Research Article

IJCRR
Section: General Science
Sci. Journal
Impact Factor 4.016
ICV: 71.54

Prediction of Potential Lead Molecules through Systematic Integration of Multi-omics Datasets - A Mini-Review

Ashok Kumar T.[1], Rajagopal B.[2]

[1]Department of Bioinformatics, Noorul Islam College of Arts and Science, Kumaracoil – 629 180, Tamil Nadu, India; [2]Department of Zoology, Government Arts College, Dharmapuri – 636705, Tamil Nadu, India.

ABSTRACT

Prediction of a novel or potential lead molecules for a therapeutic drug target without adverse effects is a challenging task in the drug designing, discovery, and development process. The systematic integration of multi-omics data from various data/knowledge bases through computational techniques enables to identify potential lead molecules and study the therapeutic properties. Over the last decades, several drug discoveries using multi-omics and huge dataset integration methods proven with successive results. In this paper, we present different types of computational approaches for prediction of potential lead molecules through the systems-level integration of multi-omics datasets.

Key Words: Systematic Integration, Multi-omics Datasets, Drug Discovery, Lead Identification, Big Data Analysis

INTRODUCTION

In drug discovery, lead is a chemical compound that binds to active site regions of the biological target molecule and hence minimizes the binding free energy. Leads may be a natural product, synthetic, or semi-synthetic compound which has therapeutic effects[1].Natural product (or natural drug) consists of bioactive compounds which were produced by the living organisms that are present in nature. Plants, minerals, and animals (including microorganisms) are the common sources of natural products[2,3,69,74,75]. Natural products can also be developed by chemical synthesis (both semi-synthesis and total synthesis) and have been placed a major role in the development of potential synthetic targets. But synthetic and semi-synthetic compounds are chemically synthesized by the humans in the laboratory using *in silico* and/or experimental approaches[4,5].

Developing a potential lead molecule by using the experimental method is tedious, complicated, expensive, time-consuming, and trial-and-error process[6]. Recently, many advanced computational techniques analogous to wet-lab techniques were introduced to reduce the problem. Modern computer-aided drug design and discovery (CADDD) involve virtual screening, testing, and validation of lead molecules in a short time span using large datasets and software [73]. The resulting lead molecule further undergoes a series of preclinical and clinical studies to test the toxicity and adverse effects. The successful drug candidate is released in different dosage forms in the market after passing the food and drug administration (FDA) verification process[7,8].

MULTI-OMICS AND BIG DATA INTEGRATION

Multi-omics is a new approach for analyzing biological problems in various aspects through combining multiple-*omics* datasets[3,9]. The common types of omics include genomics, proteomics, metabolomics, epigenomics, phytochemomics, interactomics, and microbiomics[10-12]. Integration of multiple omics data in a systematic way enables to study the functional relationship or identify the key problem in an efficient manner. An association of large datasets or complex datasets of multi-omics data is a difficult task and must have sound knowledge in all areas of omics.The pattern matching (or regular expression) is a general and most popular technique for extraction of knowledge from the datasets. Analyzing the large multi-omics datasets involves big data handling.

Corresponding Author:
Ashok Kumar T., Assistant Professor and Head, Department of Bioinformatics, Noorul Islam College of Arts and Science, Kumaracoil, Thuckalay – 629 180, Tamil Nadu, India; Mobile: +919655307178; E-mail: ashok@biogem.org

ISSN: 2231-2196 (Print) **ISSN:** 0975-5241 (Online) **DOI:** 10.7324/IJCRR.2017.9194

Received: 23.08.2017 **Revised:** 09.09.2017 **Accepted:** 24.09.2017

Due to rapid growth in data size, diversity, and complexity of datasets in the biological databases, big data were introduced to analyze, manage, and derive knowledge from the datasets. Big data (aka huge data or massive data) refer to a very large volume of data or data storage, which cannot be processed using traditional computing devices and applications. Size of big data ranges from petabytes(1 PB = 10^{15} bytes) to exabytes (1 EB = 10^{18} bytes), or even more[13-15].Even though the big data analysis is a hot topic today, the concept has evolved over many years ago in IT and R&D sector. Next-generation sequencing (NGS) and drug discovery are the two most popular areas of biological sciences which currently implement big data analysis in knowledge discovery[16-18].

Comprehensive Data Integration Methods

Integrating comprehensive and related datasets from various biological databases or other external sources increases the accuracy in lead prediction, and also reveals hidden functions and interrelationship within the molecules[19].There are three types of approaches adopted to combine comprehensive data and reduce data size (Table 1): (*i*) semantic web approach – searching, retrieving, or annotating data from other external data sources through metadata or a RESTful APIweb services [20,21]; (*ii*) data warehousing approach – extracting data from other external sources and combining into a global dataset[19,22]; and (*iii*) data mining approach – extracting data or knowledge from different types of large datasets through suitable patterns[23,24].

Most of the popular three-dimensional (3D) molecular structure databases such as RCSB Protein Data Bank[25], NCBI PubChem[26], EMBL-EBI ChEBI [27], Drug Bank[28], etc. have implemented REST ful API web services or SOAP to share or integrate data in the form of FTP, HTML, XML, JSON, plain text, or AWK commands[29].Moreover, cloud computing services were offered to handle, analyze, or interpret big datasets through various remote applications/servers. There are many cloud servers such as Cloud BLAST[30], Myrna[31], Cloud Burst[32], Hadoop-BAM[33], GPU-BLAST[34], Hydra[35], Peak Ranger[36],Crossbow[37], etc. were available over cloud for analyzing different types of big datasets [38-41].

Unsupervised Data Analysis and Analytics

Handling big dataset or multi-omics data is a difficult task, because it is often very comprehensive and available in real time. In Bioinformatics, sequence (alphabets) and structure (XYZ coordinates) are the major data used for big data analysis and analytics. An effective lead identification and functional interrelationship prediction require integration of very large datasets of3D chemical libraries and disease-target-ligand interaction network. Usually unsupervised multi-omics/big datasets are integrated using clustering and grouping technique. The different types of dataset integrations are target-ligand interactions, intermolecular interactions, disease-target interactions, disease-disease relationships, protein-protein interactions, target-disease-metabolic pathways, drug-side effect relationships, gene interactions, structure-function relationships, etc. [42-44].

The network model graphical representation of biological data interrelations and various types of unsupervised dataset integration methods are [44,56]:(*i*) network-based methods – graphical representation of interrelations using the network (distance) datasets [45,46],(*ii*) Bayesian methods – probabilistic graphical representation of interrelations using the probability distribution datasets [47-51], (*iii*) correlation-based methods – multivariate graphical representation of interrelations using the partial least squares datasets [52,53], (*iv*) matrix factorization methods – graphical representation of interrelations using the product and rank of the two matrix datasets [54], and (*v*) kernel-based methods – graphical representation of interrelations using the pattern datasets predicted from kernel matrix [55].

Big Data Accessing Methods

Accessing large datasets requires high-performance computing (HPC) infrastructure and a suitable big data framework [14,15]. The common methods for big data handling are cloud computing, graphics processing unit (GPU) computing, Xeon Phi computing, grid computing, and cluster computing [57,58]. Large datasets can be accessed from various data sources using big data framework, which is based on client-server technology [59]. There are many types of big data processing frameworks used for accessing datasets through a pipeline, among which popularly used frameworks and programs are: Apache Hadoop [76], Apache Spark [77], Apache Flink [78], Apache Storm [79], Apache Samza [80], Apache Cassandra [81], NoSQL [82], R [83], and Python[84].

SYSTEMATIC MULTI-OMICS DATA INTEGRATION

A successful drug discovery requires exact compound or most suitable compound which can fit all pocketsin the active site of the target molecule and brought to a stable state [7,8]. The systematic integration of theoretical and experimental datasets of multi-omics, target-ligand interaction network, physicochemical properties, and functional properties leads to design a safe and efficient therapeutics [60].

Integrative Systems Biology Approach

To design an effective drug molecule, it is most essential to understand the nature and causes of the disease [61].Integrative systems biology advances thorough study of biological phenomenon of a system (organism, e.g. human) in a systematic way (Figure 1).The complex interaction networks

in a system can be combined through either top-down or bottom-up approaches using multi-omics datasets [62,63]. Currently there are many bioinformatics databases and tools were available for collection of various omics data and hence can design a new virtual system.

Computational Methods for Lead Identification

A lead molecule can be identified by integrating or comparing target data with large datasets using computational and statistical approaches. The common computational lead identification techniques using large datasets include:

i. *Multiple sequence alignment* –It is a popular method to find local similarity, homology, and phylogenetic relationship between different genes or protein sequences [41]. The sequence similarity through structure-based sequence alignment enables to find the similar target-ligand interacting molecules. Structural superposition is another alternative approach to compare similar protein structures based on the root mean square deviation (RMSD) calculation [64]. Moreover, systematic integration of large datasets of target-ligand molecular interaction network data with multi-omics data enables to predict or design a potential lead molecule [60].

ii. *Maximum common substructure* – It is a widely used method in CADD for finding similar 3D structures through structured-based or ligand-based virtual screening [60].Maximum common substructure search using SMILES (Simplified Molecular Input Line Entry System)pattern is commonly used to find structural similarity between large chemical datasets [65].The substructure search with compounds in the phenotype linked target-ligand interacting network datasets integrated with multi-omics data enables to predict or design a novel and potential lead molecule [66-68].

iii. *Molecular interaction network* – It is the modern and most successive approach to find a novel drug by systematic integration of large datasets of multi-omics data [60].Data scientists integrates big data into complex network in the order of phenotype → target →target-ligand←ligand←chemical library to predict or design a novel and potential lead molecule (Figure 2). Recently, many big pharmaceutical companies and R&D organizations have renewed their interest in discovering potential lead compounds from the natural products due to the structural diversity and medicinal properties [3,69,70].

CONCLUSION

Biological systems are analogous to the computer system in disease/target identification and drug design. To troubleshoot hardware issues in the computer, we must have the complete circuit diagram and the component to fix the problem [71]. In contrast, through increasing the volume of multi-omics datasets and systematic integration of large datasets, it is possible to design an effective drug molecule [72]. Recent research advances in cloud computing, big data analysis, multi-omics data integration, and virtual screening and testing technology have reduced the cost and time in predicting potential lead molecules.

ACKNOWLEDGEMENT

Authors acknowledge the immense help received from the scholars whose articles are cited and included in references of this manuscript. The authors are also grateful to authors / editors / publishers of all those articles, journals and books from where the literature for this article has been reviewed and discussed.

Conflict of Interest

The authors declare that there is no conflict of interest regarding publication of the paper.

REFERENCES

1. S.Z. Tasker, P.J. Hergenrother, Natural products: Taming reactive benzynes, *Nat. Chem.* 9 (2017) 504–506.
2. M. Lahlou, Screening of natural products for drug discovery, *Expert Opin. Drug Discov.* 2 (2007) 697–705.
3. T. Ashok Kumar, B. Rajagopal, PDTDB – An Integrative Structural Database and Prediction Server for Plant Metabolites and Therapeutic Drug Targets, *Int. J. Curr. Res.* 9(2017) 46537–46541.
4. All natural, *Nat. Chem. Biol*. 3 (2007) 351–351.
5. A.M. Lourenço, L.M. Ferreira, P.S. Branco, Molecules of natural origin, semi-synthesis and synthesis with anti-inflammatory and anticancer utilities, *Curr. Pharm. Des.* 18 (2012) 3979–4046.
6. F. Ooms, Molecular modeling and computer aided drug design. Examples of their applications in medicinal chemistry, *Curr. Med. Chem.* 7 (2000) 141–158.
7. I.M. Kapetanovic, Computer-aided Drug Discovery and Development (CADDD): *in silico*-chemico-biological approach, *Chem. Biol. Interact*. 171 (2008) 165–176.
8. G. Sliwoski, S. Kothiwale, J. Meiler, E.W. Lowe, Computational Methods in Drug Discovery, *Pharmacol. Rev.* 66 (2013) 334–395.
9. A. Ebrahim, E. Brunk, J. Tan, E.J. O'Brien, D. Kim, R. Szubin, J.A. Lerman, A. Lechner, A. Sastry, A. Bordbar, A.M. Feist, B.O. Palsson, Multi-omic data integration enables discovery of hidden biological regularities, *Nat. Commun*. 7 (2016) 13091.
10. M. Bersanelli, E. Mosca, D. Remondini, E. Giampieri, C. Sala, G. Castellani, L. Milanesi, Methods for the integration of multi-omics data: mathematical aspects, *BMC Bioinformatics.* 17 (2016) 167–202.
11. C. Bock, M. Farlik, N.C. Sheffield, Multi-Omics of Single Cells: Strategies and Applications, *Trends in Biotechnol.* 34 (2016) 605–608.
12. C. Vilanova, M. Porcar, Are multi-omics enough?, *Nat. Microbiol.* 1 (2016) 16101.
13. M. Swan, The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery, *Big Data*. 1 (2013)

85–99.
14. H. Mohanty, P. Bhuyan, D. Chenthati, eds., *Big Data*, Springer India, New Delhi, 2015.
15. V. Marx, Biology: The big challenges of big data, *Nature*. 498 (2013) 255–260.
16. L. Leyens, M. Reumann, N. Malats, A. Brand, Use of big data for drug development and for public and personal health and care: Leyens et al., *Genet. Epidemiol.* 41 (2017) 51–60.
17. R.S. Kim, N. Goossens, Y. Hoshida, Use of big data in drug development for precision medicine, *Expert Rev. Precis. Med. Drug Dev*. 1 (2016) 245–253.
18. R. Tripathi, P. Sharma, P. Chakraborty, P.K. Varadwaj, Next-generation sequencing revolution through big data analytics, Front. *Life Sci.* 9 (2016) 119–149.
19. C. Chen, P.B. McGarvey, H. Huang, C.H. Wu, Protein Bioinformatics Infrastructure for the Integration and Analysis of Multiple High-Throughput "omics" Data, Adv. *Bioinformatics*. 2010 (2010) 1–19.
20. K.-H. Cheung, A.K. Smith, K.Y.L. Yip, C.J.O. Baker, M.B. Gerstein, Semantic Web Approach to Database Integration in the Life Sciences, in: C.J.O. Baker, K.-H. Cheung (Eds.), *Semantic Web*, Springer US, Boston, MA, 2007: pp. 11–30.
21. L.J.G. Post, M. Roos, M.S. Marshall, R. van Driel, T.M. Breit, A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data, *Bioinformatics*. 23 (2007) 3080–3087.
22. H. Chai, G. Wu, Y. Zhao, A Document-Based Data Warehousing Approach for Large Scale Data Mining, in: Q. Zu, B. Hu, A. Elçi (Eds.), Pervasive Computing and the Networked World, Springer Berlin Heidelberg, *Berlin, Heidelberg,* 2013: pp. 69–81.
23. J. Han, M. Kamber, D*ata mining: concepts and techniques*, Morgan Kaufmann Publishers, San Francisco, 2001.
24. M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*, 2nd ed, John Wiley : IEEE Press, Hoboken, N.J, 2011.
25. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, *Nucleic Acids Res*. 28 (2000) 235–242.
26. S. Kim, P.A. Thiessen, E.E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B.A. Shoemaker, J. Wang, B. Yu, J. Zhang, S.H. Bryant, PubChem Substance and Compound databases, *Nucleic Acids Res*. 44 (2016) D1202–D1213.
27. K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner, ChEBI: a database and ontology for chemical entities of biological interest, Nucleic Acids Res. 36 (2008) D344–D350.
28. V. Law, C. Knox, Y. Djoumbou, T. Jewison, A.C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z.T. Dame, B. Han, Y. Zhou, D.S. Wishart, DrugBank 4.0: shedding new light on drug metabolism, *Nucleic Acids Res*. 42 (2014) D1091-1097.
29. T. Ashok Kumar, B. Rajagopal,BLASTphp: a PHP wrapper for NCBI BLAST API, *Int. J. Comp. Bio*. 6(2017) 31–33.
30. A. Matsunaga, M. Tsugawa, J. Fortes, Cloud BLAST: Combining Map Reduce and Virtualization on Distributed Resources for Bioinformatics Applications, in: *IEEE*, 2008: pp. 222–229.
31. B. Langmead, K.D. Hansen, J.T. Leek, Cloud-scale RNA-sequencing differential expression analysis with Myrna, *Genome Biol.* 11 (2010) R83.
32. M.C. Schatz, CloudBurst: highly sensitive read mapping with MapReduce, *Bioinformatics*. 25 (2009) 1363–1369.
33. M. Niemenmaa, A. Kallio, A. Schumacher, P. Klemelä, E. Korpelainen, K. Heljanko, Hadoop-BAM: directly manipulating next generation sequencing data in the cloud, *Bioinformatics*. 28 (2012) 876–877.
34. P.D. Vouzis, N.V. Sahinidis, GPU-BLAST: using graphics processors to accelerate protein sequence alignment, *Bioinformatics*. 27 (2011) 182–188.
35. S. Lewis, A. Csordas, S. Killcoyne, H. Hermjakob, M.R. Hoopmann, R.L. Moritz, E.W. Deutsch, J. Boyle, Hydra: a scalable proteomic search engine which utilizes the Hadoop distributed computing framework, *BMC Bioinformatics*. 13 (2012) 324.
36. X. Feng, R. Grossman, L. Stein, PeakRanger: A cloud-enabled peak caller for ChIP-seq data, *BMC Bioinformatics*. 12 (2011) 139.
37. B. Langmead, M.C. Schatz, J. Lin, M. Pop, S.L. Salzberg, Searching for SNPs with cloud computing, *Genome Biol.* 10 (2009) R134.
38. C. Yang, Q. Huang, Z. Li, K. Liu, F. Hu, Big Data and cloud computing: innovation opportunities and challenges, *Int. J. Digit. Earth*. 10 (2017) 13–53.
39. B.T. Moghadam, J. Alvarsson, M. Holm, M. Eklund, L. Carlsson, O. Spjuth, Scaling Predictive Modeling in Drug Development with Cloud Computing, *J. Chem. Inf. Model.* 55 (2015) 19–25.
40. D. D'Agostino, A. Clematis, A. Quarati, D. Cesini, F. Chiappori, L. Milanesi, I. Merelli, Cloud Infrastructures for In Silico Drug Discovery: Economic and Practical Aspects, *BioMed Res. Int*. 2013 (2013) 1–19.
41. J. Daugelaite, A. O' Driscoll, R.D. Sleator, An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics, *ISRN Biomath*. 2013 (2013) 1–14.
42. M.W. Gonzalez, M.G. Kann, Chapter 4: Protein Interactions and Disease, *PLoS Comput. Biol.* 8 (2012) e1002819.
43. K. Sun, N. Buchan, C. Larminie, N. Pržulj, The integrated disease network, Integr. Biol. 6 (2014) 1069–1079.
44. V. Gligorijević, N. Pržulj, Methods for biological data integration: perspectives and challenges, J. R. Soc. Interface. 12 (2015) 20150571.
45. F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, Y. Tang, Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference, *PLoS Comput. Biol.* 8 (2012) e1002503.
46. X. Guo, L. Gao, C. Wei, X. Yang, Y. Zhao, A. Dong, A Computational Method Based on the Integration of Heterogeneous Networks for Predicting Disease-Gene Associations, *PLoS ONE*. 6 (2011) e24171.
47. C.J. Needham, J.R. Bradford, A.J. Bulpitt, D.R. Westhead, A Primer on Learning in Bayesian Networks for Computational Biology, *PLoS Comput. Biol.* 3 (2007) e129.
48. I. Ben-Gal, Bayesian Networks, in: F. Ruggeri, R.S. Kenett, F.W. Faltin (Eds.), Encyclopedia of Statistics in Quality and Reliability, John Wiley & Sons, Ltd, Chichester, UK, 2008.
49. E.E. Schadt, S.H. Friend, D.A. Shaywitz, A network view of disease and compound screening, Nat. Rev. Drug Discov. 8 (2009) 286–295.
50. O. Gevaert, F.D. Smet, D. Timmerman, Y. Moreau, B.D. Moor, Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks, Bioinformatics. 22 (2006) e184–e190.
51. R. Jansen, A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data, Science. 302 (2003) 449–453.
52. E. Parkhomenko, D. Tritchler, J. Beyene, Sparse Canonical Correlation Analysis with Application to Genomic Data Integration, Stat. *Appl. Genet. Mol. Biol.* 8 (2009) 1–34.
53. J. Chen, S. Zhang, Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data, Bioinformatics. 32 (2016) 1724–1732.

54. D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature*. 401 (1999) 788–791.
55. B. Schölkopf, K. Tsuda, J.-P. Vert, eds., *Kernel methods in computational biology,* MIT Press, Cambridge, Mass, 2004.
56. S. Huang, K. Chaudhary, L.X. Garmire, More Is Better: Recent Progress in Multi-Omics Data Integration Methods, Front. Genet. 8 (2017).
57. D.E. Baz, IoT and the Need for High Performance Computing, in: *IEEE*, 2014: pp. 1–6.
58. H. Peréz-Sánchez, A. Fassihi, J.M. Cecilia, H.H. Ali, M. Cannataro, Applications of High Performance Computing in Bioinformatics, Computational Biology and Computational Chemistry, in: F. Ortuño, I. Rojas (Eds.), Bioinformatics and Biomedical Engineering, Springer International Publishing, *Cham*, 2015: pp. 527–541.
59. A. Bhadani, D. Jothimani, Big Data: Challenges, Opportunities, and Realities, in: Manoj Kumar Singh, G. Dileep Kumar (Eds.), *Effective Big Data Management and Opportunities for Implementation,* IGI Global, Pennsylvania, USA, 2016: pp. 1–24.
60. H. Yu, J. Chen, X. Xu, Y. Li, H. Zhao, Y. Fang, X. Li, W. Zhou, W. Wang, Y. Wang, A Systematic Prediction of Multiple Drug-Target Interactions from Chemical, Genomic, and Pharmacological Data, *PLoS ONE*. 7 (2012) e37608.
61. B. Chen, A. Butte, Leveraging big data to transform target selection and drug discovery, *Clin. Pharmacol. Ther.* 99 (2016) 285–297.
62. F.J. Bruggeman, H.V. Westerhoff, The nature of systems biology, *Trends Microbiol.* 15 (2007) 45–50.
63. H.-C. Schneider, T. Klabunde, Understanding drugs and diseases by systems biology?, *Bioorg. Med. Chem. Lett.* 23 (2013) 1168–1176.
64. A.D. McLachlan, Rapid comparison of protein structures, *Acta Cryst. A*. 38 (1982) 871–873.
65. Y. Cao, T. Jiang, T. Girke, A maximum common substructure-based algorithm for searching and predicting drug-like compounds, *Bioinformatics*. 24 (2008) i366–i374.
66. G.R. Zimmermann, J. Lehár, C.T. Keith, Multi-target therapeutics: when the whole is greater than the sum of the parts, *Drug Discov. Today*. 12 (2007) 34–42.
67. K.A. O'Connor, B.L. Roth, Finding new tricks for old drugs: an efficient route for public-sector drug discovery, Nat. Rev. *Drug Discov.* 4 (2005) 1005–1014.
68. T.T. Ashburn, K.B. Thor, Drug repositioning: identifying and developing new uses for existing drugs, *Nat. Rev. Drug Discov.* 3 (2004) 673–683.
69. M. Lahlou, The Success of Natural Products in Drug Discovery, *Pharmacol. Pharm.* 04 (2013) 17–31.
70. V. Aparna, K. Dineshkumar, N. Mohanalakshmi, D. Velmurugan, W. Hopper, Identification of Natural Compound Inhibitors for Multidrug Efflux Pumps of *Escherichia coli* and *Pseudomonas aeruginosa* Using *In Silico* High-Throughput Virtual Screening and *In Vitro* Validation, *PLoS ONE*. 9 (2014) e101840.
71. Y. Lazebnik, Can a biologist fix a radio?—Or, what I learned while studying apoptosis, *Cancer Cell.* 2 (2002) 179–182.
72. T. Katsila, G.A. Spyroulias, G.P. Patrinos, M.-T. Matsoukas, Computational approaches in target identification and drug discovery, *Comput. Struct. Biotechnol. J.* 14 (2016) 177–184.
73. S.J.Y. Macalino, V. Gosu, S. Hong, S. Choi, Role of computer-aided drug design in modern drug discovery, *Arch. Pharm. Res.* 38 (2015) 1686–1701.
74. B. Oyon (2016, August 9), What Are the Pharmaceutical Sources of Drugs? *HealDove*. Retrieved August 21, 2017, from https://healdove.com/health-care-industry/Where-do-drugs-come-from-Sources-of-Drugs
75. Natural product (n.d.), *Wikipedia*. Retrieved August 21, 2017, from https://en.wikipedia.org/wiki/Natural_product
76. Apache Hadoop – https://hadoop.apache.org/
77. Apache Spark – https://spark.apache.org/
78. Apache Flink – https://flink.apache.org/
79. Apache Storm – https://storm.apache.org/
80. Apache Samza – http://samza.apache.org/
81. Apache Cassandra – https://cassandra.apache.org/
82. NoSQL – https://www.oracle.com/database/nosql/index.html
83. R – https://www.r-project.org/
84. Python – https://www.python.org/

Approaches	Advantages	Disadvantages
Semantic web	• Occupies less storage space • Provides more information • Provides updated information • High quality of data • Multiple access options	• Non-uniform data from external sources • Sometimes links may be broken • The data access format may be changed • Sometimes data may be ambiguous • Interlinking is not possible • Sometimes data process may timeout • Interrelationship study is not possible
Data warehousing	• Provides more information • High quality of data • Uniform access options • Interlinked to the target source • Can predict interrelationships • Can add more features	• Occupies more storage space • Provides outdated information • Manual data synchronization
Data mining	• Provides updated information • Uniform access options • Interlinked to the target source • Can predict interrelationships • Can add more features	• Occupies more storage space • Provides less information • Less quality of data

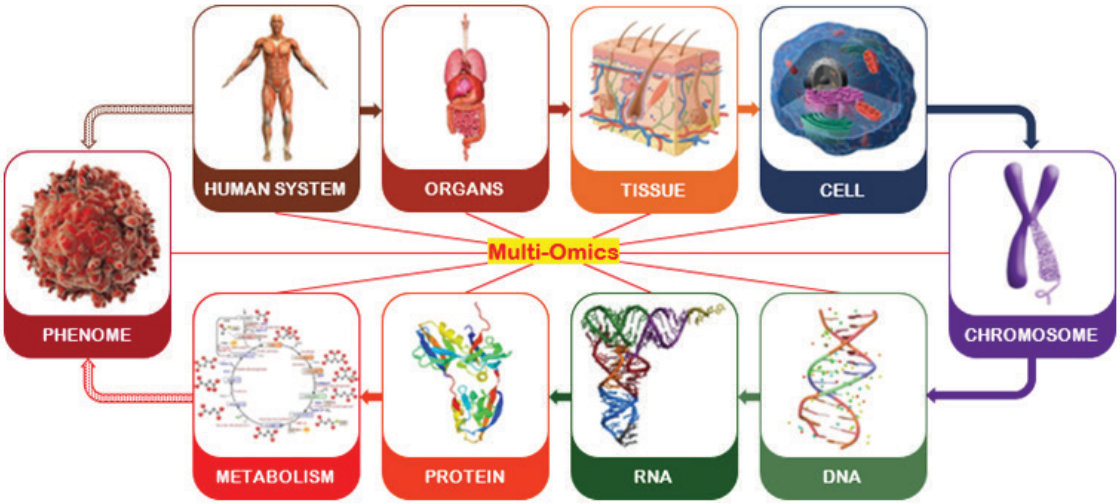

Figure 1: An illustration of multi-omics data integration through integrative systems biology approach.

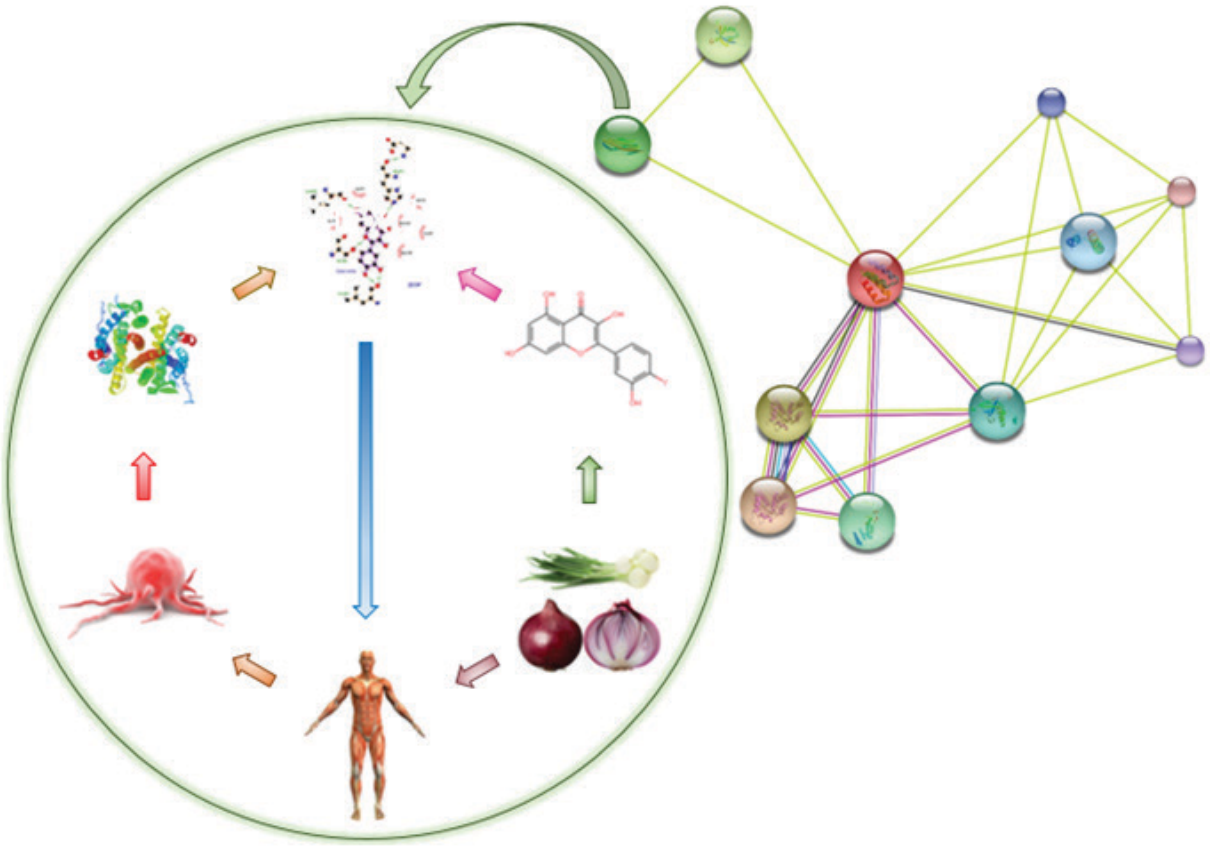

Figure 2: An illustration on multiple target and phytochemical interaction network.