Corresponding Author:
dr.syedrizwan@kiu.edu.pk

DARSGAH-E-AHLEBAIT

JOURNAL OF NATURAL SCIENCES [ISSN:2308-5061]

CHEMOINFORMATICS ALGORITHM "UNLOCKING NATURE'S POTENTIAL

MEHWISH BAHADUR[1] AND SYED RIZWAN ABBAS[2]

[1]Department of Animal Sciences; Zoology; Karakuram International University; Gilgit.
[2]Department of Plant Sciences; Karakuram International University; Gilgit.

ABSTRACT

Chemoinformatics, a broad field of study that focuses on manipulating chemical structural information while using computational techniques to solve chemical problems.The field combines expertise from, among others in physics,chemistry,biology,mathematics, statistics and informatics, has emerged as a significant field of computational molecular sciences. Chemoinformatics is a field that uses a wide variety of computational algorithms from computer science and mathmatics in addition to algorithms specifically designed for chemoinformatics-related problems. An algorithm is a series of steps that a machine must follow in order to carry out a mathematical operation. The term Artificial Intelligence (AI) was first used in 1955 to describe the idea that a machine could be taught to mimic human intellect. Although AI has not yet achieved its objective, advanced machine learning algorithms have brought us one step closer to it. Algorithms based on electroencephalography (EEG) that are used to design brain-computer interface (BCI) systems. It outlines the key characteristics of frequently used algorithms and gives an overview of them. The development of robust chemoinformatics algorithms is pivotal in areas such as drug discovery, material science, and chemical data analysis. This review article represents chemoinformatics in machine learning algorithm, genetic algorithm, evolutionary algorithm, optimization based algorithm, single objective and multi-objectives optimization, similarity searching and Dijsktra Algorithm.The revolutionary significance of Artificial Intelligence and Algorithm in transforming several procedures including machine learning and multi-task learning. This article reviews key algorithmic design paradigms and gives an overview of chemoinformatics algorithms and algorithmic complexity.

KEYWORDS: Chemoinformatics; Algorithm; Mathmatical operations; Artificial Intellegence (AI); Machine learning algorithms; Computational algorithm; Algorithm complexity.

INTRODUCTION

Frank K. Brown first developed the word "**chemoinformatics**" in 1998 with the goal of speeding up drug discovery and development. But , today, biochemistry, chemistry, and biology

all depend on chemoinformatics **(Brown, 2023).** By enabling computers to analyse data and make predictions without depending only establishing clear guidelines or formulas for mathematics, machine learning (ML) has brought about a paradigm change in comparison to previously well-established stand-alone models in the fields of statistics, mathematics, and physics. These algorithms are capable of finding intricate relationships and patterns in data on biological activity and 3D chemical structures, adapting their models in response to feedback, and generalising from training sets of data to produce precise predictions on unknown data. By using data-driven methods, it is now possible to optimise drug-target interactions, improve the accuracy and efficiency of molecular modelling, target-based drug discovery, chemical library screening, mechanics, dynamics, prioritising possible therapeutic candidates, and predicting possible toxicological reactions of biologics. The present status of research and the possible incorporation of ML-driven chemoinformatics are covered in this study.

EXPLORATION OF CHEMOINFORMATICS

Chemoinformatics, which uses Inductive learning to anticipate chemical processes, has become a powerful field in drug development at the nexus of chemistry and informatics **(Gasteiger J., 2023).** Researchers' ability to investigate, evaluate, and forecast the characteristics and actions of molecules has been completely transformed by the use of chemoinformatics using machine learning (ML). It has expedited the procedure tenfold in contrast to a few decades prior. The areas of emphasis include chemical space exploration, molecular engineering, molecular manipulation, scaffold analysis, pharmacophore, compound database searching, library design, and molecular graph mining (**Willett**, **2005**)**.**

`

FUNDAMENTAL OF CHEMOINFORMATICS

Based on chemical training data that is supplied as numerical representations or mathematical equations, machine learning models carry out prediction tasks. A sophisticated, multilayer computational procedure is involved in converting compound structures into chemical data that is appropriate for machine learning. The procedure includes creating descriptors, creating molecular graphs, creating fingerprints, analysing similarities, locating chemical spaces, simulating molecular dynamics, and more. Every layer is intricately linked to the ones that came before it, greatly impacting how the machine learning models read the chemical data and improving their ability to anticipate. Chemical databases are used in chemoinformatics to store and retrieve chemical data, which is necessary for training machine learning (ML) models. These databases make it possible to analyse big chemical datasets or search for particular substances. Chemical databases, which hold enormous volumes of information on chemicals, such as compound structures, biological activity, and other pertinent physiochemical aspects, are crucial for managing and utilising during the training process of models. For the purpose of target prediction, these databases make data mining, knowledge discovery, and information retrieval easier. specialised databases of substances found in nature **(Rutz, et al., 2021).**

CHEMICAL DATA REPRESENTATION

Data must be converted into a computer-understandable format before models are trained on it due to developments in machine learning modelling and the abundance of chemical and biological data available. Molecular graphs, fingerprints, descriptors, and other tools can be used to represent empirical, molecular, and structural data in the context of chemical data representation**.** Chemoinformatics is a multidisciplinary field that analyses and understands the

behaviour and properties of molecules by fusing chemical knowledge with computational techniques. Large datasets of chemical compounds are processed and interpreted using machine learning techniques, which allows the prediction of their characteristics and potential applications **(Haghighatlari, et al., 2020) (David, Thakkar, Mercado, & Engkvist, 2020).**

In chemoinformatics, machine learning techniques are essential because they let researchers find patterns and connections in these datasets. Because they work well with complex chemical data, a number of algorithms, including k-Nearest Neighbours (kNN), Random Forests (RF), and Support Vector Machines (SVM) are commonly utilised. Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR) modelling projects, which seek to identify correlations between molecule structures and their characteristics, these methods are frequently utilised.
Chemoinformatics applications involve features for ADME (absorption, distribution, metabolism, and excretion) prediction, which are essential for drug research and discovery, demonstrate the significance of this field. The field has grown dramatically, as evidenced by the large number of articles and citations it has received in prestigious journals, which demonstrate the scientific community's growing acknowledgment and relevance of it **(Feinberg, 2020)**.

ALGORITHM AND ARTIFICIAL INTELLEGENCE

A machine follows an algorithm, which is a series of sequential instructions, to perform a mathematical operation. The term artificial intelligence (AI) was first used in 1955 to describe the idea that a machine could be taught to mimic human intellect. Although AI has not yet achieved its objective, advanced machine learning algorithms have brought us one step closer to it. **Muhammad ibn Musa al-Khwarizmi** was a Persian astronomer and mathematician, is credited with coining the term "**algorithm**" in the ninth century when he authored a treatise on arithmetic rules. In computer science and mathematics, an algorithm is a notion that specifies a limited set of instructions to offer a practical way to solve a certain problem. A well-designed algorithm, despite the word's intimidating connotations, will be composed of precisely specified statements arranged through order, selection, and repetition (including recursion) that abstract a problem definition and eliminate ambiguity **(DE., 1997).**

MACHINE LEARNING ALGORITHM

The scientific study of statistical models and techniques used by computer systems to carry out specified tasks without explicit programming is known as machine learning, or ML for short. Understanding algorithms for a variety of everyday applications. One of the reasons an online search engine like Google performs so well each time it is used to search the internet is because it has an algorithm that is constantly learning how to rank web sites. Predictive analytics, image processing, data mining, and other uses for these algorithms are just a few. The primary benefit of machine learning is its ability to enable algorithms to do tasks automatically once they have learnt how to handle data. A branch of "Multi-task learning" is the term for machine learning looks for ways to leverage task similarities in order to solve several tasks at once. This can serve as a regularize and increase learning efficiency. Formally speaking, using the knowledge found in every task, Multitask Learning (MTL) will help improve the learning of a particular model. This is due to the fact that traditional deep learning techniques seek to employ a single model to accomplish a particular job. In this scenario, tasks or a portion of them are related to one another but not precisely the same **(W. Richert).**

GENETIC ALGORITHM

Genetic algorithms, sometimes known as GAs, are optimisation methods derived from natural selection and genetics. Numerous disciplines, including biology, computer science, engineering, economics, and more, have made extensive use of them. Genetically inspired operators (i.e., selection, crossover, and mutation) are used by GAs to evolve a population of possible solutions across generations **(Cheng et al., 2016).** Genetic algorithms are a viable substitute for conventional optimisation methods as they employ directed random searches to identify the best solutions in intricate environments. We review current topics in GA theory and application and demonstrate the science and art of genetic algorithms. Rather than presenting an in-depth analysis, We give a quick synopsis of the confusing field of GA research. Firstly, we make the comparison between natural search procedures and genetic algorithms. Next, we explain how GAs operate and the genetic algorithm that Holland first presented in 1975.

ALGORITHMIC COMPLEXITY

Algorithmic complexity, sometimes referred to as computational complexity, is a discipline in computer science that analyses the resources necessary for algorithms to solve a particular computational task. The primary resources evaluated are time (time complexity) and space (space complexity). Computing complexity is the study of grouping computing issues based on the degree of difficulty they have and connecting those groups. The complexity of computational algorithms varies depending on the kinds and quantity of instructions used to carry out the algorithm till it is successfully terminated. This section provides an introduction to algorithmic complexity notation and discusses some of the more challenging computational problems that are commonly related to chemoinformatics algorithms.

The worst-case situation for expressing the upper bound of an algorithm's space or running time requirements is provided by Big O notation **(DE., 1997).** The conventional method of characterising algorithmic complexity is Big O notation, also known as Big Oh notation, Landau notation, Bachmann–Landau notation, and asymptotic notation. It is employed to specify how quickly algorithms evolve as input gets closer to infinity. Generally, the algorithmic complexity upper bound is all that the big O notation indicates. In order to successfully complete the algorithm in a more realistic runtime window, factorial, exponential, and quadratic complexity algorithms usually require a large amount of optimisation or a redesign of the algorithm to fit a new design paradigm **(DE, 1997).**

DESIGN PARADIGAMS

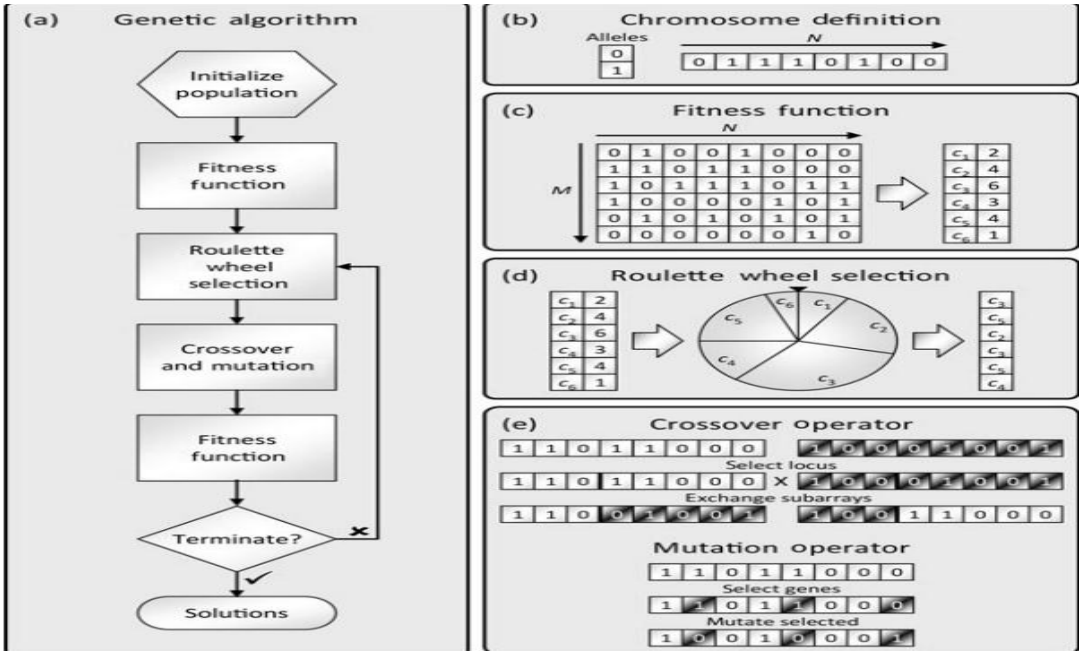
Object-Oriented Design (OOD) is one well-known design paradigm. OOD places a strong emphasis on organising software as a collection of objects, each of which stands for an instance of a class and contains behaviour and data pertinent to the object. In software development, this paradigm encourages modularity, reusability, and extensibility **(Erich Gamma, 1994).**
There are numerous techniques for creating algorithms, and each one can be used to address a variety of chemoinformatics-related problems. These algorithmic paradigms rationalise the amount of computation performed and minimise any potential duplication that may be present but not immediately noticeable, which helps in the design of algorithms.

DYNAMIC OVERLAPPING

In mathematics and computer science, overlapping sub problems are a frequent problem where, as the name suggests, information related to other solutions is generated during the computation and evaluation of one answer. Dynamic programming was first described by its inventor, Richard Bellman, as "nesting smaller choice issues inside larger decision **(Bellman, 1957) .** Memorization is a technique used by dynamic programming to save expensive recomputation by recording sub-problems that have already been calculated. Runtimes can be greatly decreased by using these recorded sub-problems in a computationally efficient way. However, because of the stored sub-problems, there will be a decrease in memory efficiency, therefore this needs to be closely watched. Runtimes can be greatly decreased by using these recorded sub-problems in a computationally efficient way. However, because of the stored sub-problems, there will be a decrease in memory efficiency, therefore this needs to be closely watched.

EVOLUTIONARY ALGORITHM

As was previously said, there are numerous algorithms that, in practice, have very long computational runs (especially exponential and polynomial algorithms), necessitating the use of various heuristic techniques to lower the expense of computation. In chemoinformatics, Specifically, evolutionary algorithms (EAs) have proven to be quite helpful. Possibly the most popular class of nondeterministic algorithms are genetic algorithms (GAs), which are based on natural evolutionary processes **(Michie, 1968, 218:19–22) .**Fontain was the first to apply EAs in chemoinformatics **(Fontain, 1992)** . Almost any chemoinformatics issue that can be thought of has been solved with the use of EAs, including ligand docking, molecular design, library design, and model creation (see infra) **(Erich Gamma R., 1994).**



OPTIMIZATION BASED ALGORITHM

Representatives and diversity analysis

Computational chemistry and large-scale screening methods have made it possible to synthesise and evaluate a far greater number of drug-like molecules. Even yet, just a small portion of the chemical space that was available could be synthesised and tested. Consequently, logical methods for designing screening libraries—and especially libraries made up of varied compounds, or compounds that differ significantly from one another—have been developed as a result (**Cheng et al., 2016).**

Identified a number of diversity functions to formulate the task of choosing a varied a portion of a parent data set used as an optimisation issue. Specifically, using Equation, The square of the minimal distance, dij2, is calculated using the MaxMin function using all of the (i, j) pairs that comprise the selected set **(OptiSim, 1997)**

$$\text{Mini}\neq j(\text{dij2})) \quad (4)\ E4 = \text{MaxMin}$$

Derivation of predictive QSAR models

QSAR models are constructed by a variety of machine-learning techniques. Starting with a modelling set, the modelling process builds a model of the activity as a function of the descriptors through regression- or classification-based analysis. In this field, machine learning techniques are widely utilised due to their ability to handle extremely complicated interactions between activities and descriptors **(Cherkasov A., 2013)**

SINGLE OBJECTIVE AND MULTI-OBJECTIVES OPTIMIZATION

Numerous issues in the fields of material and chemoinformatics can be framed as optimisation challenges, as was previously mentioned. A target (or objective) function(s) (f) and a set of variables (X1, X2, X3, Xn) connected to the scientific subject of interest must be defined in order to formulate such a formulation. Together, these variables create a complex, multi-dimensional surface whose distribution of optima is unknown beforehand. Finding the global optimum—or, better still, a set of optima—is the next stage, since events in these domains are rarely under the control of a single solution **(Cheng, 2016).**

SIMILARITY SEARCHING

Finding the compounds most comparable to a certain chemical of interest (or probe) is an alternate strategy to looking up a complex library in order to find a specific defined substructure **(Willett, 1998)**. One of the most often used vector formats in chemoinformatics is the daylight fingerprint, which has several implementations, including the fingerprinting algorithm **(Brown, 2005).**

Efficiency

Large compound libraries can be quickly reduced to a smaller group of candidates who are more likely to display the necessary biological features.

Versatility

Since similarity searching is not restricted to any particular kind of molecule or activity, it can be widely used in a variety of chemistry and biology domains.

SHORTEST PATHS ATOM PAIRS (DIJKSTRA ALGORITHM)

In a graph, the Dijkstra's algorithm is a widely used technique for finding the shortest paths between nodes. In cheminformatics, this technique can be very helpful in figuring out the

shortest pathways between atoms in a molecule, which can aid in the analysis of molecular structures and properties **(Dijkstra, 1959).**
Chemoinformatics benefits greatly from an algorithm that finds The shortest path between two atoms in a molecule or two nodes in a graph. In topological graphs, the algorithm for the shortest path is frequently employed. The vector known as the Chemically Advanced Template System (CATS).

GRAPH REPRESENTATION

Draw a graph to represent the molecule. Every atom is a node, and every bond has a weight (such as bond length) corresponding to its edge.

RUN DILKSTRA's ALGORITHM

The shortest paths between every given atom can be found using Dijkstra's approach (beginning node) and every other atom in the molecule.

CONCLUSION

An overview of the field of chemoinformatics is given in this review, which uses computational methods and chemical expertise to study and comprehense the behaviour and characteristics of molecules. The paper emphasises how crucial machine learning techniques are to the field of chemoinformatics, especially for applications like data mining, data representation, and molecular property prediction. Additionally, it covers a number of algorithmic design paradigms that are utilised to maximise the efficiency of chemoinformatics algorithms, including as dynamic programming and object-oriented design.
It talks about how different algorithmic design paradigms, such dynamic programming and object-oriented design, are employed to maximise the efficiency of chemoinformatics algorithms. The importance of chemoinformatics in drug development and discovery is emphasised in the review, along with its uses in drug-target interaction optimisation, ADME property prediction, and target-based drug discovery efficiency enhancement. It also discusses how evolutionary algorithms—like genetic algorithms—are used in chemoinformatics and how well they work to tackle challenging issues.
Overall, the review concludes that chemoinformatics has become a powerful tool in the fields of chemistry, biology, and biochemistry, enabling the rapid analysis and prediction of molecular properties, and possesses the capacity to significantly speed up the effectiveness of the drug discovery process.

REFERENCES

Brown, F. K. Chemoinformatics: what it is and how does it impact drug discovery? Ann. Report. Med. Chem. 1998, 33, 375.

Cheng, Y.-L., Lee, C.-Y., Huang, Y.-L., Buckner, C. A., Lafrenie, R. M., Dénommée, J. A., Caswell, J. M., Want, D. A., Gan, G. G., Leong, Y. C., Bee, P. C., Chin, E., Teh, A. K. H., Picco, S., Villegas, L., Tonelli, F., Merlo, M., Rigau, J., Diaz, D., … Mathijssen, R. H. J. (2016).

DE., K. (1997). Algorithms for chemoinformatics. Application of Genetic Algorithms in the Field of Constitutional Similarity, 32:748-752(5), 716–726.

Willett, P. (2005). Chemoinformatics techniques for data mining in files of two-dimensional and three-dimensional chemical molecules . FIS2005, The Third Conference on the Foundations of Information Science.

Haghighatlari, M.; Li, J.; Heidar-Zadeh, F.; Liu, Y.; Guan, X.; Head-Gordon, T. Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods. Chem 2020, 6, 1527–1542. [CrossRef] [PubMed]

David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular representations in AI-driven drug discovery: A review and practical guide. J. Cheminform. 2020, 12, 56. [CrossRef] [PubMed]

Brown, N. M. (2005). Fingal: a novel algorithm for the generation of geometric fingerprints. QSAR & Combinatorial Science.

Cheng, Y.-L. (2016). We are IntechOpen , the world ' s leading publisher of Open Access books Built by scientists , for scientists TOP 1 %. Intech, 13. Retrieved from https://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics

Dijkstra, E. W. (1959). A Note on Two Problems in Connexion with Graphs. Numerische Mathematik.

Fontain E. Application of genetic algorithms in the field of constitutional similarity.
J Chem Inf ComputerSci 1992, 32:748–752

Erich Gamma, R. H. (1994). Design Patterns: Elements of Reusable Object-Oriented Software.

Faulon, J.-L. (n.d.). Department of Biology at the University of Evry in France.

Feinberg, e. a. (2020). Improvement in ADMET Prediction with Multitask Deep Featurization. J. Med. Chem.

Gasteiger J, Engel T, eds. Chemoinformatics: A Textbook. Weinheim: John Wiley & Sons-VCH; 2003

Gasteiger, J. (2023). Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review. International Journal of Molecular Sciences.

Gasteiger, J. (New York, NY, USA, 2003.). Handbook of Chemoinformatics. Wiley.

Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning.

Haghighatlari, M., Li, J., Heidar-Zadeh, F., Liu, Y., Guan, X., & Head-Gordon, T. (2020). Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods. Chem.

Handschuh S, Wagener M, Gasteiger J. Superposition of three-dimensional chemical structures allowing for conformational flexibility by a hybrid method. J Chem Inf Comput Sci 1998, 38:220–232

Hansch C., Fujita T. p-sd-pi analysis. A method for the correlation of biological activity and chemical structure. Journal of the American Chemical Society. 1964;86(8):1616– 1626.

Hassan M., Bielawski J., Hempel J., Waldman M. Optimization and visualization of molecular diversity of combinatorial libraries. Molecular Diversity. 1996;2(1):64–74

Knuth DE. The Art of Computer Programming, 3rd ed. Reading, MA: Addison-Wesley; 1997.

Michie, D. (1968, 218:19–22). Memo" functions and machine learning.

Polanski, J. (Elsevier, Amsterdam, The Netherlands, 2020; pp. 635–676.). 4.26-Chemoinformatics. Comprehensive Chemometrics, 2nd ed.

Research, P. (24 May 2023).). Small Molecule Drug Discovery Market Size, Report By 2032. Precedence Research.

Rutz, A., Sorokina, M., Galgonek, J., Mietchen, D., Willighagen, E., Gaudry, A., . . . al., e. (2021). The LOTUS initiative for open natural products research: Knowledge management through Wikidata. bioRxiv.

Sipser, M. (2012). Introduction to the Theory of Computation.

W. Richert, L. P. (n.d.). Building Machine Learning Systems with Python. Packt Publishing Ltd.

Willett, P. B. (1998). Chemical similarity searching. Journal of Chemical Information and Computer Sciences.

Clark R.D. OptiSim: an extended dissimilarity selection method for finding diverse representative subsets. Journal of Chemical Information and Computer Sciences. 1997;37(6):1181–1188.

Cherkasov A., Muratov E.N., Fourches D., Varnek A., Baskin I.I., Cronin M., et al. QSAR modeling: where have you been? Where are you going to? Journal of Medicinal Chemistry. 2013;57(12):4977–5010