International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217112

Integration of Cassandra and Spark in Computer Aided Drug Design

Nitha V R
Guest Lecturer, Department of Computer Science, Sree Narayana College, Cherthala, Alappuzha (Dt) Kerala, India

Article Info
Volume 7, Issue 1
Page Number: 68-73
Publication Issue :
January-February-2021

Article History
Accepted : 06 Jan 2021
Published : 16 Jan 2021

ABSTRACT

The primary purpose of this paper is to provide feasibility study of Cassandra and spark in Computer Aided Drug Design (CADD). The Apache Cassandra database is a big data management tool which can be used to store huge amount of data in different file formats. A huge database can be designed with details of all known molecules or compounds that are existing on earth. The information regarding the compounds such as selectivity, solubility, synthetic viability, affinity, adverse reactions, metabolism and environmental toxicity along with the 3 D structure of molecule can be stored in this big database. A data analytics tool "spark" can be efficiently used in mining and managing huge data stored in the database. Integrating big data in CADD helps in identifying the candidate drugs within minutes, not years. It may take eight to fifteen years to develop a new drug traditionally. Spark is written in Scala Programming Language which runs on Java Virtual Machine (JVM) and it supports Scala, Java and Python Programming languages .Cassandra can provide connectors to different programming languages, hence it's very easy to integrate any other molecular modeling tool with Spark. A python based molecular modeling tool called Pymol can be easily implemented with Spark. CADD helps in identifying new drugs by computational means thus eliminating unnecessary cost incurred in chemical testing of drugs.

Keywords : Cassandra, Spark, Resilient Distributed Dataset, API ,CADD , JVM, Pymol

I. INTRODUCTION

Apache Spark is an open source big data processing framework built around speed, ease of use, and sophisticated analytics. Spark gives us a comprehensive, unified framework to manage big data processing requirements with a variety of data sets that are diverse in nature (text data, graph data etc).

The spark connector will not be able to directly access the Cassandra tables. Instead an abstract version of Cassandra data tables is created. This abstraction is known as a "Cassandra RDD/ Resilient Distributed Database". The spark Cassandra creates Cassandra RDDs, which is a fault tolerant collection of elements that can be operated. The data in this database is spread across multiple servers so that if one server fails, the data will be safe in another server. Even if redundant data is kept on multiple clusters,

it's distributed for safety. Cassandra can cluster as many number of clusters for data handling.

Apache Cassandra is based on Amazon dynamo and Google big Table. Since Cassandra is open source software, anybody can download it, use it freely and modify it according to their business specification. Hence there's no need to purchase the software and there's no licensing fee and maintenance fee. Spark provides programmers an API by providing Resilient Distributed Dataset (RDD) which is a collection of data items distributed over a cluster of machines and is maintained in a fault-tolerant way.

Basically RDDs must be broken into partitions, which relate to data that is locally stored on nodes in the cluster. The spark partitions must represent every piece of data on the node. Spark can store any type of data hence it can store many different properties of a molecule, including the 3D structure. Spark SQL allows you to perform relational queries over data stored in Cassandra clusters, and executed using Spark. Spark SQL is a unified relational query language for traversing over Spark Resilient Distributed Datasets (RDDs) which support a variation of the SQL language used in relational databases. Spark SQL uses a special type of RDD called SchemaRDD, which are similar to tables in a traditional relational database.

Hadoop is extensively used to analyze big data. Since Hadoop framework is based on MapReduce (a simple programming model) it can be widely used to process huge data which enable scalability, flexibility, fault-tolerance and cost effectiveness. Spark was introduced by Apache for speeding up the Hadoop computational software process. Spark has its own cluster management and can't be considered as a modified version of Hadoop. Hadoop can be considered as one of the ways to implement Spark. Spark uses Hadoop in two ways – one is storage and second is processing. Since Spark has its own cluster management computation, it uses Hadoop for storage purpose only. Apache Spark is designed for fast computation. It is based on Hadoop MapReduce and it extends the MapReduce model to efficiently use it for more types of computations.

Spark helps to run an application in Hadoop cluster, up to 100 times faster in memory, and 10 times faster when running on disk. This is possible by reducing number of read/write operations to disk. It stores the intermediate processing of data in memory. Spark provides built-in APIs in Java, Scala, or Python. Therefore, you can write other applications such as molecular modeling in different languages. Connectors can be written in any programming language so that the data base can be accessed from any front end application. Spark facilitates interactive querying. Spark not only supports 'Map' and 'reduce'. It also supports SQL queries, Streaming data, Machine learning (ML), and Graph algorithms. Spark SQL includes a server mode with industry standard JDBC and ODBC connectivity.

Spark can be deployed on top of HDFS (Hadoop Distributed File System) and space is allocated for HDFS. Spark provides In-Memory computing and referencing datasets in external storage systems. Spark SQL is a component on top of Spark Core that introduces a new data abstraction called SchemaRDD, which provides support for structured, unstructured and semi- structured data.

Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. Each dataset in RDD is divided into logical partitions. RDDs can contain any type of Python, Java, or Scala objects, including user-defined classes.

Computer Aided Drug Design (CADD)

Computer Aided Drug Design (CADD) represents the integration of computational methods and graphical

resources to facilitate the design and discovery of new therapeutic drugs. CADD involves the invention of new medications (or modification of existing medicines) based on the knowledge of a biological target that uses computational methods to simulate drug receptor interactions to determine if a given molecule will bind to a target and what its affinity would be. Drug design involves the design of small molecules that are complementary in shape and charge to the bio-molecular target with which they interact and therefore will bind to it. CADD methods are heavily dependent on bioinformatics tools which include software applications, information technology, information management and databases.
A drug target is a key molecule or protein involved in a particular metabolic pathway that is specific to a disease condition. A drug may inhibit the functioning of the pathway in the diseased state by binding a key molecule in the active site thereby stopping its functioning. Another approach may be to enhance the normal pathway by promoting specific molecules in the normal pathways that may have been affected in the diseased state. There are various computer modeling techniques available to design a drug. CADD requires significantly less cost and time for high throughput screening without compromising the quality of lead discovery. Binding of ligands to the receptor may occur via hydrophobic, electrostatic, and hydrogen-bonding interactions.

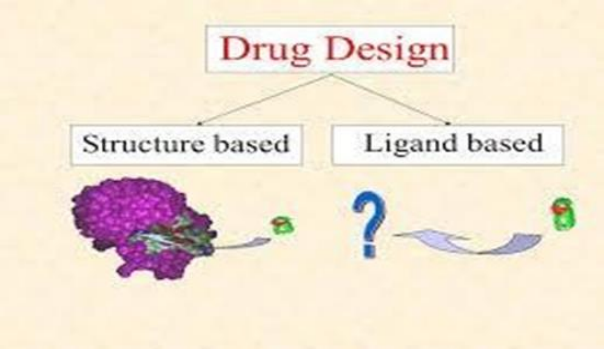
Drug design can be done in two different ways.

1) Structure based drug design (SBDD)
2) Ligand Based Drug Design (LBDD)

Normally Structure based drug design is done based on the knowledge of 3 D structure of biological target that is obtained through X-ray Crystallography or NMR Spectroscopy. Template modeling of proteins can be done by Homology Modeling which involves modeling a protein 3D structure using a known experimental structure of a homologous protein (the template). By the usage of Cassandra and Spark the 3D structure of already known compounds can be kept in the big data base. If a molecule is known for toxicity or inactivity it can be easily eliminated from preliminary selection while selecting the candidate drug.

The basic drug design involves the invention of a new drug based on the knowledge of a biological target.

1. Identify a compound which is mostly organic, complementary in shape to target molecule and oppositely charged to the bio-molecular target
2. This molecule interacts with target, bind the target, activates or inhibits the function of a bio molecule (protein).



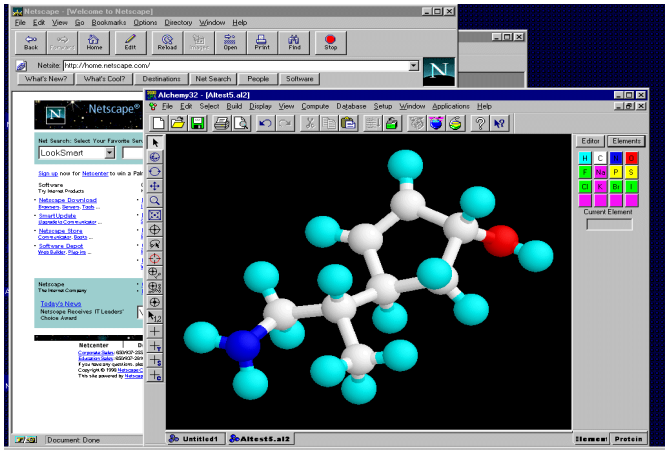
In Ligand based Drug Design the receptor structure is not known .But the ligand and their biological activities are known. Based on that information atoms can be randomly selected to identify a candidate drug using computer aided drug design.

In the case of Ligand based drug design, ligand based computational methods can be employed in screening new drugs. It relies on knowledge of other molecules that bind the biological target. Ligand based approach is used when the receptor is not known. LBDD can be used to identify characteristics common to known ligands which help in screening new or improved drugs. If a set of active ligand molecules is known but no structural information exists for the target, ligand

based computational method can be employed. There are two methods used in LBDD.

1) QSAR (Quantitative Structure Activity Relationship): This is based on the assumption that there is an underlying relationship between the molecular structure and biological activity. On this assumption QSAR attempts to establish a correlation between various molecular properties of a set of molecules with their experimentally known biological activity.
2) Pharmacophore method: This method identifies a pharmacophore which is an important step in the interaction between a receptor and a ligand.

In LBDD, the molecule can be designed by using computational techniques by selecting each atom from the screen and the user can make combinations of atoms through computer.



A drug is a chemical entity which when consumed or injected results in the eradication of a particular disease or infection. Drug discovery involves the pipeline process involved in the evolution of drugs. CADD is a drug discovery process that starts with an analysis of binding site in target proteins or an identification of structural features common to active compounds. The process ends with the generation of small molecule called “leads” suitable for further chemical synthetic work.

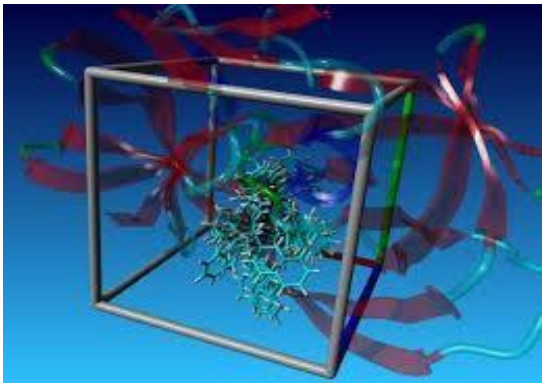
CADD uses several bioinformatics tools & related fields like chemi informatics & combinational chemistry. DRUGS are small molecules that bind, interact and modulate the activity of specific biological receptors. The initial leads are unlikely to be the final products. Complex evaluations are necessary and typically the initial hit is modified atom by atom to become the drug molecule .The choice of lead structure is important for successful drug development. Impact of structural bioinformatics on drug discovery is it speeds up the key steps in combing aspects of bioinformatics, structural biology and structure based drug design.

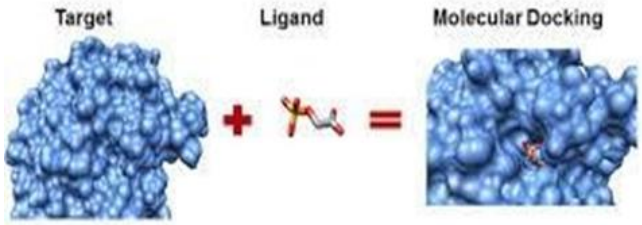
CADD uses computational chemistry to discover, enhance or study drugs and related biologically active molecules. The CADD is not to find the ideal drug but to find the lead compounds which save some experiments. CADD help to find out the lead molecule which is compounds having biological activity on a validated molecular target. A molecule is considered as a lead compound if it exceeds a specific potency threshold against the target. Techniques like NMR base screening, high throughput docking and QSAR can be done to identify a lead molecule. HITS are the chemical compounds that produce biological activity which represent therapeutic potential. Steps involved in CADD are:

- ✓ Identify disease- One has to determine the biochemical basis of disease process.
- ✓ Target Selection – Knowledge of the molecular basis of the disease is important in order to select a target which disrupts the process.
- ✓ Target structure determination – crystal structure of target protein can be taken from database.
- ✓ Determine active site of target protein- Identify the active site where a ligand can bind. Drugs bind to a protein by lock and key mechanism.
- ✓ Selection of ligand /drug- A ligand or a candidate drug is selected which has a tendency to bind.
- ✓ Molecular docking – Docking is a method which predicts the preferred orientation of one

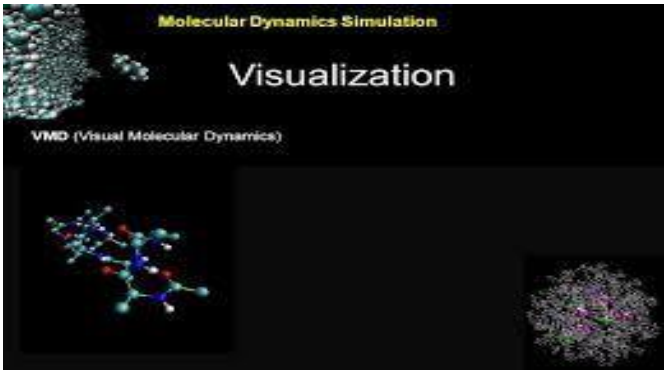
molecule to a second when bound to each other to form a stable complex. There are some docking programs and algorithms that enable the user to predict favorable biological target-ligand complex structures with reasonable accuracy and speed. Docking involves scanning a database of known molecules for those likely to bind well to the receptor. Hence Docking can be used to generate possible binding geometries and can evaluate using a scoring function. Docking algorithms can be used to find ligands and binding conformations at a receptor site close to experimentally determined structures. Docking algorithms are used to identify multiple proteins to which a small molecule can bind.

- ✓ AutoDock – It's a suite of automated docking tools. It's designed to predict how small molecules such as candidate drugs bind to a receptor of known 3D structure.





- ✓ Visualization of docked complex-The docked complex is visualized and studied using a software like VMD(Visual Molecular Dynamics)



- ✓ Retrieving 3 D structure- first step for protein visualization is to extract the protein structure from a structure database, which will act as input for 3D visualization programs.

Benefits of CADD:

- ✓ Elimination of compounds with undesirable properties like poor activity or high toxicity.
- ✓ It enables screening upto 10000 compounds a day for activity against a target protein .Hence CADD can rapidly produce vast numbers of compounds.
- ✓ The usage of computer graphics ,virtual screening and molecular modeling help predict activity

Current trends in CADD:

- ✓ Data storage and Retrieval: End to end information about a molecule and their properties can be kept in a big database so that all the required information about a particular molecule is available under a roof. Information about Molecular toxicity, solubility, stability, synthetic viability, metabolism and excretion can be made available in a huge database.
- ✓ Visualizing molecules: Similarities and differences b/w the drugs acting in the same way can be easily identified using CADD. The interaction between drugs and receptors can be easily visualized.

- ✓ Calculations: The strength of interaction between drugs and protein molecule can be identified efficiently by CADD.
- ✓ CADD helps to identify the toxicological properties, excretion and metabolism of drugs in advance before doing experimentally.

II. CONCLUSION

Traditional method of new drug design and identification will take 10- 15 years and the estimated cost to bring a new drug to market is $800-1000 $ million. Nowadays drug companies are not ready to spend their precious time, money and years for development of new drugs.20 % of drug manufacturing cost increase per year. CADD helps to increase speed, and reduce the time of development. Since there's no experimental testing of drugs no unnecessary chemical waste is produced. Hence we are not worried about the disposal of chemical waste produced. CADD helps to meet the strategies to overcome toxic side effects of new drugs which help to design drugs in a cost effective manner. CADD enable computational scientists to manipulate molecules on screen rather than manual testing of chemicals. Druggability of drugs can be easily identified in CADD than chemical testing. Efficient analysis and interpretation of Big Data opens new window in CADD. Big data analytics is the term used to describe the process of researching massive amounts of complex data in order to reveal hidden patterns or identify secret correlations. Integration of Cassandra and Spark help in CADD help to identify the exact molecule having druggability. Big data specifically refers to data sets that are so large or complex that traditional data processing applications are not sufficient.

III. REFERENCES

[1]. http://cbb.sjtu.edu.cn/~qinxu/files/papers2013/xuke-JCAMD2011.pdf

[2]. W. G. Richards, "Computer-Aided Drug Design," Pure and Applied Chemistry, Vol. 66, No. 8, 1994, pp. 1589- 1596. Doi:10.1351/pac199466081589

[3]. https://www.um.edu.mt/data/assets/pdf_file/0004/84307/WhatMakesaMoleculeDrug-

[4]. Like.pdf http://file.scirp.org/pdf/OJMC_2012123116390139.pdf

[5]. https://academy.datastax.com/resources/getting-started-apache-spark-and-cassandra

Cite this article as :

Nitha V R, "Integration of Cassandra and Spark in Computer Aided Drug Design", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 1, pp. 68-73, January-February 2021. Available at doi : https://doi.org/10.32628/CSEIT217112
Journal URL : http://ijsrcseit.com/CSEIT217112