Turbo Analytics: Applications of Big Data and HPC in Drug Discovery

Rajendra R. Joshi, Uddhavesh Sonavane, Vinod Jani, Amit Saxena, Shruti Koulgi, Mallikarjunachari Uppuladinne, Neeru Sharma, Sandeep Malviya, E. P. Ramakrishnan, Vivek Gavane, Avinash Bayaskar, Rashmi Mahajan and Sudhir Pandey

Abstract In this current age of data-driven science, perceptive research is being carried out in the areas of genomics, network and metabolic biology, human, animal, organ and tissue models of drug toxicity, witnessing or capturing key biological events or interactions for drug discovery. Drug designing and repurposing involves understanding of ligand orientations for proper binding to the target molecules. The crucial requirement of finding right pose of small molecule in ligand–protein complex is done using drug docking and simulation methods. The domains of biology like genomics, biomolecular structure dynamics, and drug discovery are capable of generating vast molecular data in range of terabytes to petabytes. The analysis and visualization of this data pose a great challenge to the researchers and needs to be addressed in an accelerated and efficient way. So there is continuous need to have advanced analytics platform and algorithms which can perform analysis of this data in a faster way. Big data technologies may help to provide solutions for these problems of molecular docking and simulations.

Keywords Drug discovery · Drug repurposing · Hadoop · Big data
Molecular dynamics simulations

Abbreviation

PCA	Principal component analysis
RMSD	Root-mean-square deviation
RMSF	Root-mean-square fluctuation
MR	MapReduce

R. R. Joshi (✉) · U. Sonavane · V. Jani · A. Saxena · S. Koulgi
M. Uppuladinne · N. Sharma · S. Malviya · E. P. Ramakrishnan
V. Gavane · A. Bayaskar · R. Mahajan · S. Pandey
High Performance Computing-Medical & Bioinformatics Applications Group,
Centre for Development of Advanced Computing (C-DAC),
Savitribai Phule Pune University Campus, Pune 411007, India
e-mail: rajendra@cdac.in

C. G. Mohan (ed.), *Structural Bioinformatics: Applications in Preclinical Drug Discovery Process*, Challenges and Advances in Computational Chemistry and Physics 27, https://doi.org/10.1007/978-3-030-05282-9_11

1 Introduction

This decade has been witnessing a major shift in technologies which have been used in various sectors ranging from social media, agriculture, services, to science and technology. In the current age, new advances are being made in the field of satellites, robotics, micro- and nanotechnologies as well as revolution in computing. The stream of science has been impacted by this revolution. All disciplines of science have been generating and building newer technologies and different approaches for scientifically accurate experimentation. All these developments in various scientific disciplines are also changing our social life, health, environment, etc. One of the major streams of science is life sciences, which has been strongly affected and accelerated due to all these advancements in techniques and technologies.

Various technologies like next-generation sequencing (NGS) in genomics, high-throughput assays, and supramolecular chemistry are revolutionizing the life sciences and applied areas of human health, agriculture, livestock, and many more [1–4]. The robotics-based automation is generating volumes of data from various experiments and characterization techniques. The next-generation biology has been driven heavily by wet laboratory experimentation as well as dry laboratory computation.

Technologies like next-generation sequencing (NGS) enable sequencing of genomes of thousands of species in plants and animals at an extremely rapid rate [5–7]. Today, many genome sequencing centers are producing data of about terabytes per week. This results in petabytes of data of sequencing information per year. The figure is expected to grow exponentially and very soon will be facing challenges of storage and analysis of exabytes of sequence data [5–7]. To extend this further, there is already a race to sequence the genomes of all living species on the planet including humans, plants, animals, microbes to name a few. It is expected that this gigantic exercise will result in zetabytes to yottabytes of sequence data. Such large volumes of sequence data will be the genomic ocean of tomorrow [7–9].

Similarly, structural database of biomolecules like protein, nucleic acids, lipids, and membranes is also growing rapidly (shown in Fig. 1) due to methods like cryo crystallization, high-frequency NMR, and other characterization techniques along with computational modeling techniques [10]. Computational modeling and simulation of biomolecules have been drastically improving due to the advancement in high-performance computing (HPC) [11] and development of advanced enhanced sampling methods [12, 13]. It has paved the way for mimicking long timescale events occurring in different biological systems more efficiently. Owing to the better computing paradigm, today structural data generation is no more the major challenge, but analyzing this huge data has become one. Computer simulations help to determine mechanism of action of biomolecules in a cell, thereby suggesting their implication in various diseases and discovering their potential use in therapeutics. Hence, the computational techniques generate biomolecular structural and dynamical data via very long time scale simulations. Likewise, detailed and

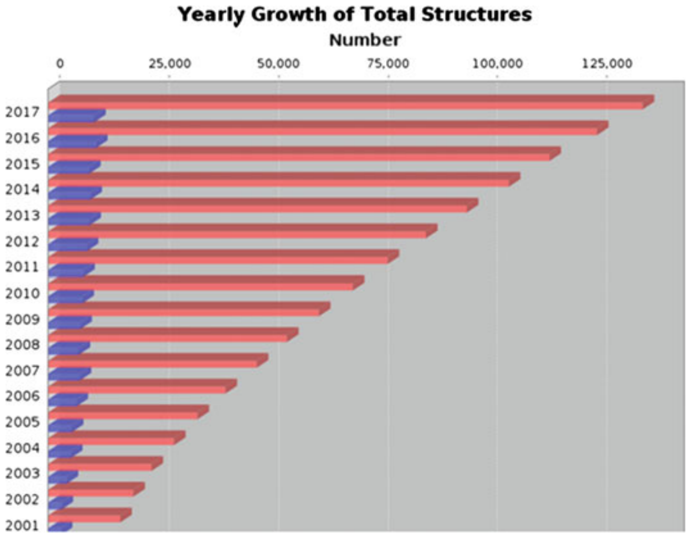


Fig. 1 Growth of structural data from 2001 onwards. *Source* https://www.rcsb.org/pdb/statistics

systematic analysis of data becomes an important part of any study, as it would further help to understand the entire mechanism of biomolecular action. Advances in crystallization, NMR, and computational methods are directly influencing and accelerating the drug discovery process.

2 Drug Discovery Process

Discovering a new drug is a very complex, time-consuming, expensive, and high-risk process for R&D and pharmaceutical laboratories [14–16]. It is also a multi-step process involving target identification, target validation, and screening of small molecules for validated targets. These steps need to be made easy, cost-effective, and fast. Computational method like computer aided drug discovery is one such process that involves identifying new ligand molecules for a particular target protein, which is an important step in drug discovery. Historically, the drug discovery process was involved extraction of chemical compounds from natural resources and testing them in the cell for disease treatment [17]. With the advancement of technology and ability to chemically synthesize small chemical moieties, various drug databases came into existence. The availability of vast structural resource of small molecules has made high-throughput screening of these databases against target protein a more feasible practice. Also, increasing affinity and reducing toxicity of already available ligand molecules needs to be addressed in drug discovery process.

Drug discovery process involves the following steps: (1) target identification, (2) validation of target protein, (3) creation of small molecule database,

(4) screening of small molecules against target protein, i.e., hit to lead identification, (5) lead optimization, (6) preclinical testing, and (7) clinical testing.

Almost all these steps generate huge data from experimental laboratory and computational laboratory experimentations and need better way of handling data with fast and better analytics approaches. Target identification and validation involve selection of protein molecules whose activity when blocked or enhanced can affect the particular disease-related cellular pathway. This involves a systems biology approach wherein an understanding of all the proteins involved in the pathway or finding possibility of any alternate pathway available, role of particular protein in particular pathway and identifying side effects of the target protein. Second most important thing is to have database of lakhs of small molecules which can be screened against the target protein. The source of these small molecules can be microbial metabolites, plant origin, and chemically synthesized. There are various drug molecule databases, i.e., Chemspider [18], DrugBank [19], ZINC [20] to name a few which are already available.

The technique to screen these lakhs of molecules to a target protein is performed using molecular docking. The screening process should be fast enough, which demands the use of and better computational or programming techniques. Each of these molecules tends to have conformational flexibility which in turn makes the docking process more time-consuming. Choice of efficient force field and scoring methodologies also plays an important role in screening of these molecules. In order to achieve this, high-throughput docking methods have been developed. Although, the analysis of these docked conformations to choose the best ligand becomes a big data analytics problem as it involves finding of various parameters and several interactions between the target protein and the docked ligand.

Docking or screening projects a static picture of the binding of ligand with the receptor [21]. However, the dynamic picture would be obtained from the molecular dynamics simulations which provide an understanding of the flexibility of protein and ligand. Molecular dynamics simulation gives an insight about various intermolecular interactions and binding affinities between protein-ligand complex, thereby ensuing binding efficiency [16]. Molecular docking followed by simulations generates huge molecular trajectories data. Thus, the management and fast analytics of this data have become the need of the hour.

The upcoming area of drug repurposing is again proving to be a bigger computational task, and it has the potential to deliver a drug molecule for a chosen disease [22, 23]. Various pharmaceuticals and R&D laboratories are working on drug repurposing which involves docking of already approved FDA drugs on new target protein. The involvement of FDA-approved drugs suggests that they have been already tested on humans for their toxicity and pharmacology. Hence, rejection of such drugs due to toxicity is ruled out, and entire duration required for the drug discovery process can be shortened by few years. HPC-based molecular docking and molecular dynamics simulations pose a challenging role in this area of drug repurposing.

In order to manage this rapidly increasing data and efficient analysis, there is need to develop tools with parallelization and thereby enhance the overall

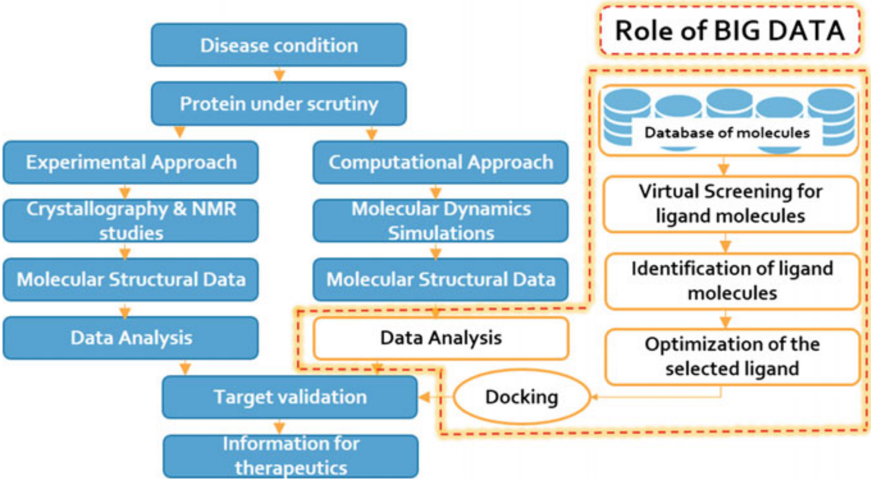

Role of BIG DATA
Disease condition
Protein under scrutiny
Experimental Approach
Computational Approach
Crystallography & NMR studies
Molecular Dynamics Simulations
Molecular Structural Data
Molecular Structural Data
Data Analysis
Data Analysis
Database of molecules
Virtual Screening for ligand molecules
Identification of ligand molecules
Optimization of the selected ligand
Docking
Target validation
Information for therapeutics

Fig. 2 Role of big data analytics in drug discovery

performance. This denotes a continuous need to have advanced analysis platform and algorithms which can perform analysis of the biological data in a faster way. Big data technologies may help to provide solutions for these problems of molecular docking and simulations (Fig. 2).

3 Big data Technologies: Challenges and Solutions

The context of big data is dependent on the problems and the existing technologies. Today’s big data can be tomorrow’s small data as the technologies and methods that are handling the data may become more advanced in the future. The big data is the data that cannot be handled using the existing traditional methods and requires specialized methods to solve the big data problem.

Big data is categorized by its three main properties, viz. volume, velocity, and variety [24]. Volume denotes the huge data that needs to be analyzed, velocity tells about the rate at which the data is generated of the data, and variety tells about the different types of data that can be generated by the various sources using different formats of data generations and exchange. Big data usually expands rapidly in the unstructured form and varies to such an extent that it becomes difficult to maintain the data in traditional databases. In such cases, specialized techniques like NoSQL [25] can be used to handle the problems of the unstructured data. Big data technologies are capable of managing huge data generated in different formats. Advancements in technologies like cloud computing offer a unified platform to store and retrieve the data. The Internet speed has increased to several manifolds, and the cloud technologies have effectively exploited the Internet capabilities to offer a scalable, multi-user platform for big data analytics in the field of Bioinformatics. The use of big data in the Bioinformatics is an emerging field which presents new opportunities to medical researchers and paves the way toward prediction of personalized medicines. The greatest challenge lies in designing a strategy to acquire the data followed by filtering it to meet the appropriate decision-making demands.

This can be achieved by bringing together experts from clinical medicines, computer science, bioinformatics, biotechnology, and statistics and address the challenge of the data management and analytics solutions toward precision biology. Hadoop [26]-based platform with MapReduce and spark-based algorithms may be useful to make all the analysis optimized with fast calculation. Hadoop- and MapReduce [27]-based algorithms implemented on scalable architecture have been discussed further along with drug repurposing big data case study for cancer protein.

4 Big data Technology Components

Hadoop

Apache Hadoop is an open-source software framework for storage and large-scale processing of datasets on clusters of commodity hardware. Hadoop has gained lots of popularity among the peer parallel data processing tools because of its simplicity, efficiency, cost, and reliability. Hadoop can be built on the commodity hardware. Hadoop has major three components. Hadoop Distributed File System (HDFS), YARN scheduler and resource negotiating framework and the MapReduce [27] programming framework. A typical framework of Hadoop test bed is shown in Fig. 3.

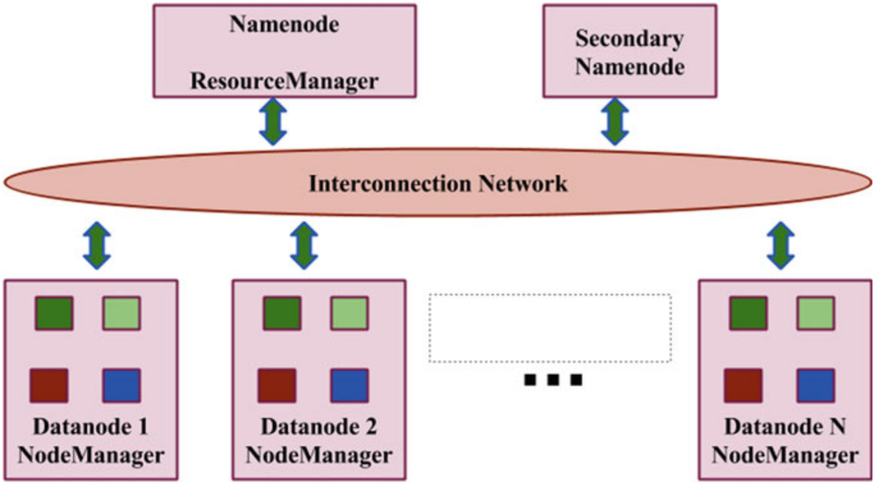


Fig. 3 Basic architecture diagram of hadoop test bed

I. *HDFS*

Hadoop Distributed File System (HDFS) is built to provide high-throughput, reliable, efficient, and fault-tolerant file system. It can provide streaming reads and writes for large files. The basic architecture diagram of HDFS is shown in Fig. 4. As shown in the figure, the HDFS has two main components, namenode, and datanode. HDFS is mainly designed for low-cost hardware, and hence, it can be built on cluster of commodity hardware. In HDFS, the file is divided into fixed size blocks or chunks of 128 MB each except the last chunk. The fixed size 128 MB can be configured with various needs. Namenode contains the metadata information of all the files. It stores information regarding the block of file stored on datanodes, while datanodes actually store the block of data. Each block is stored on three datanodes of the cluster. This policy provides reliability at the cost of redundancy. Generally, two copies of blocks are stored on two different datanodes of the same rack of cluster, while the third copy is stored on the datanodes of the different rack of the same cluster. These two racks are connected by a very high-speed network switch. This policy ensures the reliability of the HDFS file system. In case, if any two nodes fail, still the data can be accessed from the datanode having this third copy of the data. Datanodes periodically updates their state to the namenode so that namenode can be aware of the overall state of cluster. While scheduling MapReduce [27] job, the hadoop framework ensures with most possibility that the mapper task should run on the same datanode where the actual data is residing. This avoids significant network overhead. This policy of hadoop improves the performance of the overall cluster.

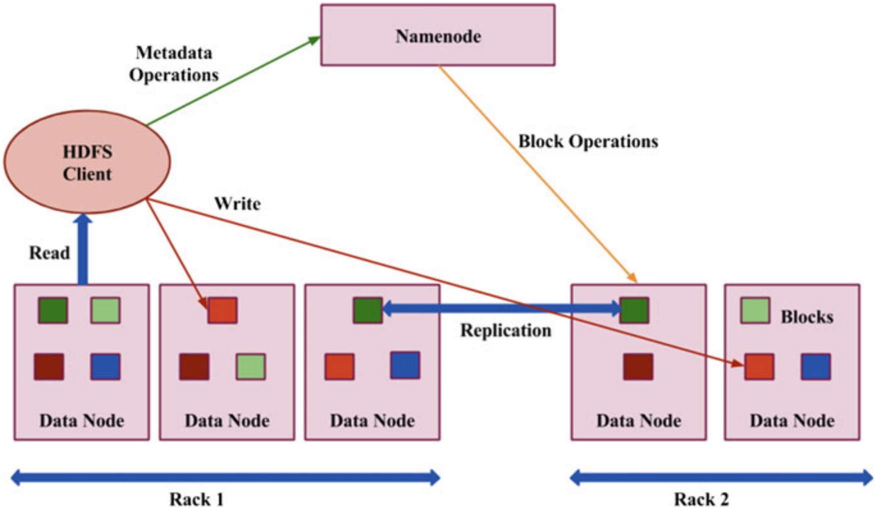


Fig. 4 Basic architecture diagram of Hadoop Distributed File System (HDFS)

HDFS major components:

(i) **Namenode**
Namenode stores the metadata about the file. It has the complete view of the distributed file system. It tracks which datanode is active and which are node. In case of any datanode failure, it initiates the operation regarding maintaining the replication factor by copying the data stored the failed nodes to the active datanodes. In case, namenode fails, the complete HDFS file system gets crashed.

(ii) **Datanode**
It stores the actual data. It performs the read and write operation once it receives the command from the namenode. It is responsible for block creation, deletion, and replication. It periodically sends the heartbeat signal to the namenode.

II. *Map Reduce*

Hadoop MapReduce is the programming framework. It is one of the major parts of the Apache Hadoop project. It provides the programing model for data parallel application. The basic flow of MapReduce algorithm is shown in Fig. 5. MapReduce programming model makes use of HDFS and makes the application performance very efficient and fast. The MapReduce framework with the help of Hadoop framework places the mapper job on the datanode where the actual data resides. It improves the performance and removes the network bottleneck while processing huge amounts of data. The major phases of the MapReduce program are mapper, partitioner, combiner, shuffle and sort, and reducer.

The mapper reads the data from HDFS and processes it. This is followed by the partitioner ensuring that the processed data is sent to be the desired reducer. The

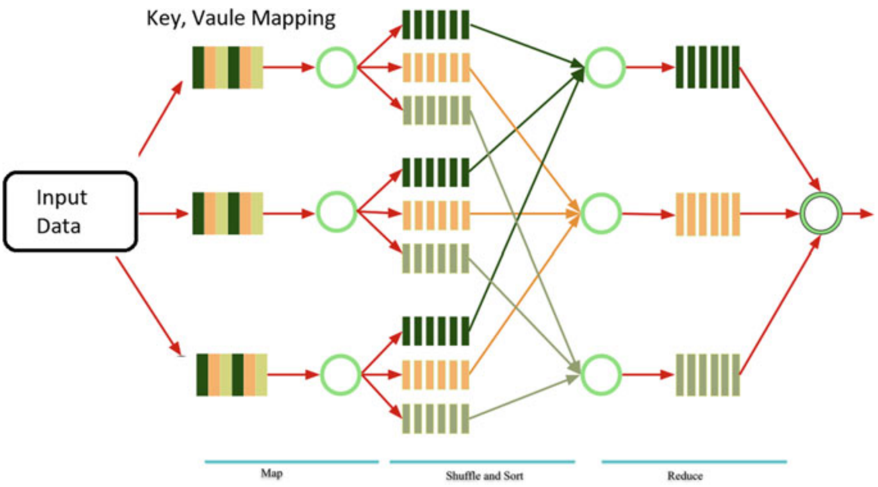


Fig. 5 Basic flow of MapReduce algorithm execution

data before being sent to the reducer is shuffled and sorted so that the reducer can easily process it. Finally, the reducer performs an operation of reduction or aggregation on the final data and this is followed by writing the final output to the HDFS. The combiner does a similar task as the reducer but at the mapper lever providing local lever aggregation or reduction.

III. *YARN*

Apache YARN stands for Yet Another Resource Negotiator. Before Hadoop 2.x, the only framework which could run on Hadoop platform is MapReduce. The job scheduling and resource negotiation is integrated with the MapReduce framework and shared by Hadoop framework. The YARN provides the separate layer for job scheduling and resource negotiation. It provides the platform for other programming framework like spark and storm, and many can run on Hadoop framework. The basic architecture of YARN is shown in Fig. 6.

YARN has ResourceManager, NodeManager, Container, and ApplicationMaster. Each container on datanode is specified with amount of CPU and memory, and it is configurable. ResourceManager is run on namenode, and NodeManagers are run on datanodes. Whenever a job is submitted, one container is allocated by a ResourceManager on any datanode. This container process is called as ApplicationMaster. This ApplicationMaster is responsible for all job management and resource negotiation with ResourceManager. With the help of ResourceManager,

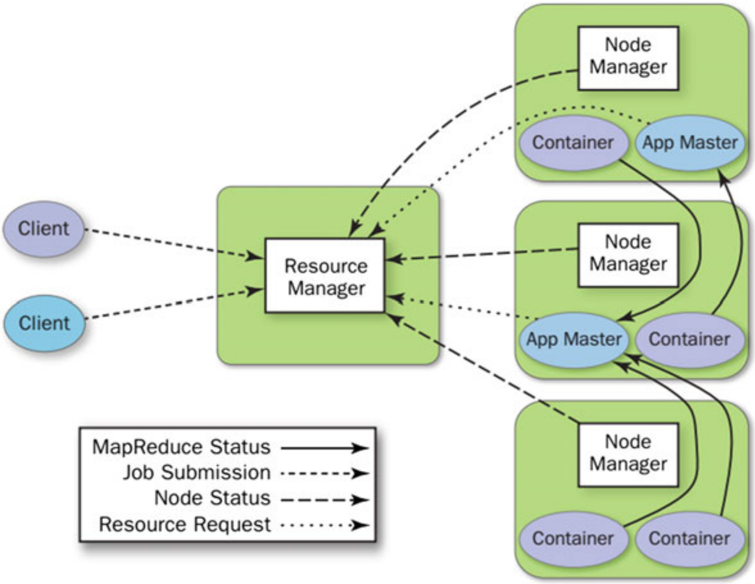


Fig. 6 Basic architecture of YARN showing various components

this ApplicationMaster allocates Containers from NodeManager for MapReduce task. This approach reduces the load on ResourceManager and distributes it across ApplicationMasters on the datanodes for each job. This way, using YARN the hadoop cluster can grow up to 10,000 nodes. Earlier benchmark without YARN on Hadoop 1.x was up to 4000 nodes. This way YARN provides scalability to the Hadoop cluster along with different programming platforms to be incorporated in hadoop framework.

5 Big data Tools Development for Drug Discovery

There have been efforts by various scientific groups to use HPC, grid technologies for drug discovery. Multiple docking tools like DOCK6 [28], Gold [29], Autodock Vina [30], and some others are already available in the parallel mode on HPC platform. Most of these tools are fast and robust; however, they have their own scoring functions based on molecular mechanics force fields and other geometrical descriptors. Although, improvements are still going on in enhancing the scoring function and guiding it further toward higher efficiency and accuracy. Docking with the concept of flexible ligand and protein still remains to be time-consuming calculation. Docking of multiple ligands to single protein or multiple ligands with multiple proteins may be some of the future challenges in docking area. Understanding the flexibility of both the proteins and ligands has been taken care by some of the currently available molecular simulation packages like AMBER [31], CHARMM [32], GROMACS [33], and NAMD [34]. All these packages are known to be scalable on the HPC platform. Although molecular simulations are time-consuming, they still prove to be the best in understanding the allowed flexibility of proteins, ligands, active sites, and other biomolecular entities. The advent of cloud and big data technologies promises to accelerate the drug development process using MapReduce [27] and spark methods coupled with machine learning and deep learning analytics. The tools like DIVE [35], HiMach [36], and HTMD [37] have been developed for molecular simulations as well as trajectory visualization and analysis. Many more tools may be getting developed using these newer technologies.

Bioinformatics group at C-DAC, Pune, has been addressing the issue on data analytics and visualization of trajectories in structural biology domain using HPC technologies combined with big data technologies. Various analytics tools have been developed and tested on Hadoop platform using MapReduce as shown in Fig. 7. At this stage, analytics tools for multiple molecular trajectories include hydrogen bond calculations, identifying water molecules and bridged water-mediated interactions. Other big data analytics tools for RMSD, 2DRMSD, RMSF, water density, WHAM-based free energy calculations are in the process of development. Few of the big data analytics tools which have been already developed proved to be useful in the process of drug discovery. These tools have described below.

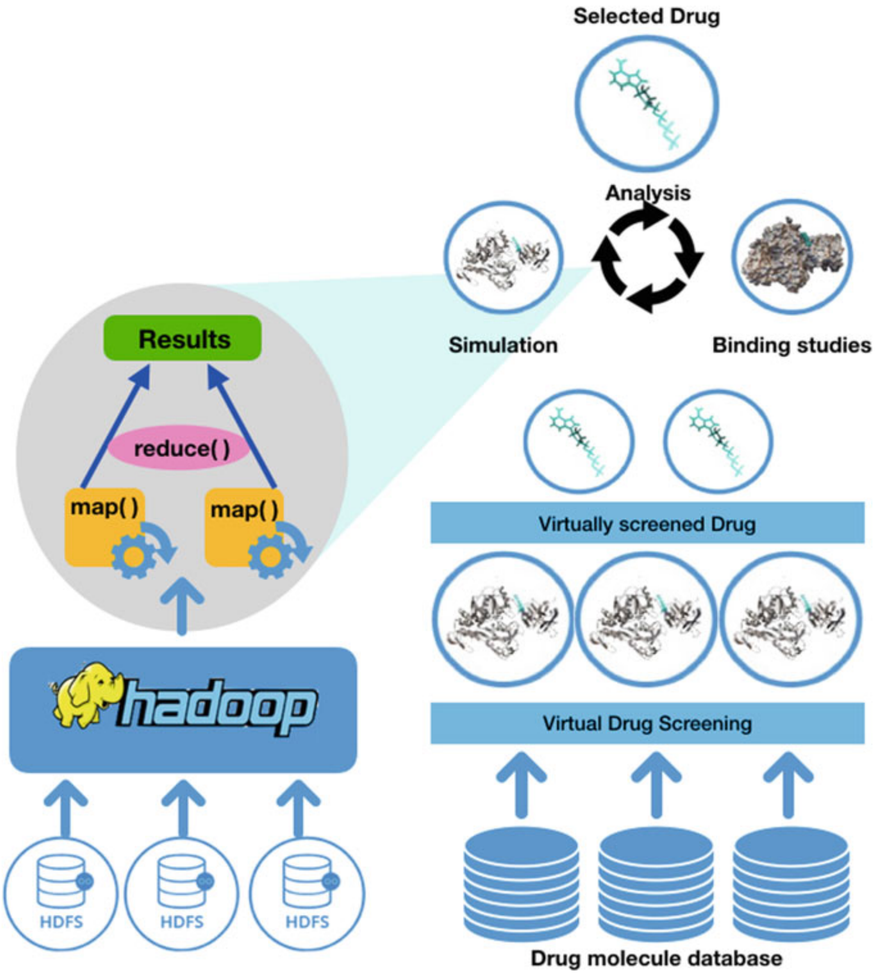


Fig. 7 Schematic representation of role of Hadoop and MapReduce paradigm in drug discovery process

5.1 Hydrogen Bond Big data Analytics Tool (HBAT)

The molecular dynamics (MD) simulations generate large trajectories which would be in the size of GBs to TBs depending on the size of the molecule and length of the simulation time. Many of the MD simulations use explicit solvation models in which water molecules are added explicitly to the solute to mimic the natural system. This increases the size of the system drastically in terms of number of atoms, and the analysis of such system becomes more compute intensive, iterative, and time-consuming. There are various analysis programs (ptraj, cpptraj [38], VMD [39], etc.) available corresponding to the different MD simulation packages.

All these programs have modules written for performing different analyses like RMSD, RMSF, radius of gyration, PCA [40], distance calculations, H-bond analysis, and MMGBSA [41] free energy calculations. However, many of these programs are either inefficient or very slow in calculating the H-bond interactions within solute and especially between the solute and the solvent (water molecules). These programs are highly time-consuming and also have constraint in dealing with the large size data for example 500 GB or beyond. This drawback of the existing tools suggests a strong need for the development of water-mediated H-bond analysis tool which is capable of handling a very large size of trajectories and also be executed parallel to reduce the time. The water molecules added to the system may play a crucial role in the activity or functioning of that particular molecule. Hence, understanding the role and mechanism of such water molecules and their interactions with the solute (protein/RNA/DNA or drug) molecules is very important [42, 43]. In order to achieve this, a big data analytics tool for hydrogen bond calculation was developed by Bioinformatics group C-DAC.

The MapReduce algorithm for H-bond calculation was developed and ported/ tested on Hadoop cluster. The algorithm flow has been shown in Fig. 8a for H-bond calculation using the MapReduce approach. The HDFS file system was used to store the multiple molecular trajectories data. The current version of tool can analyze trajectory data in the PDB format generated using molecular dynamics packages like AMBER [31], GROMACS [33], CHARMM [32]. The tool is scalable or portable on any distributed computing platform and can find out H-bonds between all types of residues including water. However, the tool requires a significant amount of time for executing the preprocessing stage where, the PDB files are generated from the trajectories and copied on the distributed HDFS storage. Despite this overhead, the overall performance of the tool is better than currently existing tools such as CPPTRAJ or PTRAJ [38], especially for trajectories with a large number of water molecules. The benchmarking of H-bond tool is shown in Fig. 8b. The benchmarking of up to 5.5 TB data is carried out, and it shows near linear scale up. Additionally, the tool can also help identify water-mediated interactions such as water bridges easily.

5.2 Molecular Conformation Generation on Cloud (MOSAIC)

Drug databases usually contain millions of ligands, and for each ligand, there can be billions of conformations [44, 45]. Such billions of conformations need to be docked on to a target which is a generally a protein molecule. Generation and optimization of such billions of ligand conformations is a huge computational problem, since it involves the use of advanced methods like molecular mechanics, semi-empirical and quantum techniques [46, 47]. The application of an embarrassingly parallel approach accompanied by virtualized resource scaling and an efficient structure optimization tool can handle billions of conformations with the help of cloud computing technologies.

The Bioinformatics group of C-DAC has developed a tool called MOSAIC, which stands for MOlecular Structure generator In the Cloud. MOSAIC is an OpenStack [48] cloud-based conformation search tool to explore potential energy

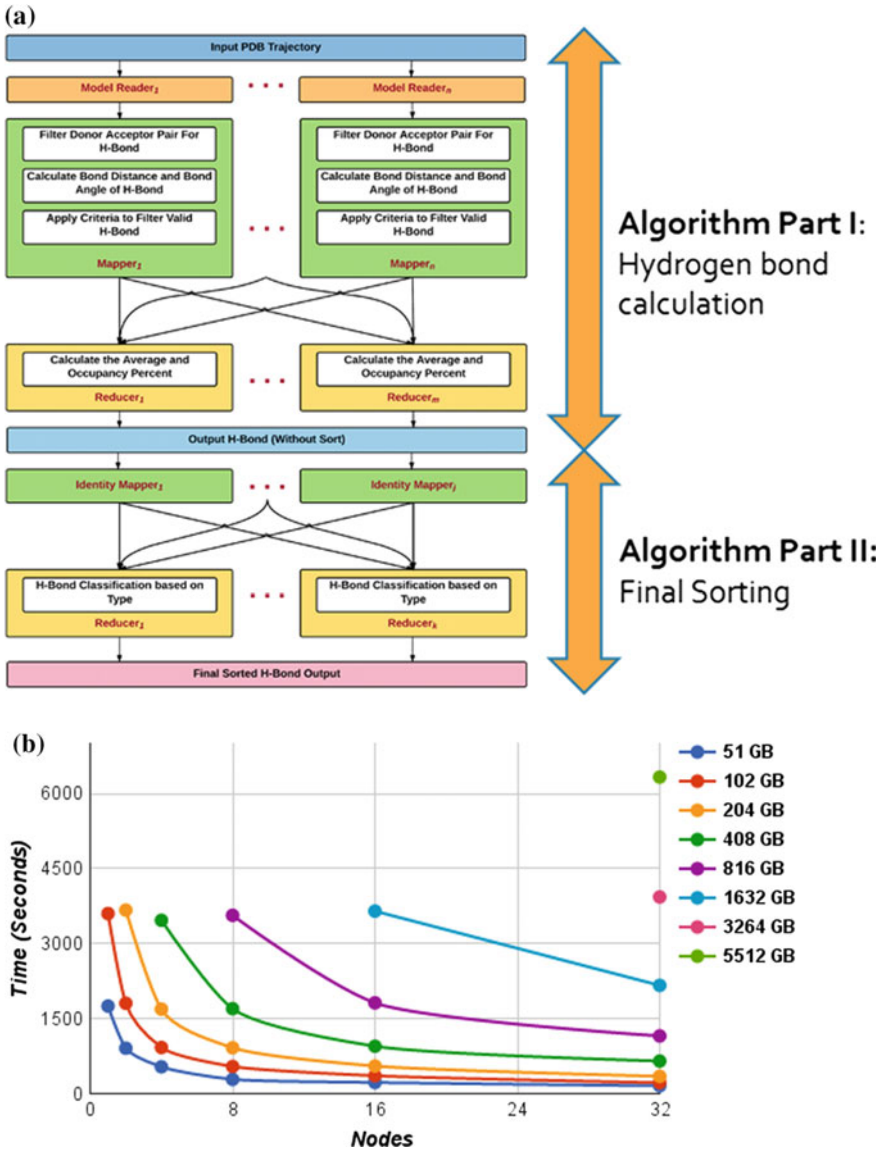

(a)
Input PDB Trajectory
Model Reader$_1$
Model Reader$_n$
Filter Donor Acceptor Pair For H-Bond
Calculate Bond Distance and Bond Angle of H-Bond
Apply Criteria to Filter Valid H-Bond
Mapper$_1$
Mapper$_n$
Calculate the Average and Occupancy Percent
Reducer$_1$
Reducer$_m$
Output H-Bond (Without Sort)
Identity Mapper$_1$
Identity Mapper$_j$
H-Bond Classification based on Type
Reducer$_1$
Reducer$_k$
Final Sorted H-Bond Output
Algorithm Part I: Hydrogen bond calculation
Algorithm Part II: Final Sorting
(b)
Time (Seconds)
0
1500
3000
4500
6000
Nodes
0
8
16
24
32
51 GB
102 GB
204 GB
408 GB
816 GB
1632 GB
3264 GB
5512 GB

Fig. 8 **a** MapReduce algorithm for H-bond calculation implemented in MapReduce paradigm **b** Benchmarking of HBAT tool for data up to 5.5 TB surface of biomolecules of interest in parallel mode using semi-empirical method. Molecular Orbital PACkage (MOPAC) is a general purpose semi-empirical molecular orbital package for the study of molecular structures and their energies [49]. The high-throughput energy calculations of the small molecules database can be done by MOPAC using hadoop and cloud technologies. Multiple instances of MOPAC are created for energy calculations of small molecules database. The tool can screen a database of millions of small drug-like molecules and understand their energetics and electrostatic behavior. The tool is useful for finding the target drug ligands. The torsion angle-driven conformational search method is useful in a range of chemical design applications [50], including drug discovery and design of targeted chemical hosts. MOSAIC has an easy-to-use interface for the bioinformatics community over Software as a Service (SaaS) platform. A user-friendly Web interface has been developed for MOPAC-based energy calculation of small molecule database. The Web interface has the capability of configuring any OpenStack-based cloud and managing multiple users to submit the jobs on dynamically created cloud VM. The Web interface has been developed using LAMP (Linux, Apache, Mysql, and PHP) framework [51]. The Web interface is shown in Fig. 9a, b. The application is deployed on OpenStack kilo version which provides platform for running the MOPAC with resources allocated virtually in the cloud. OpenStack cloud infrastructure provides scalable computational resources and scalable storage capacity.

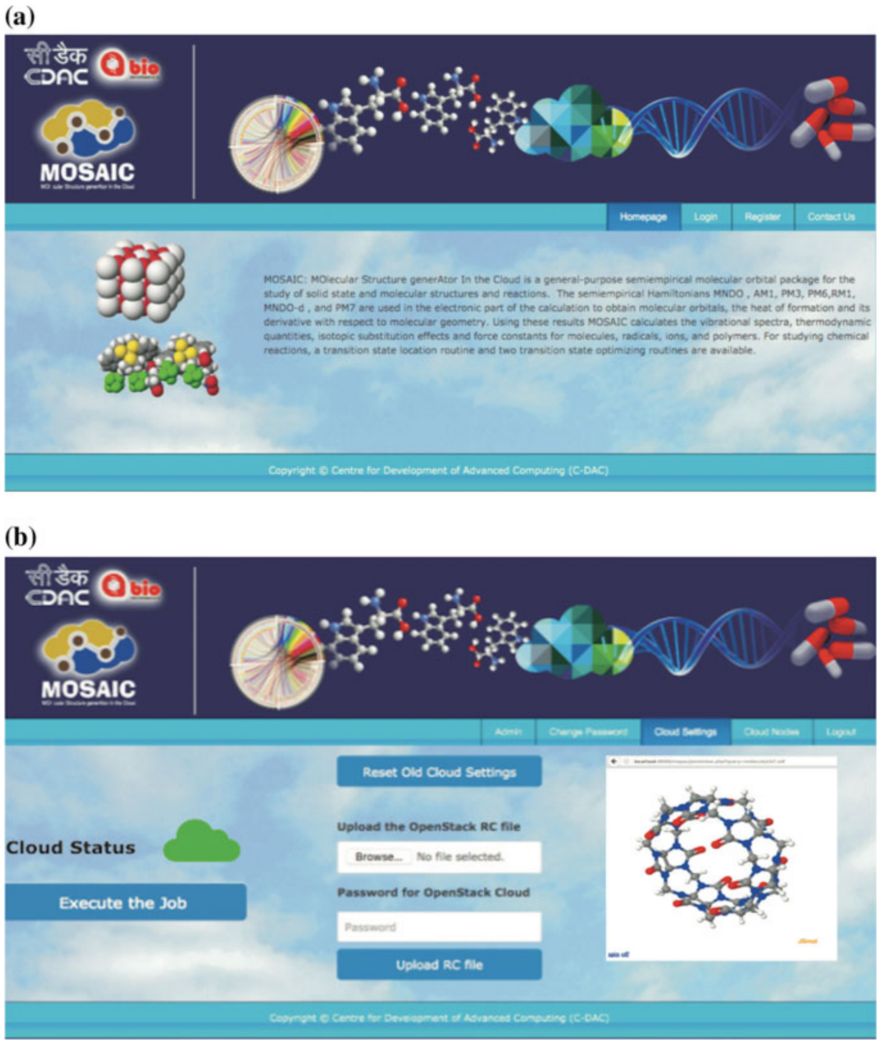
The details of cloud configurations are as follows:

The cloud infrastructure is installed using multi-nodes architecture. The cloud test bed is deployed using following configurations:

- Controller node: 1 processor, 2 GB memory, and 5 GB storage and 2 NIC.
- Network node: 1 processor, 512 MB memory, and 5 GB storage and 3 NIC.
- Compute node: 1 processor, 2 GB memory, and 10 GB storage and 2 NIC.

To synchronize the clusters, there is a need to set up NTP server. The controller node acts as NTP server, and rest of the network along with compute nodes would be synchronize with this controller node. All the nodes in the cluster except controller node have mysql client service, and on controller mysql databases have been installed. Controller node also contains the messaging server for passing message across the nodes, and we have used the RabbitMQ [52] server. The configuration is depicted in Fig. 10.

MOSAIC is executed using underlying Open Stack-based cloud to distribute millions of molecules in .mop format across the cloud nodes. The cloud nodes can be dynamically scaled to accommodate the computing load. The drug database is in the sdf format having different conformations of the same molecule and containing millions of such molecules. The sdf is converted into the desirable input file, i.e., . mop format which is used by the code for semi-empirical optimizations. The output files generated are parsed based on the energy value, and a few best optimized ligand molecules are selected based on the energy profile. The best few optimized ligands may further be scrutinized for possible drug target. This tool may have


(a)
सीडैक CDAC
bio
MOSAIC
Homepage
Login
Register
Contact Us
Copyright © Centre for Development of Advanced Computing (C-DAC)
(b)
सीडैक CDAC
bio
MOSAIC
Admin
Change Password
Cloud Settings
Cloud Nodes
Logout
Reset Old Cloud Settings
Cloud Status
Execute the Job
Upload the OpenStack RC file
Browse... No file selected.
Password for OpenStack Cloud
Password
Upload RC file
Copyright © Centre for Development of Advanced Computing (C-DAC)

MOSAIC: MOlecular Structure generAtor In the Cloud is a general-purpose semiempirical molecular orbital package for the study of solid state and molecular structures and reactions. The semiempirical Hamiltonians MNDO , AM1, PM3, PM6,RM1, MNDO-d , and PM7 are used in the electronic part of the calculation to obtain molecular orbitals, the heat of formation and its derivative with respect to molecular geometry. Using these results MOSAIC calculates the vibrational spectra, thermodynamic quantities, isotopic substitution effects and force constants for molecules, radicals, ions, and polymers. For studying chemical reactions, a transition state location routine and two transition state optimizing routines are available.

Fig. 9 **a** MOSAIC tool homepage **b** MOSAIC tool job submission page

tremendous potential in terms of ligand optimization, i.e., finding the best posture not just for one molecule but for ligand database. The tool can be easily deployable on any OpenStack-based cloud platform. MOSAIC has an easy-to-use interface for the scientific community as it abstracts the complexity of cloud-based job submission. It has a user-specific work area for managing secured private data and outputs. It has a configurable orchestration mechanism for virtual hardware configuration. The result is shared in the form of a few selected molecules favorable for drug target. It is anticipated that MOSAIC will accelerate the process of drug discovery by using high-throughput optimization of Ligand databases in parallel

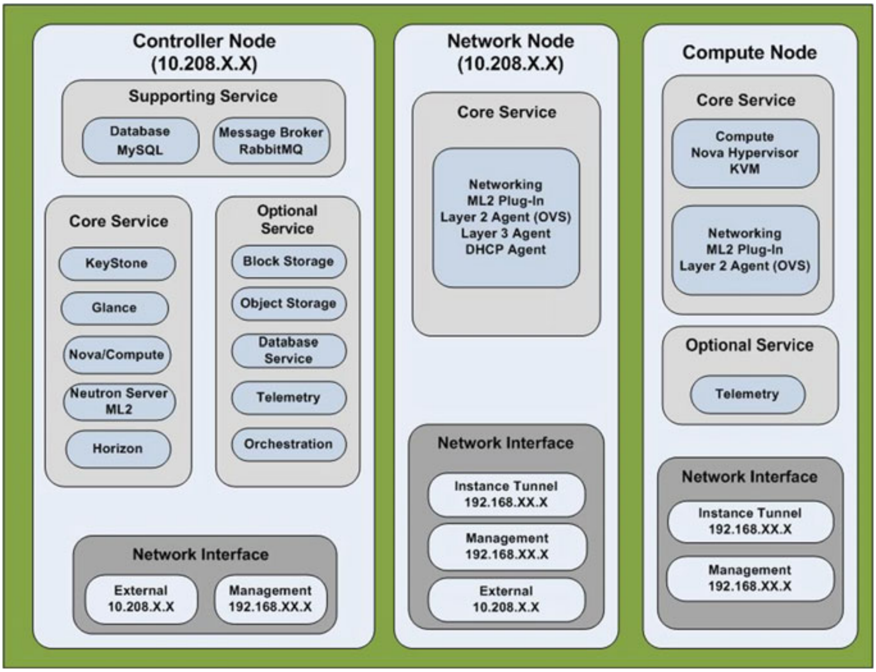


Fig. 10 Cloud configuration of MOSAIC tool

manner using distributed cloud environment. MOSAIC helps in high-throughput optimization of ligand database in parallel manner using distributed cloud environment. It will accelerate the scientific research by carrying out high-throughput virtual screening and docking in parallel manner. The tool uses the advantages of cloud computing like dynamic scaling and on-demand computing reducing the overall cost and helpful in finding optimized ligands. The workflow as discussed is shown in Fig. 11.

The tool has following features:

- Easy to use for the bioinformatics community which abstracts the complexity of cloud-based job execution.
- It is supported by a user-friendly interface with user-specific storage area with login time stamp features.
- Cloud-based high-throughput optimization of ligand database in parallel using distributed environment.
- Integrated browser-based visualization for optimized ligand molecules.
- OpenStack-based cloud environment facilitates users with on-demand scalable virtualized resources.
- Configurable orchestration mechanism for virtual hardware configuration.
- Generalized configurable solution for any OpenStack-based cloud using openrc script.

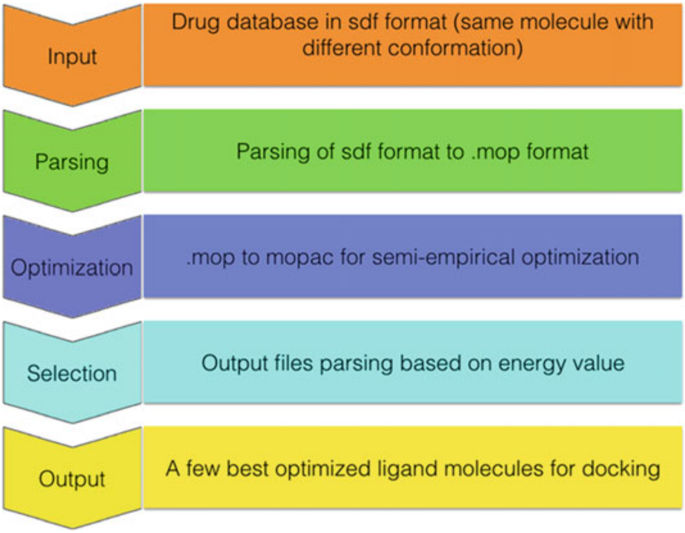

Input
Drug database in sdf format (same molecule with different conformation)
Parsing
Parsing of sdf format to .mop format
Optimization
.mop to mopac for semi-empirical optimization
Selection
Output files parsing based on energy value
Output
A few best optimized ligand molecules for docking

Fig. 11 MOSIAC tool workflow for cloud-based MOPAC implementation

5.3 Embarrassingly Parallel Molecular Docking Pipeline

Molecular docking or high-throughput screening has become increasingly important in the context of drug discovery [45]. High-throughput screening may be the only way to identify correct inhibitors of the specific target. However, high-throughput drug docking is cost-effective and very fast and could be very useful for pharmaceutical industry. An attempt has been made to develop a scalable workflow as shown in Fig. 12, for high-throughput conformational search and docking on the high-performance computing, Hadoop or cloud-based clusters. The workflow is divided into two sections. The first section performs conformational search, and the second section performs the molecular docking. The objective of the conformation search is to find the most stable conformation of the molecule along with alternative stable conformations. The semi-empirical program like MOPAC [49] is used for finding the stable structures as described in the previous section of MOSAIC. After getting the stable structures of the small molecule, docking is carried out in the parallel manner with protein of interest in the next part of the workflow. Docking of either multiple small molecules with one protein or multiple molecules with multiple proteins docking facility is available in the workflow. The testing of the workflow has been done for the drug repurposing strategy in the cancer. A test case/example of usage of this tool is given in the Sect. 6 below in the cancer K-Ras drug repurposing studies.

This tool is also deployable on any HPC, Hadoop, or cloud platform available worldwide. The current version is deployed on the computing resources of BRAF (Bioinformatics Resources & Applications Facility), C-DAC, Pune, India.

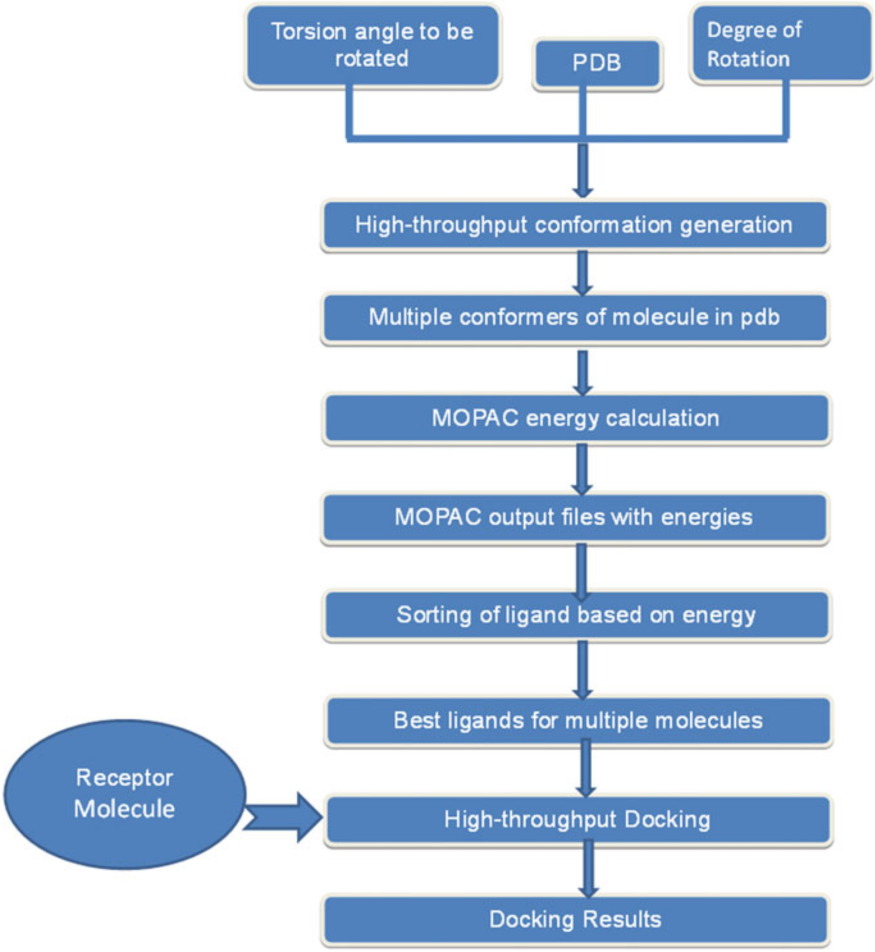

Torsion angle to be rotated
PDB
Degree of Rotation
High-throughput conformation generation
Multiple conformers of molecule in pdb
MOPAC energy calculation
MOPAC output files with energies
Sorting of ligand based on energy
Best ligands for multiple molecules
Receptor Molecule
High-throughput Docking
Docking Results

Fig. 12 High-throughput conformation generation and drug docking pipeline

5.4 Parallel Molecular Trajectories Visualization & Analytics (DPICT)

In any computational study of biomolecular systems, analysis and visualization play a pivotal role in understanding and interpretation. Molecular dynamics (MD) simulation studies of biomolecular systems, including proteins, nucleic acids, are no exceptions to this rule. The recent advances in MD techniques like REMD [53] generate multiple trajectory files whose size ranges in few gigabytes (GBs). The present-day tools often find it difficult to load a trajectory of a few GB size as it tends to occupy the entire CPU memory. The same problem is faced for loading multiple trajectories simultaneously, since most of the codes do not support parallel

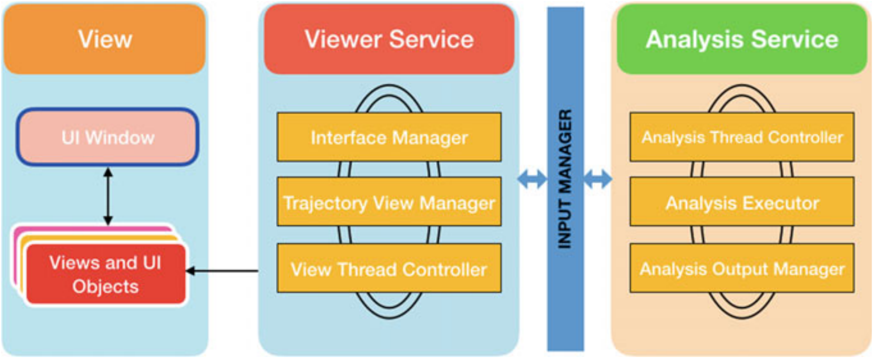


Fig. 13 Flowchart of the DPICT tool

architecture. Redundancy also occurs when the same set of calculations need to be carried out for all the trajectories individually. This often becomes a bottleneck in the research work, since recoding these programs to suit one's purpose is quite cumbersome. One often grapples around for an appropriate program/software, for analyzing and visualizing the multiple MD simulations data. And in the absence of a good program, one has to resort to writing codes and scripts. Also, loading trajectory files for visualization and analysis using the present tools often becomes extremely slow, since most of the codes are meant for serial processing and do not support multiple processors. VMD [39] tries to solve this issue by means of multi-threading, but the process becomes unresponsive when more than one trajectory is to be loaded at a time and visualized. The development of visualization and analysis tool capable of analyzing terascale and petascale data along with high-end visualization screens would accelerate the drug discovery process. Here, an attempt has been made to develop a new visualization and analysis tool capable of reading various file formats like AMBER [31], GROMACS [33] and doing most of the required analyses for a simulation in a parallel environment. The flowchart of the DPICT tool is shown in Fig. 13.

The tool has two distinct modules: one for visualization and rendering and the other for analysis of the MD simulations. The tool is an entirely GUI-based software meant to be run on Unix/Linux operating systems. The entire software tool is coded in C/C++ and OpenGL [ref] programming may be incorporated.

Features of DPICT:

- A tool to elucidate the visualization of huge molecular dynamics trajectories simultaneously for better understanding of the simulation data
- Supports visualization of nine molecules simultaneously
- Different rendering options for biomolecules like ribbon, cartoon, ball, and stick can be viewed
- Works in synchronous manner, where in nine trajectories may be handled simultaneously to perform certain operations

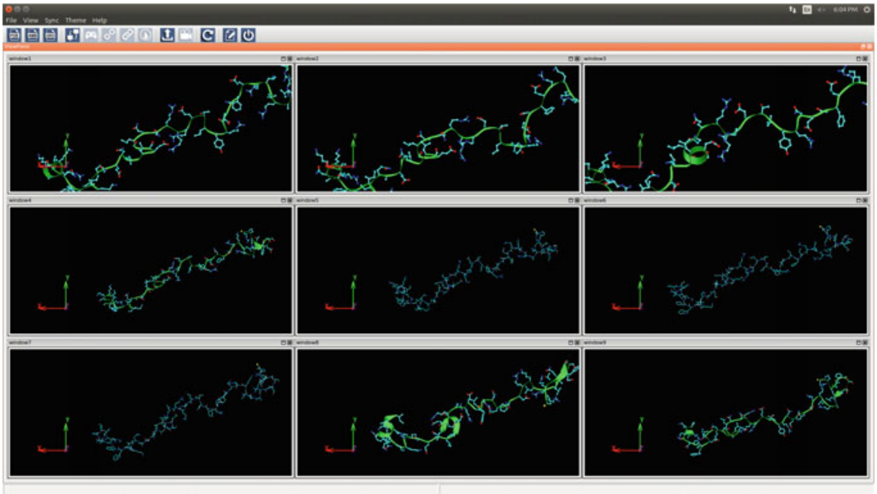


Fig. 14 DPICT tool showing simultaneous multiple trajectory visualization

- Widely used file formats of PDB, AMBER, and GROMACS are supported
- SSH feature enables the users to handle the transfer of large files from remote to local HPC clusters and vice versa.

DPICT tool in its current version is able to manage big data of multiple trajectories as shown in Fig. 14. However, future versions would be targeted to reach the goal of big data visualization.

Bioinformatics group at C-DAC has used the above tools on docking, simulations, and analytics for the drug repurposing studies for cancer protein. The details of it have been described below.

6 Drug Repurposing Study Using Big data Analytics

The drug repositioning or repurposing is a strategy to find new action mechanism of the FDA-approved drug for other disease protein than those for which it was originally intended. The repositioned drug need not go through complete drug development cycle of many years [54]. However, it can directly enter the preclinical testing and clinical trials, thereby reducing risk, time, and costs. One of the well-known examples of repurposed drug is sildenafil citrate (viagra), which was repositioned from a common hypertension drug to a therapy for erectile dysfunction [55, 56]. Similarly, use of off-label FDA-approved drugs for cancer medical practice is also known and accounts for 50–75% of drugs or biologic therapies for cancer in the USA [57, 58]. Owing to computational drug repurposing strategy, a large number of receptors can be tested with already FDA-approved drug, thereby increases the chance of identifying cure for disease within shortened time [59]. One of the proteins crucial Ras in a center pathway has been discussed as a case study.

RAt Sarcoma (RAS) protein is a crucial member of the protein family known as G-proteins. The protein Ras is encoded by one of the most common oncogene in humans. Ras belongs to GTPase class of the proteins, which possess an inherent property of GTP hydrolysis activity. Depending on its association with GDP/GTP, the protein is classified in two distinct conformations: GDP-bound inactive state and GTP-bound active state [60–62]. The malfunctioning of this protein is known to play a crucial role in human cancers, especially pancreatic cancer and various developmental disorders like Costello syndrome, Noonan syndrome [63–65]. The normal functioning of Ras plays pivotal role in the processes of cell proliferation, development, differentiation, and signal transduction [63]. The most common of the Ras mutations are found in pancreatic cancers. Most of the cancers causing mutations are reported to belong to the conserved switch (Sw I and Sw II) and GEF-binding regions of the protein. As these regions are involved in protein–protein interactions and other crucial features, and such mutations directly affect the Ras protein interaction with other proteins [66, 67]. Studies to understand the activation and deactivation Ras pathways and comparative studies of wild type and mutant have been carried out by various groups. A significant low-energy barrier in case of mutant counterparts of Ras is also well established by various experimental and computational studies. To further explore the crucial mutations and further comparison with the wild-type counterpart, computational studies are required to provide more insight about their dynamics and conformational features. Furthermore, for K-Ras which is inherently a less druggable molecule, the current trend of the drug discovery efforts is now directed toward the development of inhibitors of Ras downstream effectors. Related studies suggest that need of dual site inhibitors to effectively block oncogenic Ras signaling. Also, triple site inhibitors are also gaining more importance for improved cancer therapeutics. Considering this as a reference, simulations have been performed to explore and understand the dynamics of activation pathway of the reported hotspot mutants of Ras [68]. Similarly, the GTP hydrolysis-mediated inactivation pathways of the mutant Ras complexes have also been explored. This has helped to provide more information on the energetics of the mutant Ras complexes by calculating the energy barrier between the end states of the protein [69]. Molecular docking studies were carried out on Ras using the approach of drug repurposing with FDA-approved drug molecules database. The literature has suggested three active sites for Ras as shown in Fig. 15 where ligands can be docked [70]. The residues involved in three sites are (SITE1) residue 29–37, (SITE2) residue 68–74 and 49–57, (SITE3) residue 58–74 and 87–91. High-throughput docking has been done using the DOCK6 software employed in embarrassingly parallel molecular docking pipeline. Docking-based drug repurposing and simulation study is being carried out on four Ras systems, namely the wild type, Q61L, G12 V, and G12D mutants, each for 37 ligands. The multiple trajectories for these systems were visualized using parallel trajectory visualizer tool, DPICT. For understanding the ligand (drug candidate) properties, multiple conformations (Fig. 16) were generated using

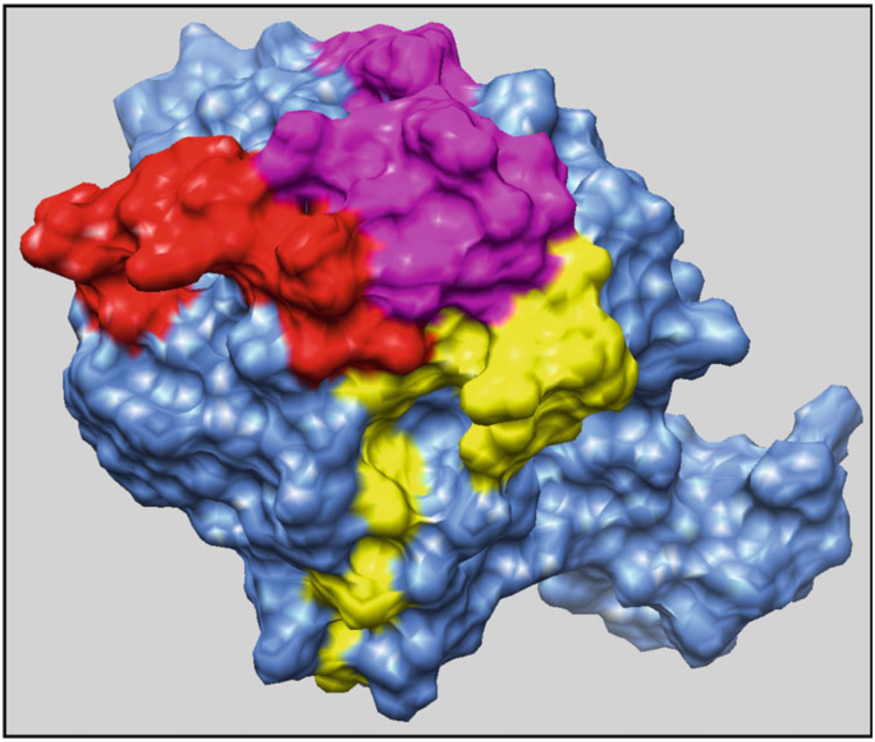


Fig. 15 KRas docking sites: SITE 1 (red): residue 29–37, SITE2 (yellow): 68–74 and 49–57, SITE3 (pink): 58–74 and 87–91

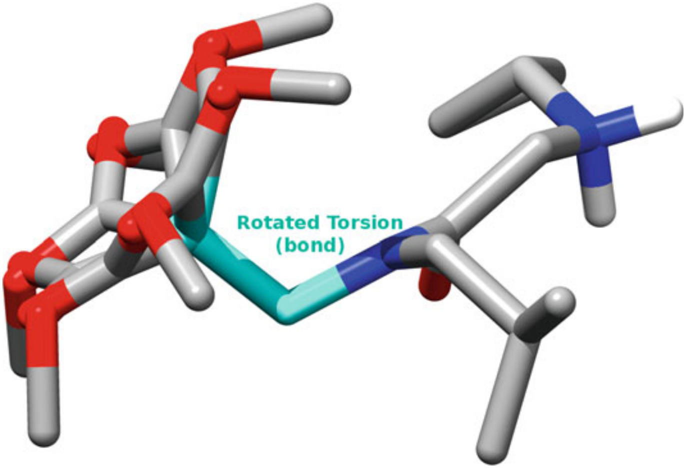


Fig. 16 Conformations generated for docking

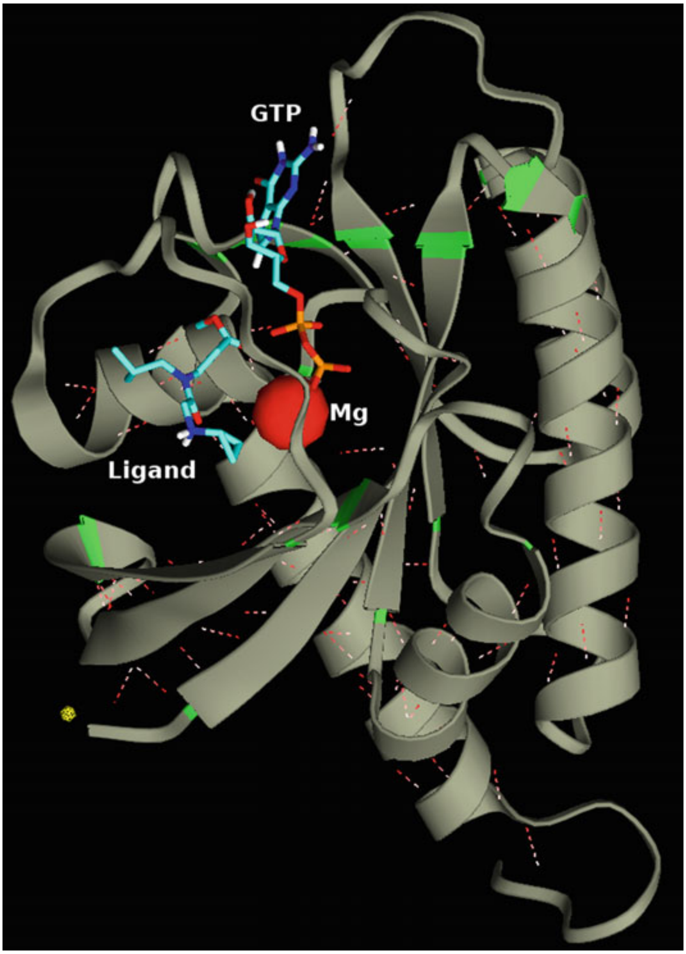


Fig. 17 KRas protein with ligand docked at SITE2

high-throughput conformation generator tool. Moreover, to study the protein–ligand complexes for the simulated systems, in-house developed tool was used. Docked pose of one of ligand is shown in Fig. 17. Preliminary analyses have been completed for the systems. The hydrogen bond and water density analyses have been performed using the in-house developed big data analytics tool, HBAT. MSM analyses are also being carried out for the same, and the results are being compared with the wild-type counterpart. Further, MD simulations were carried out for the best molecule per site in order to check the binding of the molecule with Ras (data unpublished). Classical simulations have been carried out using GROMACS software on Bioinformatics Resource and Applications Facility (BRAF). The standard protocol has been followed for minimization, heating, equilibration, and production run.

Various tools discussed earlier in this chapter have been used for parallel visualization and efficient and fast analysis of Ras docking and simulation trajectories data. In-house computational facility BRAF has been used where these tools are already deployed and tested. The results would help the experimentalist to select the better ligand for further steps of drug development.

7 Latest Development in Big data

Bioinformatics is a technology-driven science. There have been major technological shifts which are driving the data-driven science. With the ever-increasing data, the storage and analysis of huge data are becoming very tedious and most of the data remains unanalyzed. For example, the sequencing of genomes of various organisms is generating petabytes to zetabytes of data. Also, the development of new sequencing technology like nanopore is capable of producing long reads generating huge data [71]. The assembly of such genomes put out a huge challenge on the Big Data technologies. The Apache Hadoop has also enhanced to tackle such challenges like Yarn which allows different data processing engines including graph processing, stream processing as well as batch processing. The MapReduce framework provided by Apache Hadoop is good for batch processing. In case of iterative processing where the data need to be read many times, the MapReduce is not efficient. MapReduce relies heavily on disk input/output so it is slow. The Apache Spark addresses this limitation of Hadoop and provides in memory computing but reducing disk input/output. Spark supports in memory computing and optimizes disk performance by lazy loading and cache mechanism. Hence, spark is suitable for iterative computing.

Recent progressions have empowered the most precision analytics strategies at the “single cell” level. The sequencing of single cell brings about enormous volume and complexities of information and presents an extraordinary chance to comprehend the cell level heterogeneity. The latest developments highlight the inherent opportunities and challenges in Big Data analytics. The recently created technologies like erasure encoding mechanism [72] in Hadoop 3.x tend to resolve the difficulties postured by several big data problems like single cell transcriptome analysis in bioinformatics and present great opportunity to develop cutting-edge technologies for the future research problems. The HDFS uses redundancy for high availability of data. It provides great benefit at the cost of storage byte. Generally, with replication factor of 3, HDFS uses three times more storage data redundancy. So it is very costly in terms of storage. The erasure encoding mechanism in Hadoop 3.x provides same storage safety at the cost of 50% storage overhead. This is effective when data is more and its access frequency is less.

8 Conclusions

Future of medical science is to move toward personalized medicine for enhanced health care. The high-performance computing along with parallel and better algorithms would be generating volume of data from molecular docking and simulations. Advanced structural biology laboratories and techniques would also be generating different types of data. The only way which seems to be efficient in managing and analyzing such an extreme varied data may lie in the application of big data technologies. Similar kind of extreme data is being generated using advanced experimentation in life sciences in the area of agriculture for better crop production and reduced disease susceptibility and in the field of livestock to understand their genomics as well as protect them from various diseases. Data is also being generated in the field of microbes for genomics, drug discovery, vaccine ,and better environmental studies. The near future of biology/life sciences seems to be data-driven hypothesis rather than hypothesis-driven data generation, and newer computing paradigm of big data technologies may be very useful in this aspect.

References

1. Schmidt B, Hildebrandt A (2017) Next-generation sequencing: big data meets high performance computing. Drug Discov Today 22:712–717
2. Tripathi R et al (2016) Next-generation sequencing revolution through big data analytics. Front Life Sci 9(2):119–149
3. Taglang G, Jackson DB (2016) Use of “big data” in drug discovery and clinical trials. Gynecol Oncol 141(1):17–23
4. Leyens Lada et al (2017) Use of big data for drug development and for public and personal health and care. Genet Epidemiol 41(1):51–60
5. Richter BG, Sexton DP (2009) Managing and analyzing next-generation sequence data. PLoS Comput Biol 5(6):e1000369
6. Stephens ZD et al (2015) Big data: astronomical or genomical? PLoS Biol 13(7):e1002195
7. Zhao S et al (2017) Cloud computing for next-generation sequencing data analysis. In: Cloud computing-architecture and applications. InTech, London
8. Bhuvaneshwar K et al (2015) A case study for cloud based high throughput analysis of NGS data using the globus genomics system. Comput Struct Biotechnol J 13:64–74
9. da Fonseca RR et al (2016) Next-generation biology: sequencing and data analysis approaches for non-model organisms. Mar Genomics 30:3–13
10. https://www.rcsb.org/
11. Shaw DE et al (2008) Anton, a special-purpose machine for molecular dynamics simulation. Commun ACM 51(7):91–97
12. Bernardi RC, Melo MCR, Schulten K (2015) Enhanced sampling techniques in molecular dynamics simulations of biological systems. Biochimica et Biophysica Acta (BBA) 1850(5): 872–877
13. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314.1:141–151.APA
14. Swinney DC, Anthony J (2011) How were new medicines discovered? Nat Rev Drug Discov 10(7):507–519

15. Borhani DW, Shaw DE (2012) The future of molecular dynamics simulations in drug discovery. J Comput Aided Mol Des 26(1):15–26
16. Durrant JD, McCammon JA (2011) Molecular dynamics simulations and drug discovery. BMC Biol 9(1):71
17. Fabricant DS, Farnsworth NR (2001) The value of plants used in traditional medicine for drug discovery. Environ Health Perspect 109(Suppl 1):69
18. http://www.chemspider.com/
19. Wishart DS et al (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 34.suppl_1:D668–D672
20. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. J Chem Inf Model 45(1):177–182
21. Lengauer T, Rarey M (1996) Computational methods for biomolecular docking. Curr Opin Struct Biol 6(3):402–406
22. Sleigh Sara H, Barton Cheryl L (2010) Repurposing strategies for therapeutics. Pharm Med 24(3):151–159
23. Oprea TI, Mestres J (2012) Drug repurposing: far beyond new targets for old drugs. AAPS J 14(4):759–763
24. Sagiroglu, Seref, and Duygu Sinanc (2013) Big data: a review. In: International conference on collaboration technologies and systems (CTS). IEEE
25. Nayak A, Poriya A, Poojary D (2013) Type of NOSQL databases and its comparison with relational databases. Int J Appl Inf Syst 5(4):16–19
26. Hadoop A (2009) Hadoop. 2009-03-06. http://hadoop.apache.org
27. Zaharia M et al (2010) Spark: cluster computing with working sets. HotCloud 10(10-10):95
28. Allen WJ et al (2015) DOCK 6: impact of new features and current docking performance. J Comp Chem 36(15):1132–1156
29. Jones G et al (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267(3):727–748
30. Trott Oleg, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem 31(2):455–461
31. Case DA et al (2005) The Amber biomolecular simulation programs. J Comput Chem 26.16:1668–1688
32. Brooks BR et al (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30.10:1545–1614
33. Van Der Spoel D et al (2005) GROMACS: fast, flexible, and free. J Comput Chem 26(16): 1701–1718
34. Phillips JC et al (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26(16): 1781–1802
35. Rysavy SJ, Bromley D, Daggett V (2014) DIVE: a graph-based visual-analytics framework for big data. IEEE Comput Graphics Appl 34(2):26–37
36. Doerr S et al (2016) HTMD: high-throughput molecular dynamics for molecular discovery. J Chem Theory Comput 12(4):1845–1852
37. Tu T et al (2008) A scalable parallel framework for analyzing terascale molecular dynamics simulation trajectories. In: International conference for high performance computing, networking, storage and analysis. SC 2008. IEEE
38. Roe DR, Cheatham TE III (2013) PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. J Chem Theory Comput 9(7):3084–3095
39. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph Model 14(1):33–38
40. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemometr Intell Lab Syst 2(1–3):37–52
41. Genheden S, Ryde U (2015) The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov 10(5):449–461

42. Privalov PL, Crane-Robinson C (2017) Role of water in the formation of macromolecular structures. Eur Biophys J 46(3):203–224
43. Pace CN, Fu H, Lee Fryar K, Landua J, Trevino SR, Schell D, Thurlkill RL, Imura S, Scholtz JM, Gajiwala K, Sevcik J (2014) Contribution of hydrogen bonds to protein stability. Protein Sci 23(5):652–661
44. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. J Chem Inf Model 45(1):177–182
45. Yuriev E, Chalmers D, Capuano B (2009) Conformational analysis of drug molecules: a practical exercise in the medicinal chemistry course. J Chem Educ 86(4):477
46. Li J, Ehlers T, Sutter J, Varma-O'Brien S, Kirchmair J (2007) CAESAR: a new conformer generation algorithm based on recursive buildup and local rotational symmetry consideration. J Chem Inf Model 47(5):1923–1932
47. Lagorce D, Pencheva T, Villoutreix BO, Miteva MA (2009) DG-AMMOS: a new tool to generate 3D conformation of small molecules using distance geometry and automated molecular mechanics optimization for in silico screening. BMC Chem. Bio 9(1):6
48. Sefraoui O, Aissaoui M, Eleuldj M (2012) OpenStack: toward an open-source solution for cloud computing. Int J Comput Appl 55(3):38–42
49. Stewart JJP (1990) MOPAC: a semiempirical molecular orbital program. J Comput Aided Mol Des 4(1):1–103
50. Hawkins PC, Skillman AG, Warren GL, Ellingson BA, Stahl MT (2010) Conformer generation with OMEGA: algorithm and validation using high quality structures from the protein databank and cambridge structural database. J Chem Inf Model 50(4):572–584
51. Ware B (2002) Open source development with LAMP: using Linux, Apache, MySQL and PHP. Addison-Wesley Longman Publishing Co., Inc., Reading
52. https://www.rabbitmq.com/
53. Hukushima K, Nemoto K (1996) Exchange Monte Carlo method and application to spin glass simulations. J Phy Soc Jpn 65(6):1604–1608
54. Ashburn TT, Thor KB (2004) Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 3(8):673–683
55. Novac Natalia (2013) Challenges and opportunities of drug repositioning. Trends Pharmacol Sci 34(5):267–272
56. Smith Kelly M, Romanelli Frank (2005) Recreational use and misuse of phosphodiesterase 5 inhibitors. J Am Pharm Assoc 45(1):63–75
57. Pfister DG (2012) Off-label use of oncology drugs: the need for more data and then some. J Clin Oncol, 584–586
58. Jin G, Wong STC (2014) Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. Drug Discov Today 19(5):637–644
59. Neves SR, Ram PT, Iyengar R (2002) G protein pathways. Science 296(5573):1636–1639
60. Khrenova MG et al (2014) Modeling the role of G12 V and G13 V Ras mutations in the Ras-GAP-catalyzed hydrolysis reaction of guanosine triphosphate. Biochemistry 53(45): 7093–7099
61. Spoerner M et al (2010) Conformational states of human rat sarcoma (Ras) protein complexed with its natural ligand GTP and their role for effector interaction and GTP hydrolysis. J Biol Chem 285(51):39768–39778
62. Ma J, Karplus M (1997) Molecular switch in signal transduction: reaction paths of the conformational changes in ras p21. Proc Natl Acad Sci USA 94(22):11905–11910
63. White MA et al (1995) Multiple Ras functions can contribute to mammalian cell transformation. Cell 80(4):533–541
64. Schubbert S, Shannon K, Bollag G (2007) Hyperactive Ras in developmental disorders and cancer. Nat Rev Cancer 7(4):295
65. Gao C, Eriksson LA (2013) Impact of mutations on K-Ras-p 120GAP interaction. Comput Mol BioSci 3(02):9
66. Shurki A, Warshel A (2004) Why does the Ras switch "break" by oncogenic mutations? Proteins: Struct Funct Bioinf 55(1):1–10

67. Lu S et al (2016) Ras conformational ensembles, allostery, and signaling. Chem Rev 116(11): 6607–6665
68. Sharma N, Sonavane U, Joshi R (2017) Differentiating the pre-hydrolysis states of wild-type and A59G mutant HRas: an insight through MD simulations. Comput Biol Chem 69:96–109
69. Sharma N, Sonavane U, Joshi R (2014) Probing the wild-type HRas activation mechanism using steered molecular dynamics, understanding the energy barrier and role of water in the activation. Eur Biophys J 43(2-3):81–95
70. Wang W, Fang G, Rudolph J (2012) Ras inhibition via direct Ras binding—is there a path forward? Bioorg Med Chem Lett 22(18):5766–5776
71. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich, SB (2010) The potential and challenges of nanopore sequencing. In: Nanoscience and technology: A collection of reviews from Nature Journals, pp 261–268
72. https://hadoop.apache.org/docs/r3.0.0/hadoop-project-dist/hadoop-hdfs/HDFSErasureCoding.html