Original Article

DUSR (Distributed Ultrafast Shape Recognition): a Hadoop Based Tool to Identify Similar Shaped Ligand Molecules

Vandana Kumari, Rashmi Tripathi*, Sunil Patel, Utkarsh Raj, Pritish Kumar Varadwaj*

Department of Bioinformatics, Indian Institute of Information Technology-Allahabad, INDIA

ABSTRACT

Background: Identifying potential drug candidates through Ligand-based virtual screening is often associated with processing of huge amount of data and hence is a computational intensive task. Ultrafast Shape Recognition (USR) algorithm has been reported as a faster alternative for molecular shape comparison which maps the chemical structure of query ligand into its shape moment vector to find novel chemical scaffolds in chemical compound libraries. The USR algorithm however was devoid of the ability to discriminate ligand molecules according to their pharmacokinetic features. **Methods:** To overcome this discrepancy, a modification in the existing USR algorithm called **DUSR (Distributed Ultrafast Shape Recognition)** was carried out where chemical compounds were screened on the basis of their drug-likeliness properties prior to the molecular shape comparison followed by shape complementarity momentum measure. The DUSR due to its Hadoop implementation acts as a faster approach than the existing standalone tools, utilizing the MapReduce algorithm supporting the high throughput screening of million conformers in a much reduced time span. We further demonstrated the utility of DUSR on dataset of 2 million ligand molecules by running shape comparison based searching job on standalone and multisystem Hadoop platforms. **Results:** The result suggested that DUSR completed its job in 1h 15 m 41s, 0 h 23 m 41s and 0h13m 22s sec for 2038924 molecules on Hadoop standalone mode, 3-nodes cluster & 5-nodes cluster of distributed commodity hardware respectively.

Key words: DUSR (Distributed Ultrafast Shape Recognition), High throughput Screening, Hadoop, MapReduce, Virtual Screening.

INTRODUCTION

The conventional drug discovery process involves identification of novel drug like lead candidate through series of trial and error checks which are very tedious, cost ineffective and time consuming processes, typically requiring time span of 10-15 years and cost in range from $800 million to $1 billion. The complexity of the process can be reflected by estimating the rate of approval of these compounds which are one among thousands that enters Research and Development (R&D) pipeline and further to get the FDA approval. Therefore, to make the process accelerated and less expensive, it has become a common practice to use computational methods for docking *viz. in silico* ADME (Absorption, Distribution, Metabolism, and Elimination) studies and virtual high throughput screening of large compounds libraries to identify appropriate lead molecules. These practices of finding similar structure ligands through matching ligand candidates from backend chemical library with a query ligand is called virtual screening (VS) or to be more précised it is also known as high throughput virtual screening (HTVS).[1,2] VS pose as an effective and feasible computational measure to reduce the expensive biological tests and to tackle high fail-

Submission Date: 07-07-2016;
Revision Date: 06-09-2016;
Accepted Date: 29-09-2016

DOI: 10.5530/ijper.51.1.6

Correspondence:
Rashmi Tripathi[1],
Department of
Bioinformatics,
Indian Institute of Information
Technology
Allahabad, INDIA.
Ph.no: +91- 5322922219,
+91-7291048638
Email: rs154@iiita.ac.in
Email: kvandana1992@
gmail.com

ure rate faced by conventional drug designing strategies. VS is also helpful in similarity scoring and sub-structure searching, quantitative structure-activity relationship (QSAR), pharmacophore and three-dimensional ligand shape matching. Various tools which were proposed for VS includes Autodock,[3] PharmDock,[4] and LigBuilder,[5] LIGSIFT,[6] eSimdock,[7] Align-it,[8] *etc.* which focuses on ligand based shape similarity measures. 1D and 2D methods for representing chemical compounds may successfully identify chemical analogues but are unsuccessful in identifying differential activities among them. It is quite obvious that molecules having same shape and drug likeliness properties could be potent for same target. Therefore, shape based VS possessing 3D structural information tools (*e.g.* USR,[9] USRCAT[10]) are gaining popularity as compared to the previous existing models based on 1D and 2D information for the molecular shape comparison of molecules. However ligand based virtual screening (LBVS) solely depends on the structure of reference ligands which may be aligned or screened against a chemical scaffold databases.[11-15]

Depending on the structural properties and available information to be mined these VS methods can be safely classified as follows:

1. Atomic distance based method pose as one of the fastest and simplest way to do VS by comparing the atomic distance between pairs of atom. USR (Ultra-fast Shape Recognition)[1] and ESHAPE3D[9] are two existing methods which computes the statistical distances between the atoms for describing 3D shape complementarity. USR uses position of molecule center (CG), closest and farthest atom from CG by utilizing the distributions (distances) for calculating three moments (mean, variance, skewness). Each molecule therefore can be described as a unique 12 dimensional moment vector space representing structural intricacy of corresponding ligand 3D shape. Whereas ESHAPED3D utilizes the distance matrix calculated through the Eigen values characterizing the molecular shape of the compounds. Therefore it computes the fingerprint scores to measure the shape similarity.
2. Gaussian function based methods are molecular volume based shape recognition algorithms that generates the superimposition atomic density through spherical Gaussian function to calculate the similarity score by estimating volume of the query molecule. ShaEP[16] uses transformation matrix derived from aligned molecular feature frames portraying a graph bearing vertices. Each vertex is being represented by heavy atoms and hybridized orbital which do not participate in covalent bonding. Another Gaussian function based method ROCS[17] calculate the similarity score by capturing the chemical similarity between the superimposed molecules.
3. Surface based method tends to explore the shape of a ligand molecule through series of surface roughness descriptors utilizing the atomic position coordinate information. MSMS,[18] MOLPRINT3D[19] and BetaDock[20] are the well known surface-based molecular shape descriptor models. The MSMS represents a molecular surface by joining the vertices by means of triangular patches. Whereas, MOLPRINT3D offers a slight modification in MSMS algorithm by assigning energy values to each of the surface point to categorize the molecules through respective surface point vectors. BetaDock is another approach based on β-shape representation using a Voronoi diagram; where unlike a Voronoi diagram based method (α-shape), it is capable of recognizing the shape of heavy atoms with variable radii.
4. Field based methods primarily compares the molecular field for 3D space around a target molecule, unlike the surface based and volume based method where the molecular properties are possessed by the atoms and surface points specifically. Apart from the electrostatic properties, Van der Waals potential and hydrophobicity are also considered in this method. BRUTUS[21] is a field based method which computes electrostatic potential to identify similar molecules based on empirical charges.
5. Pharmacophore based method utilizes pharmacophoric spatial geometric knowledge in combination with presence and absence of hydrogen bonding, hydrophobicity core, anion and cation *etc.* FLAP[22] and Tuplets[23] are pharmacophore VS approaches for identifying identical features through discrete point GRID of potential energy by combining 4-feature points into a form of quadruplets. Hence the quadruplets of a ligand and database are compared against the candidates from chemical database. Unlike FLAP, Tuplets also uses Cosine similarity in place of Tanimoto coefficient for calculating similarity score.

However, the reported atomic distance based method, Gaussian function based method, surface based method, field based method and pharmacophore based method often fails in handling the exponentially increasing volume of chemical scaffold data. To overcome this problem of searching the ever increasing chemical space with vast amount of data within a reasonable time span, the concept of big data analytics comes as a handy tool. In this work we are proposing parallelization of USR algorithm using Hadoop framework to harness the processing power of cluster of commodity hardware.

Hadoop an open source implementation is an effective platform for storing, handling and investigating large scale data, through arena of reliable, scalable, distributed computing software tools. The principle of dividing large datasets into smaller blocks distributing across clusters of commodity computers by Hadoop programming libraries has provided simple programming models. Furthermore, the library keeps a check on any kind of failures at the application layer, hence ensuring the quality-based services on low-end hardware. Hadoop made it possible for a large number of machines to scale up their processing through chain of multiple local computation and storage units.[24] A highly beneficial framework is an *Ecosystem* comprising of two main architectural components viz. MapReduce and a Distributed File System (HDFS). MapReduce is responsible of parallel computing whereas HDFS is responsible for data management. Hadoop partitions a job into tasks, blocks them respectively, and assigns them to particular nodes in a form of clusters. Hadoop adopts master/slave architecture, in which a master node manages other slave nodes in the cluster. There may be multiple masters in the model for failure protection. In the Map Reduce model, the master is called Job Tracker, and each slave is acknowledged as Task Tracker. In the HDFS, the master is called Name Node, and each slave is called Data Node. Job and data distributions are managed by the master to assign nodes for computing and storing.

The distributed architecture of Hadoop greatly enhances the processing efficiency by using numerous general PCs that can build a high performance computing environment. No modifications in hardware is required while installing the software in spite of some possible changes to meet minimum recommended RAM, disk space *etc* needed on each cluster, written in JAVA and is an extremely scalable framework. Map/reduce model from Google is used in its open source implementation in Hadoop. MapReduce frames parallelize problems as a set of two functional phases: Map phase and Reduce phase. Map phase performs filtering, shuffling and sorting, whereas reduce phase performs a summary operation. Each phase has key-value pairs as input and output; the types of which may be chosen by the programmer. The programmer also specifies two functions: the map function and the reduce function.

MATERIALS AND METHODS

Apache Hadoop is an open source software framework that permits distributed analysis of enormous amount of data on a large scale and can be installed on commodity Linux cluster. We have used the standard Hadoop HDFS Java API (included in the Hadoop distribution), which allows the user to create and run jobs with any of the executable- the mapper and/or the reducer. In Hadoop, input values (with the default formatter) are lines of text read from one or more files (input keys are discarded).

Datasets

The data for the research work was downloaded from freely available database of commercial compounds, ZINC database (http://zinc.docking.org/), of varying sizes (500 MB, 1GB, 2GB, 3GB, 4GB, 5GB, 7GB and 9 GB) having 127895, 231586, 462624, 693673, 924595, 1155335, 1618035 and 2038924 molecules respectively.[25] Further, we have used two features of Chemistry Development Kit (CDK):[26] 3D geometry generation and QSAR descriptor calculation. CDK is an open source JAVA library for cheminformatics and bioinformatics data.

USR molecular shape comparison method

USR screening of ligands utilizes three dimensional coordinate points to construct 12 dimensional shape moment vector with distance measure to find the similarity between given pair of molecules. The coordinate points of individual ligand molecules were parsed from. sdf files, obtained from ZINC database. Each molecule coordinate points were processed further to calculate-molecular centroid (ctd), the closest atom from the ctd (cst), the farthest atom from the ctd (fct), and the farthest atom from the fct (ftf). Further the atomic distance of rest of the atoms of the query molecule for the four molecular locations *i.e.* ctd, cst, fct and ftf were calculated as shown in Figure 1. The network of atomic distances of all atoms of the individual molecule, from these four locations namely $\overline{X1}, \overline{X2}, \overline{X3}, \overline{X4}$ signifies mean distance from ctd, cst, fct and ftf respectively. $\overline{M1}, \overline{M2}, \overline{M3}$ corresponds to first, second and third moment of distributions of $\overline{X1}$. First moment of distribution, $\overline{M1}$, yields an approximation for molecular size. The second moment of distribution, $\overline{M2}$, which is also known as variance of the distribution, gives information about molecule denseness. The third moment or skewness of distribution, $\overline{M3}$, represents molecular asymmetry. Similarly, all nine descriptors positions are calculated from, $\overline{X2}$, $\overline{X3}$ and $\overline{X4}$. Hence, the shape of each individual molecule was encoded into 12 descriptors to form the corresponding moment vector as shown in Figure 2.

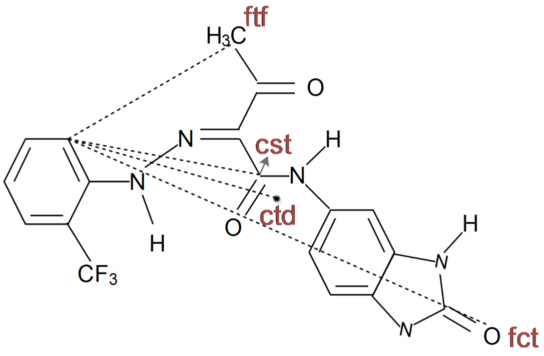

Figure 1: (A) Represents four locations: molecular centroid (ctd), nearest atom from ctd (cst), farthest atom from ctd (fct) and farthest atom from fct (ftf) in a molecule and (B) Atomic distances of an atom from these locations.

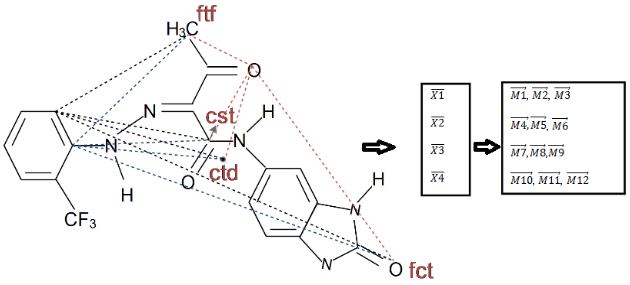
ftf
H_3C
O
N
cst
H
N
N
ctd
O
H
H
N
CF_3
N
O fct
$\overline{X1}$
$\overline{X2}$
$\overline{X3}$
$\overline{X4}$
$\overrightarrow{M1}, \overrightarrow{M2}, \overrightarrow{M3}$
$\overrightarrow{M4}, \overrightarrow{M5}, \overrightarrow{M6}$
$\overrightarrow{M7}, \overrightarrow{M8}, \overrightarrow{M9}$
$\overrightarrow{M10}, \overrightarrow{M11}, \overrightarrow{M12}$

Figure 2: Represents network of atomic distances of some atoms from these four locations. *X*1, *X*2, *X*3 and *X*4 signifies mean distance of all atoms from ctd, cst, fct and ftf respectively. The 12 descriptors are three moments of vectors generated for each of the four distributions.

However, for molecules having similar shapes but different pharmacophoric group, USR becomes inefficient. So to eliminate this pitfall, an initial screening of compounds, prior to molecular shape comparison method, was carried out on the basis of their drug-likeliness properties. These properties includes molecular mass (< 500 Daltons), number of hydrogen bond donor atoms (< 5), number of hydrogen bond acceptor atoms (< 5), partition coefficient (x log P < 5.0), total polar surface area (< 140 Å) and number of rotatable bonds (<10).

Unlike other ligand-based shape overlapping methods, the main feature of our approach is to first filter the molecules using initial screening phase and then calculating its Distance Moment Vector [1*12] in size which enhances the effectiveness of our virtual screening approach. Use of Similarity function Distance Moment Vector of active molecule was also generated and stored [1*12]. Thus, the molecule from a chemical database have $\overrightarrow{M1_1}, \overrightarrow{M1_2}, ...\overrightarrow{M1_{12}}$ as its 12 descriptors and each descriptors active molecules are represented as $\overrightarrow{MA_1}, \overrightarrow{MA_2}, ...\overrightarrow{MA_{12}}$. The similarity score between these two molecules is given in Equation.

$$\text{Similarity Score} = \frac{{}^{12}_{i=1}\overrightarrow{M1_l} \quad \overrightarrow{MA_l}}{12} \qquad \text{Eq: 1}$$

Hadoop implementation

The proposed method was implemented in two modules DistMapVector (DMV) and DistMapVectorLibScreen (DMVLS). In DMV, a MapReduce job was written consisting of both mapper and reducer phase. In this method InputFormatClass was used with SDFInputFormat which reads multi-line records from SDF file. The map input key is an object that contains a molecule in SDF format and input value class in text format, *i.e.* the whole content of a molecule remains in SDF format. Initial screening of the compounds on the basis of Lipinski's rule was carried out in mapper phase, which was further followed by the distance similarity score calculation between moment vectors on the basis of Eq.1. The output key and value of the mapper phase was the compound IDs and the corresponding distance similarity scores generated respectively. The output key and value of mapper phase acts as an input key and value of the corresponding reducer phase. In this phase, the final screening of molecules was carried out on the basis of threshold value of distance similarity score. The output of reducer phase (MapReduce job) was the final screened compounds with their respective IDs and similarity scores. Figure 3 shows the working and control flow of DMV module. In DMVLS, two MapReduce jobs were written, where former job deals with mapper phase while the later has both mapper and reducer phase. In mapper job, SDFInputFormat as described in DMV was used as InputFormatClass. In mapper phase, screening of the molecules were performed in accordance with the Lipinski's rule followed by generating DMV of these screened molecules. The output file of the mapper phase acknowledged as 'Distance moment vector library' (DMVL) consists of molecule ID's and their corresponding distance moment vector.

Reducer phase further uses the output files of the mapper phase job as the input files. The work of mapper phase is to map the compound ID's and their corresponding distance moment vector followed by finding similarity score of these compounds with reference to the active compounds. The reducer phase further screens the molecules on the basis of similarity scores and finally calling this job as '*Screening*'. The control flow for DMVLS module is explained in Figure 4.

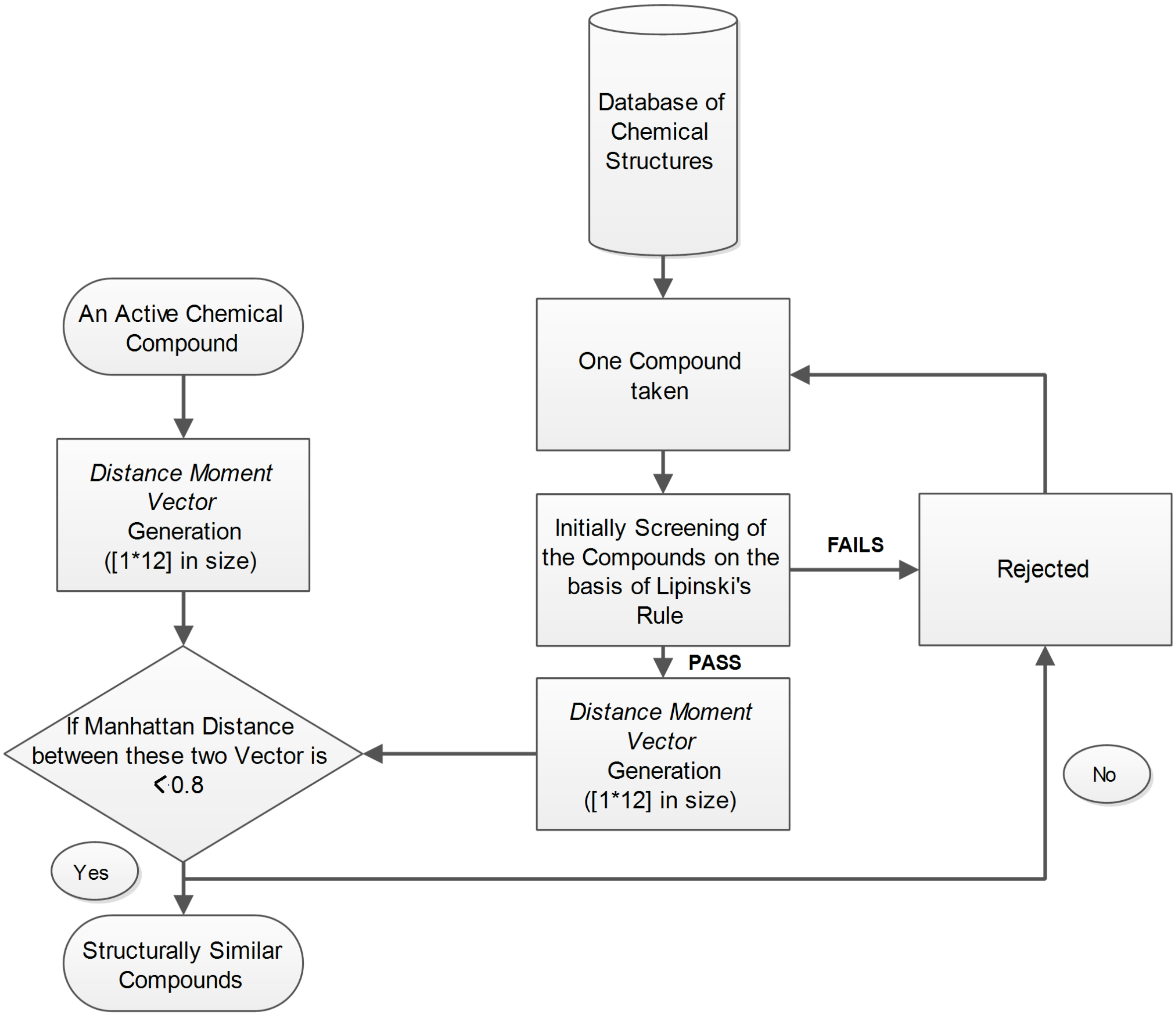
Database of Chemical Structures
An Active Chemical Compound
One Compound taken
Distance Moment Vector Generation ([1*12] in size)
Initially Screening of the Compounds on the basis of Lipinski's Rule
FAILS
Rejected
PASS
Distance Moment Vector Generation ([1*12] in size)
If Manhattan Distance between these two Vector is <0.8
No
Yes
Structurally Similar Compounds

Figure 3: A flowchart representing the implementation of DistMapVector (DMV) approach. MapReduce involves two components: (1) Mapper phase and (2) Reducer phase. The Initial screening of compounds on the basis of Lipinski's rule & Distance Moment Vector Generation comprises of the Mapper phase and checking if the Manhattan distance between two vectors passes the similarity score or not is a part of Reducer phase.

RESULTS AND DISCUSSION

The discussed DMV and DMVLS modules were performed on single node, 3 nodes and 5 nodes clusters with SDF files of 0.5GB, 1GB, 2GB, 3GB, 4GB, 5GB, 7GB and 9GB containing 127895, 231586, 462624, 693673, 924595, 1155335, 1618035 and 2038924 ligands molecules respectively. Each of these nodes was equipped with four quad-core intel3 processor and 6 GB RAM. All nodes were connected through Fast Ethernet (100 MB/sec). The HDFS block size was fixed to 128 MB, therefore blocks of various sizes 4, 8, 16, 24, 32, 40, 56 and 72 were generated from above dataset.

Figure 5 shows the run time against the number of molecules for JAVA and different modes of Hadoop (standalone mode, fully distributed 3-node and fully distributed 5-node clusters). As it is shown in Fig. 5, that execution time is reduced significantly with increasing number of nodes. However, the time taken to run the job in JAVA and Hadoop standalone mode are at par whereas there was no notable difference in the execution time for running 3-node and 5-node clusters using Hadoop fully distributed mode for 500 MB of data because of its small size. Figure 6 shows time taken to complete '*Distance moment vector library'* job in JAVA and different modes of Hadoop *via* DMVLS.

Figure 6 shows the run time against the number of molecules of Distance Moment Vector job for different modes of Hadoop (standalone mode, fully distributed 3-node and fully distributed 5-node clusters). The execution time was reduced significantly with increasing number of nodes as shown in Figure 6, which suggest the fact that a 3-node cluster is 2.5 times faster than standalone mode whereas 1.5 times slower than 5-node cluster.

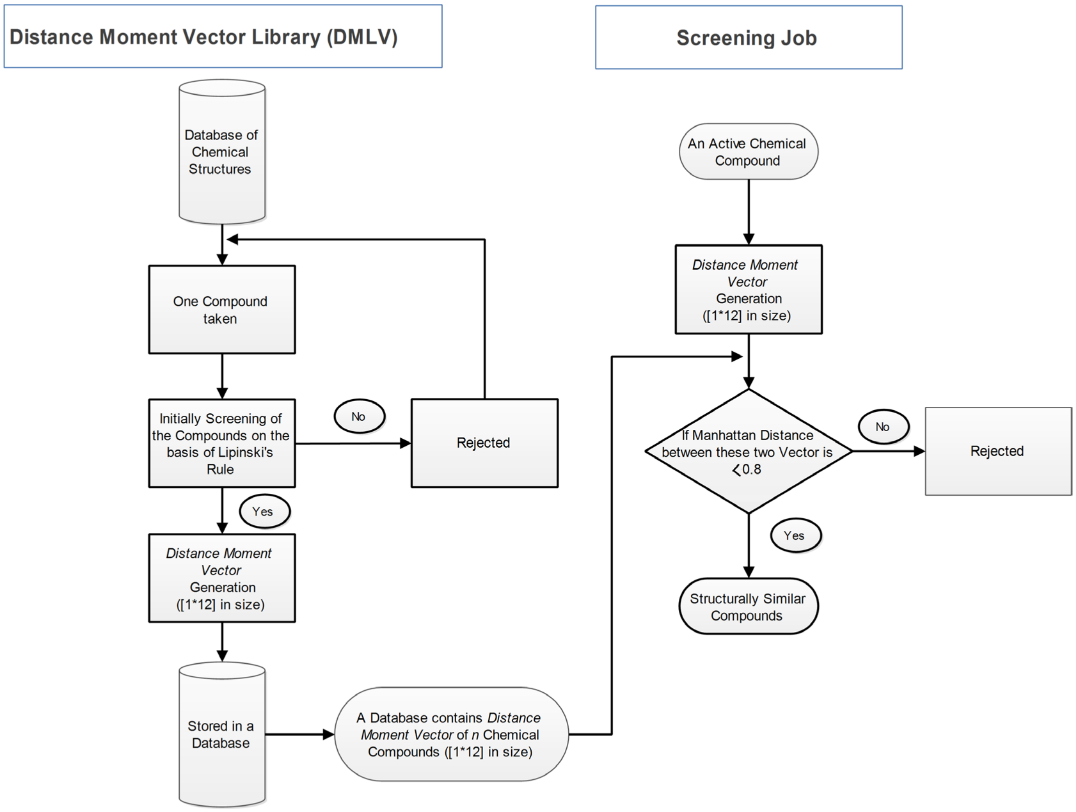
Distance Moment Vector Library (DMLV)
Screening Job
Database of Chemical Structures
One Compound taken
Initially Screening of the Compounds on the basis of Lipinski's Rule
No
Rejected
Yes
Distance Moment Vector Generation ([1*12] in size)
Stored in a Database
A Database contains Distance Moment Vector of n Chemical Compounds ([1*12] in size)
An Active Chemical Compound
Distance Moment Vector Generation ([1*12] in size)
If Manhattan Distance between these two Vector is <0.8
No
Rejected
Yes
Structurally Similar Compounds

Figure 4: A flowchart representing implementation of DistMapVectorLibScreen (DMVLS) approach. The Initial screening of compounds on the basis of Lipinski's rule & Distance Moment Vector Generation comprises of the Mapper phase and checking if the Manhattan distance between two vectors passes the similarity score or not is a part of Reducer phase.

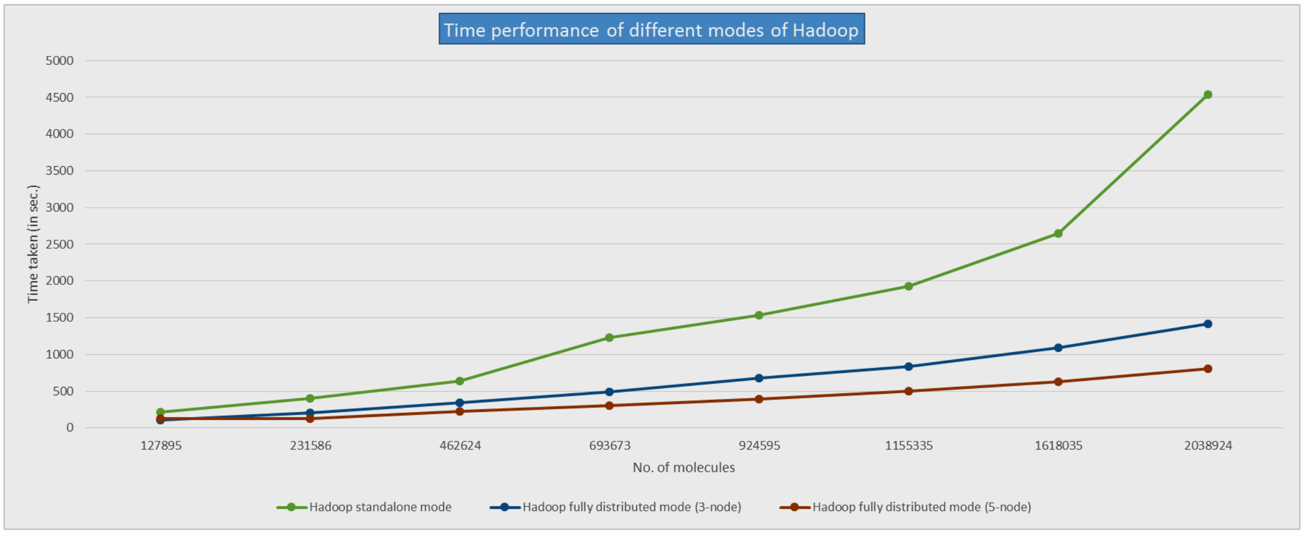

Figure 5: Time performance of JAVA and different modes of Hadoop *via* DMV.

Figure 7 shows the graphical representation of completion time against number of molecules for the Screening job in DMVLS for different modes of Hadoop. Figure 7 also suggested the fact that the performance time decreases with increasing cluster size. Running DMVL job for the first time builds a library which can be repeatedly used for running Screening jobs for different reference active molecules. This advantage of using DMVLS over DMV further minimize the time span of searching as it eliminates the process of mapping the ligand molecules into vector space, hence making it a faster alternative for the VS. For further scientific validation of the software, standard virtual screening benchmarking dataset from DUD-E

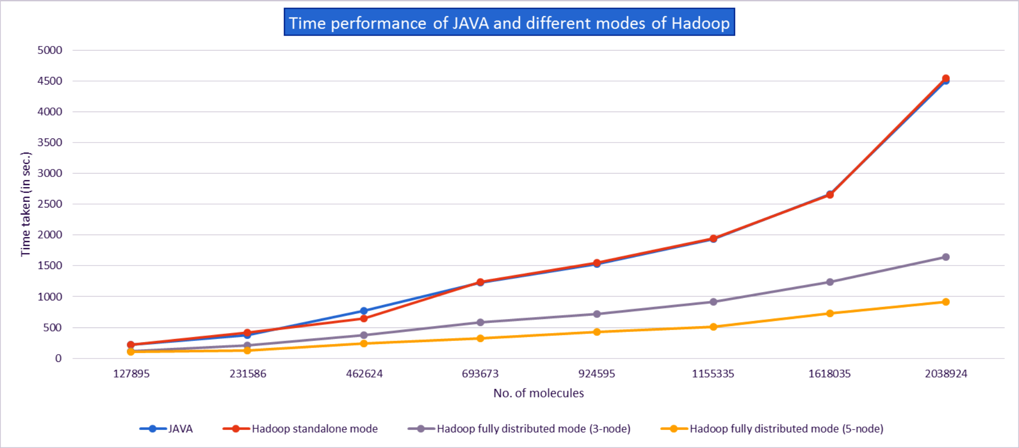

Figure 6: Time performance of DMVL job in different modes of Hadoop *via* DMVLS.

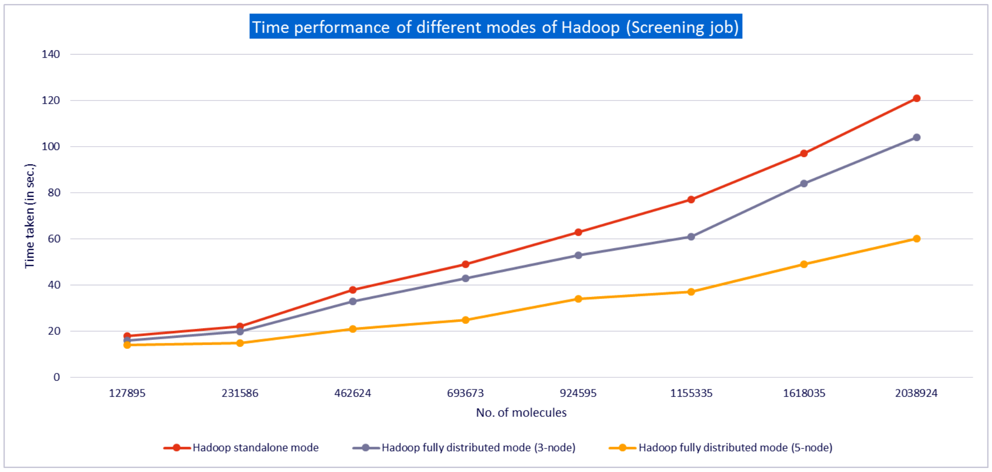

Figure 7: Time performance of 'Screening' job in different modes of Hadoop *via* DMVLS.

Table 1: Shows time taken to complete job in JAVA and different modes of Hadoop *via* DMV

Number of molecules	Time taken to complete job (in secs.)			
	JAVA	Hadoop standalone mode	Hadoop fully distributed mode (3-node cluster)	Hadoop fully distributed mode (5-node cluster)
127895	219	219	113	110
231586	376	420	206	126
462624	770	645	380	236
693673	1231	1238	582	321
924595	1533	1550	722	424
1155335	1936	1941	917	515
1618035	2660	2652	1239	732
2038924	4504	4549	1645	912

(Directory of Useful Decoys- Enhanced) was used and the performance of the virtual screening approach was tested (http://dude.docking.org/). Mapping and screening of 56692 number of ligand molecules was done. Using the framework, molecules were screened in about 29 seconds on Hadoop standalone mode.

CONCLUSION

In the present report, a modification to the existing USR algorithm was performed with modalities to screen ligand on basis of their pharmacophoric properties using a Hadoop based framework. DUSR showed better performance as compared to the previously used method in terms of screening those compounds which have drug-likeliness properties and showed structural similarities. The generation of novel and robust scoring function using Hadoop not only is useful for ligand-based virtual screening algorithms, but can also be implemented on other screening methods which may open the door for the ranking of docking results. In addition, although DUSR performed better than non-Hadoop based applications, more accurate result can be attained if users give their own input (*eg*. drug-likeliness properties) which can be achieved by modifying the values of the variants in the existing program which has been made available on http://bioserver.iiita.ac.in/dusr. As a result, efficient and reproducible screening workflows can now be implemented at lower cost and effort making preclinical drug research projects faster and feasible. The efficiency and accuracy of the Hadoop based MapReduce framework makes the DUSR method perfectly suitable for its usage on the large set of data with millions of molecules. The source code of the DUSR was released as a JAVA module under Hadoop framework and can be downloaded at http://bioserver.iiita.ac.in/dusr. Such significant improvement in efficiency will be source for structurally similar molecules derived from a well-comprehended and high standard compound library which can act as a successful approach for drug discovery. Using the framework, more than 2 million of compounds are screened in about 15 min on fully distributed mode (5-node cluster) in contrast with JAVA or Hadoop standalone mode which takes approximately 1 hur and 15 min. Hence, there is an approximately 5 fold decrease in the execution time. Further the performance has potential to increase exponentially with increasing number of nodes.

ACKNOWLEDGEMENT

We are thankful to Department of Bioinformatics, Indian Institute of Information Technology-Allahabad, India for providing the computational facility to perform the study.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

ABBREVIATION USED

USR: Ultrafast Shape Recognition; **DUSR:** Distributed Ultrafast Shape Recognition; **ADME:** Absorption, Distribution, Metabolism, and Elimination; **VS:** Virtual Screening; **HTVS:** High Throughput Virtual Screening; **QSAR**: Quantitative Structure Activity Relationship; **LBVS:** Ligand Based Virtual Screening; **HDFS**: Distributed File System; **DMV:** Dist Map Vector; **DMVLS:** DistMapVectorLibScreen; **DMVL:** Distance Moment Vector Library.

REFERENCES

1. Lavecchia A, Di Giovanni C. Virtual screening strategies in drug discovery: a critical review. Current Medicinal Chemistry. 2013;20(23):2839-60. http://dx.doi.org/10.2174/09298673113209990001.
2. Oprea TI. Virtual screening in lead discovery: a viewpoint. Molecules. 2002;7(1):51-62. http://dx.doi.org/10.3390/70100051.
3. Park H., Lee J, Lee S. Critical assessment of the automated AutoDock as a new docking tool for virtual screening. Proteins: Structure, Function and Genetics. 2006;65(3):549-54. http://dx.doi.org/10.1002/prot.21183 PMid:16988956.
4. Bingjie H, Markus AL. PharmDock: A pharmacophore-based docking program. Journal of Cheminformatics. 2014;6(1):1.
5. Wang R., Ying G, Luhua L. LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design. Journal of Molecular Modeling. 2000;6(7-8)498-516. http://dx.doi.org/10.1007/s0089400060498.
6. Roy A, Skolnick J. LIGSIFT: an open-source tool for ligand structural alignment and virtual screening. Bioinformatics. 2015;31(4):539-44. http://dx.doi.org/10.1093/bioinformatics/btu692 PMid:25336501 PMCid:PMC4325547.
7. Brylinski M, Nonlinear scoring functions for similarity-based ligand docking and binding affinity prediction. Journal of Chemical Information and Modeling. 2013;53(11):3097-112. http://dx.doi.org/10.1021/ci400510e PMid:24171431.
8. Taminau J, Thijs G, De Winter H. Pharao: pharmacophore alignment and optimization. J Mol Graph Model. 2008;27(2)161-9. http://dx.doi.org/10.1016/j.jmgm.2008.04.003 PMid:18485770.
9. Ballester Pedro J, Graham WR. Ultrafast shape recognition to search compound databases for similar molecular shapes. Journal of Computational Chemistry. 2007;28(10):1711-23. http://dx.doi.org/10.1002/jcc.20681 PMid:17342716.
10. Schreyer AM, Tom B. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. Journal of Cheminformatics. 2012;4(1)1-12. http://dx.doi.org/10.1186/1758-2946-4-27 PMid:23131020 PMCid:PMC3505738.
11. Leo S, Federico S, Gianluigi Z. Biodoop: Bioinformatics on Hadoop. Parallel Processing Workshops ICPPW'09, IEEE Computer Society. 2009;415-22.
12. Sastry GM, Adzhigirey M, Day T, Annabhimoju R, Sherman W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. Journal of Computer-aided Molecular Design. 2013;27(3):221-234. http://dx.doi.org/10.1007/s10822-013-9644-8 PMid:23579614.
13. Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. Journal of Medicinal Chemistry. 2000;43(20)3714-7. http://dx.doi.org/10.1021/jm000942e PMid:11020286.
14. White T (2012). Hadoop: The definitive guide, 3^{rd} ed. (O'Reilly Media).
15. Venkatraman V, Padmasini C, Daisuke Kihara. Application of 3D Zernike descriptors to shape-based ligand similarity searching. J Cheminformatics. 2009;1:1. http://dx.doi.org/10.1186/1758-2946-1-19 http://dx.doi.org/10.1186/1758-2946-1-1 PMid:20142984 PMCid:PMC2816862.
16. Vainio MJ, Puranen JS, Johnson MS. ShaEP: Molecular Overlay Based on Shape and Electrostatic Potential. J Chem Inf Model. 2009;49(2):492-502. http://dx.doi.org/10.1021/ci800315d PMid:19434847.
17. Hawkins PD, Skillman AG, Nicholls A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. J Med Chem. 2007;50:74-82. http://dx.doi.org/10.1021/jm0603365 PMid:17201411.
18. Sanner MF, Olson AJ, Sphener JC. Reduced surface: An efficient way to compute molecular surface. Biopolymers. 1996;38(3):305-20. http://dx.doi.org/10.1002/(SICI)1097-0282(199603)38:3<305::AID-BIP4>3.3.CO;2-8 http://dx.doi.org/10.1002/(SICI)1097-0282(199603)38:3<305::AID-BIP4>3.0.CO;2-Y.

19. Bender A, Mussa HY, Gill GS, Glen RC. Molecular Surface Point Environment for Virtual Screening and the Elucidation of Binding Patterns (MOLPRINT3D). J Med Chem. 2004;47(26):6569-83. http://dx.doi.org/10.1021/jm049611i PMid:15588092.

20. Kim DS, Seo J, Kim D, Ryu J, Cho Y, Lee C, Bhak J. BetaDock: Shape-Priority Docking Method Based on Beta-Complex. J Biomol Struct Dyn. 2011;29:219-42. http://dx.doi.org/10.1080/07391102.2011.10507384 PMid:21696235.

21. Tervo AJ, Ronkko T, Nyronen TH, Poso A. BRUTUS: Optimization of a Grid-Based Similarity Function for Rigid-Body Molecular Superposition. 1. Alignment and Virtual Screening Applications. J Med Chem. 2001;48(12):4076-86. http://dx.doi.org/10.1021/jm049123a PMid:15943481.

22. Cross S, Baroni M, Carosati E, Benedetti P, Clementi S. FLAP: GRID Molecular Interaction Fields in Virtual Screening. Validation using the DUD Data Set. J Chem Inf Model. 2010;50(8):1442-50. http://dx.doi.org/10.1021/ci100221g PMid:20690627.

23. Abrahamian E, Fox PC, Naerum L, Christensen IT, Thogersen H, Clark RD. Efficient Generation, Storage, and Manipulation of Fully Flexible Pharmacophore Multiplets and Their Use in 3-D Similarity Searching. J Chem Inf Sci. 2003;43(2):458-68. http://dx.doi.org/10.1021/ci025595r PMid:12653509.

24. Tripathi R, Sharma P, Chakraborty P, Varadwaj PK. Next-generation sequencing revolution through big data analytics. Frontiers in Life Science. 2016;1-31. http://dx.doi.org/10.1080/21553769.2016.1178180

25. Irwin JJ, Brian KS. ZINC - A free database of commercially available compounds for virtual screening. Journal of Chemical Information and Modeling. 2005;45(1):177-82. http://dx.doi.org/10.1021/ci049714+ PMid:15667143 PMCid:PMC1360656.

26. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): An open-source Java library for chemo-and bioinformatics. Journal of Chemical Information and Computer Sciences. 2003;43(2):493-500. http://dx.doi.org/10.1021/ci025584y PMid:12653513 PMCid:PMC4901983.

SUMMARY

- To overcome the discrepancy of the existing USR algorithm, a method called DUSR (Distributed Ultrafast Shape Recognition) was carried out.
- In this method, the chemical compounds were screened on the basis of their drug-likeliness properties prior to the molecular shape comparison followed by shape complementarity momentum measure.
- The DUSR due to its Hadoop implementation acts as a faster approach than the existing standalone tools.
- Utilizing the MapReduce algorithm the high throughput screening of million conformers in a much reduced time span is possible using DUSR.

About Authors

Vandana Kumari: She has done her M. Tech (IT) from Indian Institute of Information Technology-Allahabad (IIIT-Allahabad), India. Currently, she is working as Assistant System Engineer in TCS, New Delhi. Her current research interest is in structural variant analysis using NGS techniques and Big Data Analytics.

Rashmi Tripathi: She received her B. Sc. Degree in Biological Science from Ewing Christian College, Allahabad, India and M. Sc. Degree in Bioinformatics from Banasthali University, Jaipur, India. In 2014, she joined the Department of Bioinformatics, Indian Institute of Information Technology, Allahabad for pursuing Doctorate. Her current research interest is in identifying non-coding RNAs using Next Generation Sequencing techniques utilizing various Bioinformatics tools and techniques.

Sunil Patel: He received his B. Pharma degree from the Gujarat Technological University, India in 2013 and M. Tech (IT) degree in Bioinformatics at Indian Institute of Information Technology, Allahabad, India. Currently, he is working as Assistant System Engineer in TCS, Hyderabad. His current research interest is in biological network characterization using machine learning approach.

Utkarsh Raj: He received his B. Tech Degree (Biotechnology) from Amity University, Lucknow, India and M. Tech Degree in Biotechnology from GBU, Greater Noida, India. In 2014, he joined the Department of Bioinformatics, Indian Institute of Information Technology, Allahabad for pursuing Doctorate. His current research interest is in Drug Designing, Molecular Docking and Simulation utilizing various Bioinformatics tools and techniques.

Pritish Kumar Varadwaj, PhD: Currently working as Associate Professor at Indian Institute of Information Technology- Allahabad (IIIT-Allahabad), India. Dr. Varadwaj has published more than 60 research publications. Currently, he is the Coordinator of the Indo-Russian Centre of Biotechnology at IIIT-Allahabad and doing research on Cancer therapeutics utilizing various Bioinformatics tools and techniques.

Cite this article: Kumari V, Tripathi R, Patel S, Raj U, Varadwaj PK. DUSR (Distributed Ultrafast Shape Recognition): a Hadoop Based Tool to Identify Similar Shaped Ligand Molecules. Indian J of Pharmaceutical Education and Research. 2017;51(1):34-42.