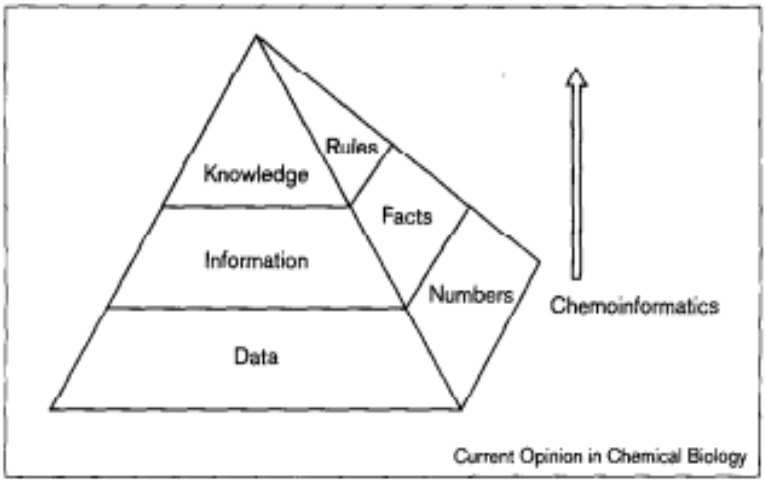

Figure 1: Concept of knowledge information (Hann & Green 1999)

The pharmaceutical aspect has benefitted greatly since researches and major developments in chemoinformatics are targeted on it, and majorly placing a primary focus of drug discovery. This has made the drug discovery aspect of chemoinformatics to be a booming field, which continues to develop till date not minding the heterogeneity associated with it. (Bajorath 2012).

Chemical structures molecules sometimes come in two-dimensional (2D) and three-dimensional (3D) structures. The art of applying computer methods to process the information which relates to this structure molecule is known as chemoinformatics. It follows the procedural pattern to knowledge discovery by transforming data into information, and further transform information into knowledge, which is hereby used in generating better ideas easily and making faster decisions based on the knowledge discovered for drug discovery. It journey into relevance as technological developments in biology and chemistry is because it offers various data mining tools which is effective and efficient in discovering new bioactive molecules (Willett n.d.2005).

In (Aktar et al. 2008), Chemoinformatics is defined as a collective term which comprises of designing, organizing, creating, retrieving, disseminating, analysing, visualizing, managing, and using of chemical information. It therefore involves the determination of those important parts of a molecular structure which is related to the needed properties for some given task. It is possible to differentiate the concerns associated with the atomic level of a drug design in a situation whereby its interaction with another entirely different molecule is of utmost importance and its physical characteristics related to absorption, distribution, metabolism and excretion (ADME). Molecular filters can be gotten by interacting with various diverse macromolecules. These filters consider only the molecular properties to create important models and calculate specific geometrical details. The discipline now has a widespread publicity and prominence so much that chemoinformatics as a topic now has to be introduced into chemistry curricula while some institutions makes it mandatory to offer chemoinformatics as a full curriculum in a bid to meet the pressing need for specialist in chemoinformatics. The representation of chemoinformatics is given in Fig. 2 in graphical form.

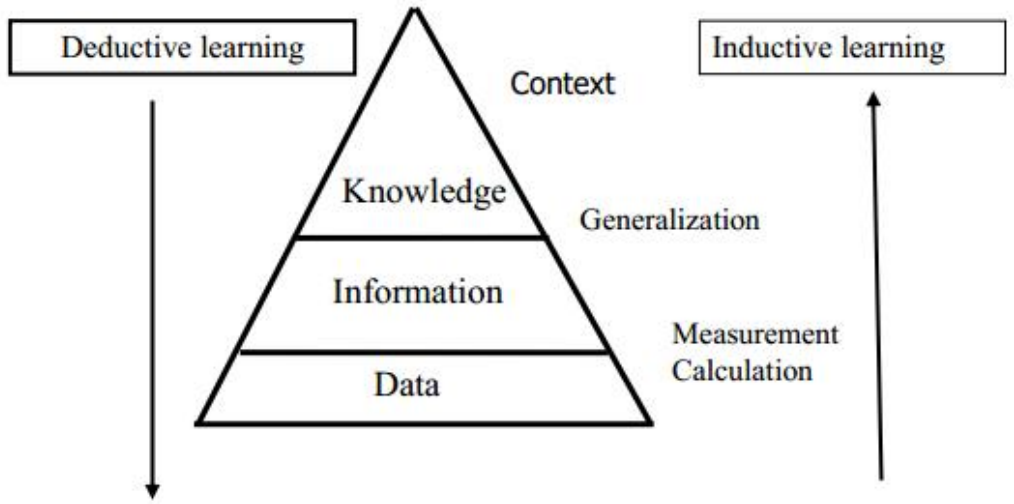

Fig. 2 Chemoinformatics representation (Aktar et al. 2008)

3. Machine Learning Methods

Machine learning methods is systematically having an extensive growing significance in the virtual screening and chemoinformatics field. Due to the increasing developmental researches being carried out on chemical and biological data, there is a consistent shift from using incapable old and traditional method to computerized approach such as machine learning methods and data mining as a significant approach for drug discovery. Asides the various statistical methods, many neural network (NN) methods such as Kohonen and counter-propagation networks, back-propagation network, and the machine learning methods such as Genetic algorithms (GAs) continue to demonstrate their worthiness in chemoinformatics application such as virtual screening and drug discovery. Support Vector Machine (SVM) which is an unsupervised method is being consistently recognized due to its ability to classify objects into two different classes as a function of their features. In (Srinivas Reddy, et al. 2007), a neural network model was developed to give an early evaluation of human cytochrome P450-mediated metabolism of drug-like compounds. The neural network model was built based on Kohonen learning technique which was unsupervised and also an already selected group of molecular descriptors.

Cytochrome P450 (CYP) isoforms has great importance in drug discovery since it aids in early detection of drugs. On this note, Hammann et al., (2009) had a classification carried out by experimenting with different machine learning method, in order to determine which method had the best accuracy, and three important isoforms were used in the classification. The Decision Tree had the best accuracy compared to

other classifiers tested. Cheng et al., (2011) on the other hand, claimed that the limitation of the models to three CYP isoforms is not effective in drug discovery research. Therefore, five important isoforms were used and the classifiers used were combined and fused with back propagation artificial neural network (BP-ANN). The resulting accuracy from the combined classifier was better than that of the single classifiers, although, compared with all the single classifiers tested, Support Vector Machine performed better. Machine Learning methods are effectively used to indicate grounded quantitative relations which exist between chemical structure and biological activity. In a bid to detect potentiators of metabotropic glutamate receptor 5 (mGluR5) as a compound with potentials of producing a novel treatment to combat the schizophrenia disease, (Butkiewicz et al. 2009) made a comparison between three different machine learning methods namely: Decision Trees (DT), Artificial Neural Networks (ANN), and Support Vector Machines (SVM). All the methods used a high predictive accuracy level although the Decision Tree method had the least predictive accuracy, while Support Vector Machine was able to produce a highest level of predictive accuracy. It is essential to note that the combination of Artificial Neural Network and Support Vector Machine was able to produce even better accuracy compared to the accuracy derived from each method as a stand-alone model. This shows that combination or ensemble of machine learning method can produce better accuracy.

Due to the better performance noticed from the combination of more than one machine learning methods, (Babajide Mustapha & Saeed 2016) combined Classification and Regression Tree (CART) and a variant of Gradient Boosting Machine to make a method known as extreme gradient boosting (Xgboost). It was used to predict the biological activity of compounds molecular structure based on the quantitative description. The resulting accuracy of the methods was better than that of other classifiers such as Support Vector Machine, Naïve Bayes, Random Forest and other classifiers with which the accuracy was compared. The method also showed effectiveness in predicting both homogeneous and heterogeneous dataset on which it was experimented.

Many examples of using machine learning techniques in the process of chemoinformatics and virtual screening have been presented [(Cheng et al. 2011); (Yan et al. 2013); (Hecht 2011)]. Several features of classification conditions and their influence on the machine learning techniques performance have been considered: one or two methods tested for the classification of compounds depicting activity towards several diverse targets (Agarwal et al. 2010) or different methods tested for one target (Hammann et al. 2009). Different molecular descriptors and fingerprints have been employed for molecular representation. For instance, (Plewczynski 2011) verified the performance of seven (7) machine learning techniques for five (5) different targets using regular atom pair descriptors, (Smusz et al. 2013) used MOLPRINT 2D circular fingerprints to examine the performance of the Inductive Logic Programming, Naïve Bayes (NB) classifier and Support Vector.

(Agarwal et al. 2010) employed the FP2 and MOLPRINT 2D fingerprints to demonstrate the usage of ranking techniques in virtual screening tasks. However, no paper presents the simultaneous influence of many factors on machine learning methods performance. Notwithstanding, (Smusz et al. 2013) used 11 machine learning techniques to examine the classification of ligands active towards five (5) different protein targets, for the training sets with two (2) distinct numbers of actives, using eight (8) fingerprints for the molecular representation, taking into account the time required for building a predictive model. However, the result shows that the effectiveness of each classifier always varies depending on classification process conditions.

4. **Prediction of Bioactivity in Herbal Medicines**

Herbal medicines (HMs) comprise the most complex chemical constituents. A trend in recent studies shows integral activity evaluation being popular for quality control of HMs or functional foods (FF). More researches focus more on the relationship which exist between bioactivity and chromatography/mass spectroscopy while easily ignoring the relationship with spectrum-activity.

To show a spectrum-activity study, (Ding et al. 2016) used the equivalent anti-inflammation activities from near infrared reflection spectra (NIRS) of a sample of Flos Chrysanthemi (FC) which were collected as a representative spectrum technology. Synergy interval partial least squares (siPLS) and partial least squares regression (PLSR) calibration models and twelve (12) various spectral pre-treatment techniques were then utilized for optimization of paramenters of the Q-markers in Flos Chrysanthemi (FC) powder. It was discovered thst bioactive strategy and the integrated NIRS was efficient for a quick quality management in Flos Chrysanthemi, and it is efficient for quality control of other botanical food.

Fingerprinting combined with chemometrics (Yang et al. 2016) and representative components determination (Gholivand et al. 2015) are two common methods to guarantee the efficacy and safety of HMs. However, researchers tend to be more concerned in the comprehensive effects of Herbal Medicines; and therefore, focus on countless constituents which leads to implausible deductions. Lack of correlation analysis with integral bioactivity evaluation is a major difficulty facing the quality management of herbal medicine (Shi et al. 2014). To overcome this issue, it is possible to make prediction of the comprehensive bioactivity using its chemical composition and to locate components which are active in herbal medicine using the study of quantitative composition-activity relationship (Froufe et al. 2011). Fingerprint-activity relationship analysis also is a technique which aids bioactivity which are germane to the quality control of Herbal Medicine. (Lucio-Gutiérrez et al. 2012). The advantages of Herbal Medicines are largely considered to produce effects through a collaborative method by multi-targets and multi-compounds (Wu et al. 2014). Therefore, some novel strategies were provided by (Long et al. 2015) in discovering major compounds which plays part in some pharmacological effects. Unfortunately, these systematic and logical methods were virtually exclusively centred on mass spectrometry or chromatography, and easily, the efficiency correlation analysis with the spectral information is ignored.

Near-infrared spectroscopy (NIRS) as a representative spectrum technology, has merits over other systematic and logical methods due to it being low-cost, used easily, and it is a fast technique which takes record of ranges of liquid and solid samples with large information and quantification of more than one components permitted (Ding et al. 2016). Also, for fast quality control in the food industry active pharmaceutical ingredients, and raw materials, NIRS is been used generally. Determining if the ingredients deduced by the spectral analysis can be associated with the function is the most critical concern presently. As a matter of fact, several researchers been conscious of the challenge, started applying NIRS in assessing the total capacity of antioxidant in the quinoa, green tea and apples (Schmutzler & Huck 2016). It was discovered that the concentration of the assessment index was mostly on the antioxidant capacity in vitro with PLSR, such as flavonoids, polyphenol and vitamin E. From the results, it is shown that antioxidant activity with these bioactive components can be anticipated with linear algorithm-PLSR (Froufe et al. 2009). However, immunoregulation and analgesic effects which are non-linear bioactivities were rarely been researched upon (Han et al. 2015).

5. Conclusion

The recent progresses in computational biology, bioinformatics and chemical genetics have changed the direction and way in which bioactive molecules are discovered and classified. The capability to use biological descriptors to illustrate small molecules has significantly developed to include different sets of data from biochemical assays to clinical adverse events, cellular phenotypes and protein expression and/or gene.

The various machine learning methods/techniques, their principles and application in drug development, herbal medicine, bioactivity of chemical compounds prediction and chemoinformatics as a broad area has been reviewed in this study. Chemoinformatics arises mainly as an upshot of technological developments in biology and chemistry, new bioactive molecules can be discovered using various machine learning methods and tools which are applied in data mining.

Many studies of using machine learning techniques in the process of chemoinformatics and virtual screening have been presented and several features of classification conditions and their influence on the machine learning techniques performance have been considered. However, the review shows that the effectiveness of each classifier always varies depending on classification process conditions. It is also shown that fingerprint-activity relationship analysis is a technique which makes bioactivity which is germane to the quality control of herbal medicine easier. Although analgesic effects and immunoregulation, which are non-linear bioactivities are rarely studied, antioxidant activity which has components that are bioactive can be expected together with linear algorithm-PLSR.

References

[1] A. Srinivas Reddy, S. Priyadarshini Pati, P. Praveen Kumar, H.N.P. and G.N.S., 2007. Virtual Screening in Drug Discovery - A Computational Perspective. Current protein & peptide science, 8(4), pp.329–351. Available at: http://www.eurekaselect.com/78566/article.

[2] Agarwal, S., Dugar, D. & Sengupta, S., 2010. Ranking chemical structures for drug discovery: A new machine learning approach. Journal of Chemical Information and Modeling, 50(5), pp.716–731. Available at: papers2://publication/uuid/62ECD148-B860-408F-81A1-27E9E23D352A.

[3] Aktar, W., Murmu, S. & Bengal, W., 2008. Chemoinformatics : Principles and Applications. Agricultural Chemistry, pp.1–28. Available at: http://www.shamskm.com/env/chemoinformatics-principles-and-applications.html.

[4] Babajide Mustapha, I. & Saeed, F., 2016. Bioactive Molecule Prediction Using Extreme Gradient Boosting. Molecules, 21(8), p.983. Available at: http://www.mdpi.com/1420-3049/21/8/983.

[5] Bajorath, J., 2012. Chemoinformatics: Recent advances at the interfaces between computer and chemical information sciences, chemistry, and drug discovery. Bioorganic & Medicinal Chemistry, 20(18), p.5316. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0968089612006736.

[6] Butkiewicz, M. et al., 2009. Application of machine learning approaches on quantitative structure activity relationships. In 2009 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. IEEE, pp. 255–262. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4925736.

[7] Cheng, F. et al., 2011. Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers. Journal of Chemical Information and Modeling, 51(5), pp.996–1011.

[8] Ding, G. et al., 2016. A rapid integrated bioactivity evaluation system based on near-infrared spectroscopy for quality control of Flos Chrysanthemi. Journal of Pharmaceutical and Biomedical Analysis, 131, pp.391–399. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0731708516305106.

[9] Froufe, H.J.C., Abreu, R.M. V & Ferreira, I.C.F.R., 2009. A QCAR model for predicting antioxidant activity of wild mushrooms. SAR and QSAR in environmental research, 20(5–6), pp.579–90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19916116.

[10] Froufe, H.J.C., Abreu, R.M. V & Ferreira, I.C.F.R., 2011. QCAR models to predict wild mushrooms radical scavenging activity, reducing power and lipid peroxidation inhibition. Chemometrics and Intelligent Laboratory Systems, 109(2), pp.192–196. Available at: http://dx.doi.org/10.1016/j.chemolab.2011.09.004.

[11] Gasteiger, J. & Engel, T., 2003. Chemoinformatics: A Textbook,

[12] Gholivand, M.B. et al., 2015. Combination of electrochemistry with chemometrics to introduce an efficient analytical method for simultaneous quantification of five opium alkaloids in complex matrices. Talanta, 131, pp.26–37.

[13] Hammann, F. et al., 2009. Classification of cytochrome P450 activities using machine learning methods. Molecular Pharmaceutics, 6(6), pp.1920–1926. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-71949083397&partnerID=40&md5=abf380aba7dd730eb2f5cbc3b10e4989.

[14] Han, Y. et al., 2015. Comparative evaluation of different cultivars of Flos Chrysanthemi by an anti-inflammatory-based NF-??B reporter gene assay coupled to UPLC-Q/TOF MS with PCA and ANN. Journal of Ethnopharmacology, 174, pp.387–395. Available at: http://dx.doi.org/10.1016/j.jep.2015.08.044.

[15] Hann, M. & Green, R., 1999. Chemoinformatics: A new name for an old problem? Current Opinion in Chemical Biology, 3(4), pp.379–383.

[16] He, J. et al., 2015. Characterization and machine learning prediction of allele-specific DNA methylation. Genomics, 106(6), pp.331–339. Available at: http://dx.doi.org/10.1016/j.ygeno.2015.09.007.

[17] Hecht, D., 2011. Applications of machine learning and computational intelligence to drug discovery and development. Drug Development Research, 72(1), pp.53–65.

[18] Kang, Y. et al., 2013. Bioactive molecules : current trends in discovery , synthesis , delivery and testing. International e-Journal of Science, Medicine & Education, 7(Suppl 1), pp.32–46.

[19] Long, F. et al., 2015. A strategy for the identification of combinatorial bioactive compounds contributing to the holistic effect of herbal medicines. Scientific reports, 5(January), p.12361. Available at: http://dx.doi.org/10.1038/srep12361.

[20] Lucio-Gutiérrez, J.R., Coello, J. & Maspoch, S., 2012. Enhanced chromatographic fingerprinting of herb materials by multi-wavelength selection and chemometrics. Analytica Chimica Acta, 710, pp.40–49. Available at: http://dx.doi.org/10.1016/j.aca.2011.10.010.

[21] Plewczynski, D., 2011. Brainstorming: Weighted voting prediction of inhibitors for protein targets. Journal of Molecular Modeling, 17(9), pp.2133–2141.

[22] Schmutzler, M. & Huck, C.W., 2016. Simultaneous detection of total antioxidant capacity and total soluble solids content by Fourier transform near-infrared (FT-NIR) spectroscopy: A quick and sensitive method for on-site analyses of apples. Food Control, 66, pp.27–37. Available at: http://dx.doi.org/10.1016/j.foodcont.2016.01.026.

[23] Shi, Z.Q. et al., 2014. Identification of effective combinatorial markers for quality standardization of herbal medicines. Journal of Chromatography A, 1345, pp.78–85. Available at: http://dx.doi.org/10.1016/j.chroma.2014.04.015.

[24] Smusz, S., Kurczab, R. & Bojarski, A.J., 2013. A multidimensional analysis of machine learning methods performance in the classification of bioactive compounds. Chemometrics and Intelligent Laboratory Systems, 128, pp.89–100. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0169743913001561.

[25] Wassermann, A.M. et al., 2015. The opportunities of mining historical and collective data in drug discovery. Drug Discovery Today, 20(4), pp.422–434. Available at: http://dx.doi.org/10.1016/j.drudis.2014.11.004.

[26] Willett, P., Chemoinformatics techniques for data mining in files of two-dimensional and three-dimensional chemical molecules .

[27] Wu, H.Y. et al., 2014. The synergetic effect of edaravone and borneol in the rat model of ischemic stroke. European Journal of Pharmacology, 740, pp.522–531. Available at: http://dx.doi.org/10.1016/j.ejphar.2014.06.035.

[28] Yan, A. et al., 2013. Classification of Aurora kinase inhibitors by self-organizing map (SOM) and support vector machine (SVM). European Journal of Medicinal Chemistry, 61, pp.73–83. Available at: http://dx.doi.org/10.1016/j.ejmech.2012.06.037.

[29] Yang, Y. et al., 2016. Quantitative and fingerprinting analysis of Pogostemon cablin based on GC-FID combined with chemometrics. Journal of Pharmaceutical and Biomedical Analysis, 121, pp.84–90. Available at: http://dx.doi.org/10.1016/j.jpba.2016.01.012..