Virtual Natural Product Library – full text searchable database.

Subhash Chandra Bose. Kotte[1]*, Pavan Kumar K.V.T.S[1]*, Ravi Kumar Tumuluri[3]*, Shriram Raghavan[1], P.K. Dubey[2], and P.M. Murali[1],

[1] Evolva Biotech Private Limited, TICEL Bio Park Limited, Taramani, Chennai - 600113, India.

[2] Department of Chemistry, Jawaharlal Nehru Technological University Hyderabad, Andhra Pradesh, 500085, India.

[3] Acharya Nagarjuna University- Dr. M.R. Appa Row Campus, NUZVID, Andhra Pradesh, 521201, India.

Abstract

Small molecules occurring in nature have special significance to mankind. They have varied applications from healthcare, food, nutrition, agriculture, personal care and well-being. These natural small molecules are from very diverse sources from the rarest plants to deep sea creatures. Recently they have assumed a lot of significance as pharmaceutical companies is constantly pushing the horizons to make them "druggable" due to their inherent bioactivities. Though they are not easy to synthesize or isolate, yet their diverse molecular scaffold confers them significance especially given the fact of prevailing resistance to drug scaffolds presently being used in the clinics. Hence it's of paramount importance to have a database of diverse natural small molecules through the present effort of creating a "Virtual Natural Products Library" (VNPL-version 0.15).

Keywords: *Database; Druggable; Natural Compounds database; SDF; Virtual Natural Product Library; VNPL.*

Availability: VNPL database is available at http://122.169.243.137/ChemStructure/Home.aspx

1. Introduction

VNPL (Virtual Natural Product Library) was created by manually curating several journals that are published in the sphere of natural products and painstakingly identifying articles with structures, followed by drawing them and storing them in a "Structure-Data" (SD) format [1]. These articles were identified over the last decade starting from 1999 till 2009 selected from dozens of key primary journals and its coverage includes isolation studies, biosynthesis, and new natural products, known compounds from new sources, structure determinations, new properties and biological activities. Around 200 graphical abstracts are contained in each monthly bulletin including structure diagrams, trivial and taxonomic names, molecular formulae and physical and biological properties. This is an ideal source to mine for natural compounds with or without specific biological activity from various natural sources of interest.

VNPL is collection of service for chemists, biochemists, pharmacists, medicinal chemists and others working with natural products. Over 80 journals are monitored, covering topics such as isolation studies, new compounds and known compounds form new sources, structure determinations, new properties and activities, and total synthesis and biosynthesis. Structure was curated from each of the peer-reviewed article and this was then allocated a unique identifier code.

VNPL differentiates itself in comparison to other public domain databases (such as Pubchem, ChemSpider, Wombat, Zinc) with regard to its depth of all the natural compound entries contained in it, which would be revealed by a simple search. Each entry is selected based on the information whether its extracted from its native source or if an attempt has been made for total synthesis of the same. Citation to any biological activity or any reference to establish bio-efficacy is another parameter in identifying the compounds. These two parameters give the edge to each compound entry with specific reference to its accessibility and activity.

2. Database description

A web interface was then constructed to give an easy access to this data. In the VNPL version 0.15, query can be based on the molecular descriptors matching the criteria including formula, molecular weight, LogD, hydrogen bond donors, hydrogen bond acceptors, number of rotatable bonds, total polar surface area. The output contains the list of hits matching one or more query parameters wherein each hit is distinguished with a unique identifier, followed by its 2D chemical structure, one dimensional SMILES notation, molecular formula, molecular weight, logD, AlogP, molecular solubility, total

polar surface area, number of aromatic bonds, number of aromatic rings, number of hydrogen bond donors, acceptors & rotatable bonds. While the structure is manually curated from the peer reviewed publication, all other molecular descriptors are computed dynamically at the back end. In addition, the statistics tab indicates the diversity based on molecular weight, several different structural clusters to which these compounds belong to, the range of hydrogen bond acceptors & donors and LogD.

While there is no specific preference of a chemotype, the fundamental criteria is to ensure its "natural" or "nature-mimicing" origin. The back end of the database uses simple "Structure-data" format files with descriptors where applicable. The search interface is built as a layer on top of this.

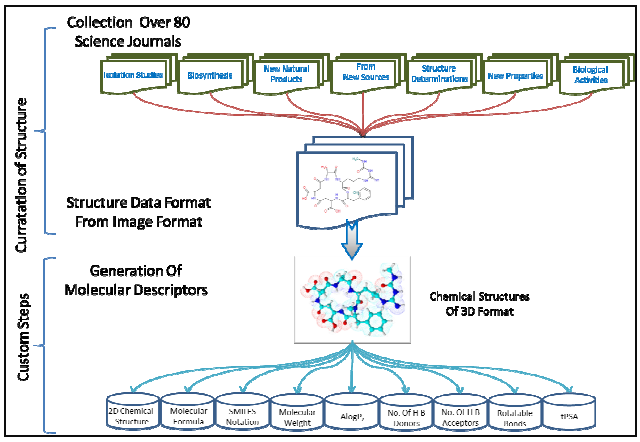

Figure 1: The scheme Diagram for the preparation workflow for database generation.

3. Database search options

A researcher may be interested in retrieving information, such as 2D or 3D structures. However, without prior knowledge of the chemical content or structure of this database, the researcher would have been unable to find these records. VNPL now allows searching for compounds by molecular weight and other text annotations. These annotations, numbering over ~18K using (in-house built Structural Library of Natural Compounds database [2]), comprise a substantial corpus of information.

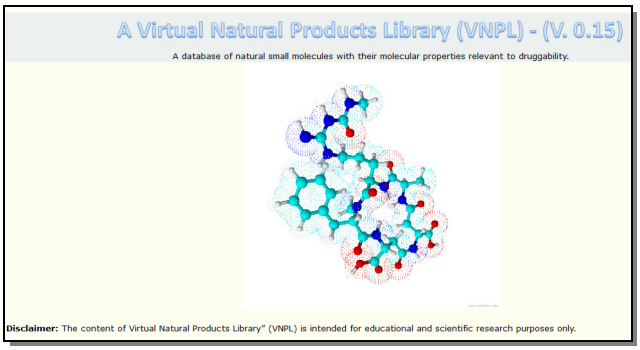

Figure 2: Main window for Virtual Natural Product Library database.

To allow users to quickly search textual information, VNPL has been updated with an annotation parser and indexer which accommodate the special syntax often found in the chemistry domain, such as Molecular Formula, SMILES, Molecular weight, H bond Donor & H bond Acceptor, Logp, Rotatable TPSA and SMILES strings.

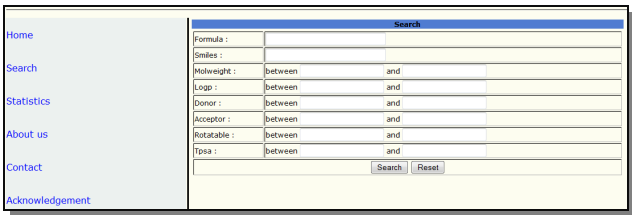

Figure 3: Search engine query Options.

These modules provide full-text indexing and sub-second searching capabilities over the database annotations, comparable to those of a typical web-based document search engine. With this tool, users can retrieve full chemical structure records given. This is an especially convenient method for finding chemicals as they are generally identified by common, non-systematic, names that can only be indexed through an electronic knowledge repository, such as the corpus of database annotations.

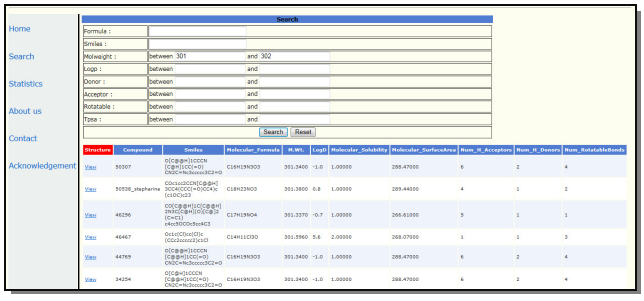

Figure 4: diversity against the whole database for each compound by Molecular Weight

Moreover, in some cases, Systematic names in VNPL can nevertheless be useful when searching for multiple substring keywords. For instance, a keyword query such as '301' as molecular weight retrieves all chemicals whose molecular weight contains all of those keywords, effectively acting as a multiple functional group search.

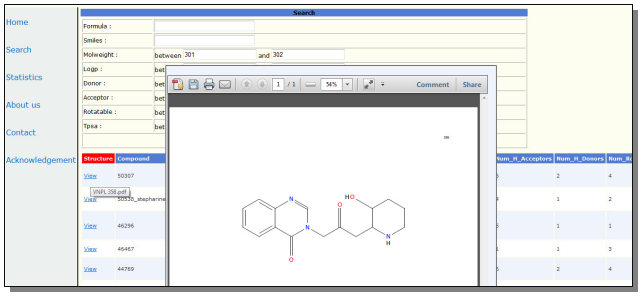

Figure 5: 2D structure view in search engine.

4. Statistics

M.Wt. Range	Number of Molecule
0 to 499	12593
500 to 999	4965
1000 to 1999	863
2000 and above	33

Table 1: Moleculare Weight range in the curration.

Lipinski [4] rules are the most widely used to identify drug-like compounds [5]. Other techniques based on artificial neural networks have also been used [6], [7], [8]. Molecular weight ≤ 3347.17, ALogP ≤ 44.301, TPSA ≤ 3045.82, H Acceptors ≤ 93, H Donors ≤ 56, Rotatable bonds ≤ 93, The definition of the reactive functions used is the modified version by Oprea [9] of the list published by Rishton [10].

A recent review of Hann and Oprea gives rules to select lead-like molecules [11]. In our database the data set was "clean-fragments" 18, 454 entries from in-house build database. All the entries in this dataset ALogP ≤ 44.301, molecular weight (MW) ≤ 3347.17, TPSA ≤ 3045.82, H bond acceptor (HBA) ≤ 93, H bond Donors (HBD) ≤ 56 and rotatable bonds ≤ 93.

4.1 Rotatable bonds: The definition of JOELib [3] is used Number of rotatable bonds, where the atoms are heavy atoms with bond orders one and a hybridization which is not one (no sp). Additionally the bond is a non-ring-bond.

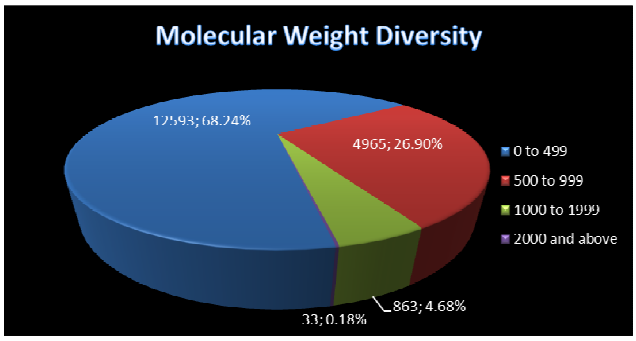

Figure 6: diversity against the whole database for each compound by Molecular Weight

The relationship between the number of compounds in a database and the diversity (number of clusters) of the chemical space covered by a database is essential information. We used the dissimilarity step of the Accelrys fingerprints to compute the number of clusters for the whole database and for each provider (figure 7). The VNPL database is clearly the most representative of the chemical space and covers 68.24 % of the chemical space of the whole database obeys Lipinski's rule of five.

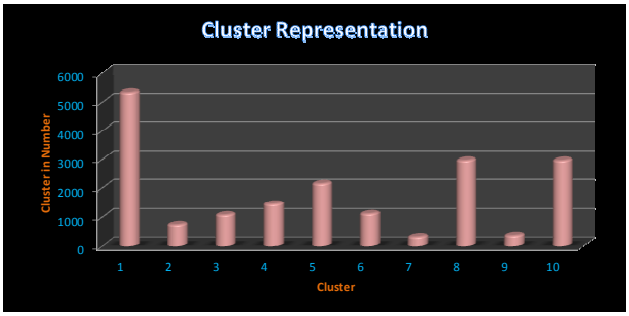

Figure 7: diversity against the whole database for each compound.

Hydrogen acceptors count from 0 to 10 are 14467 molecules, 11 to 20 are 3094 and >20 are 893 molecules.

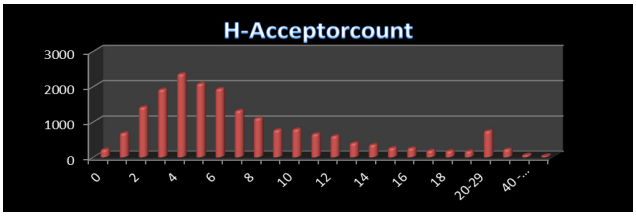

Figure 8: H-Bond Acceptors score.

Hydrogen Donors count from 0 to 10 are 17512 molecules, 11 to 19 are 789 and >20 are 152 molecules.

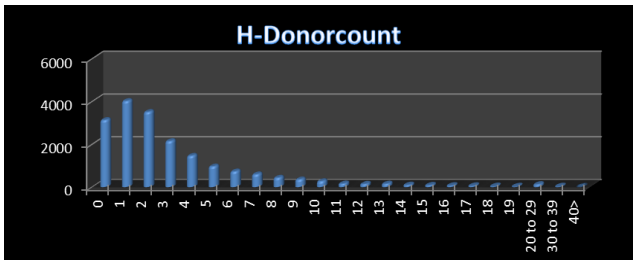

Figure 9: H-Bond Donor score

LogD values are from -11 to 0 are 3909, 0 to 5 are 11993 molecules, 6 to 10 are 2208 and >11 are 350 molecules.

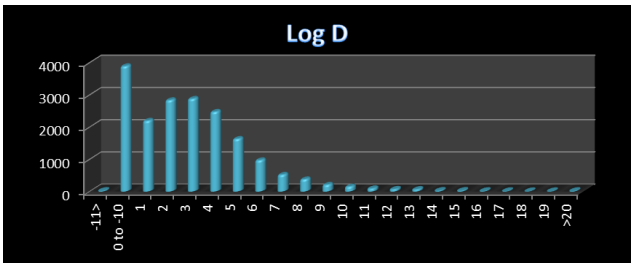

Figure 10: Log D score

TPSA Rotatable bonds from 0 to 5 are 10365 molecules, 6 to 10 are 4542 and >11 are 3567 molecules.

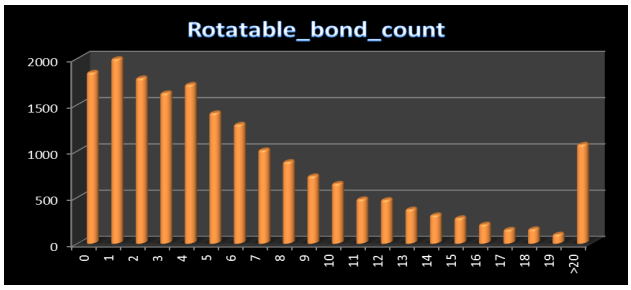

Figure 11: Number of rotatable bonds

5. Conclusion

VNPL is an attempt to create a public domain repository in this field of natural products to assist researchers working towards drug discovery including several "open source" efforts. VNPL will be shortly enriched with capabilities to search using structure & sub-structure to precisely identify the molecules having relevant pharmacophore features and molecular scaffold and the contacts tab gives the details of authors for those who seek further details and in the next update we are planning to provide journal references. Authors will make every attempt to maintain and make this database active for as long as possible and we don't have any specific timeframe. The idea is to make this available to public and help researchers such as those involved in world's first Open Source Drug Discovery (OSDD) effort.

6. Acknowledgments

We express our special thanks to Mr. Madhav K Rao and Mr. S. Narasimha Reddy of Lexicon Infotech Limited at Hyderabad for their excellent support for creating a web interface and hosting the data. VNPL databases curation is funded from our own internal resources.

7. Disclaimers

VNPL databases are provided solely for research, informational and educational purposes only.

8. Journal Source

Accounts of Chemical Research Acta Crystallographica, Section C, Acta Pharmaceutica, Angewandte Chemie, International Edition, Archiv der Pharmazie (Weinheim, Germany), Australian Journal of Chemistry Biological and Pharmaceutical Bulletin, Bioorganic and Medicinal Chemistry Letters Bioscience, Biotechnology and Biochemistry, Bulletin of the Chemical Society of Japan, Canadian Journal of Chemistry, Chemical and Pharmaceutical Bulletin, Chemical Communications, Chemical Reviews, Chemical Society Reviews, Chemistry Letters, Chemistry of Heterocyclic Compounds (Engl. Transl.), Chemistry of Natural Compounds (Engl. Transl.), Chemistry - A European Journal, Chemistry - An Asian Journal, Chinese Journal of Chemistry, Collection of Czechoslovak Chemical Communications, Croatica Chemica Acta European Journal of Organic Chemistry, European Journal of Pharmaceutical Sciences, Fitoterapia, Flavour and Fragrance Journal, Hecheng Huaxue, Helvetica Chimica Acta, Heterocycles, Insect Biochemistry and Molecular Biology, Journal of Antibiotics, Journal of Carbohydrate Chemistry, Journal of Chromatography A, Journal of Heterocyclic Chemistry, Journal of Labelled Compounds and Radiopharmaceuticals, Journal of Medicinal Chemistry, Journal of Natural Products, Journal of Organic Chemistry, Journal of Organometallic Chemistry, Journal of Synthetic Organic Chemistry (Japan), Journal of the American Chemical Society, Journal of the Chinese Chemical Society (Taipei), Korean Journal of Pharmacognosy, Magnetic Resonance in Chemistry, Mendeleev Communications, Molecular BioSystems, Monatshefte fuer Chemie, Natural Product Communications, Natural Product Reports, Natural Product Research, Natural Product Sciences, Korea, Nature, New Journal of

Chemistry, Organic & Biomolecular Chemistry, Organic Letters, Pharmazie, Phytochemistry, Phytotherapy Research, Planta Medica, Polish Journal of Chemistry, Revista Brasileira de Farmacognosia, Russian Chemical Bulletin, Russian Chemical Reviews (Uspekhi Khimii), Russian Journal of Organic Chemistry, Science (Washington, D.C.), Scientia Pharmaceutica, Steroids, SYNLETT, Synthesis, Synthetic Communications, Tetrahedron, Tetrahedron Letters, Tetrahedron: Asymmetry and Zeitschrift fuer Naturforschung, Teil B.

9. References

[1]. Arthur Dalby, James G. Nourse, W. Douglas Hounshell, Ann K. I. Gushurst, David L. Grier, Burton A. Leland, John Laufer "Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited," Journal of Chemical Information and Computer Sciences. vol. 32, 1992, pp. 244-255.

[2]. Subhash Chandra Bose. Kotte, Pavan Kumar K.V.T.S, Ravi Kumar Tumuluri, Shriram Raghavan, P.K. Dubey and P.M. Murali; Structural Library of Natural Compounds; IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, 2011, pp. 490-493.

[3]. J. K. Wegner. JOELib. http://joelib.sourceforge.net/.

[4]. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Deliv. Rev. vol 23, 1997, 3-25.

[5]. Lipinski, C.A. Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov. Today, 1, 2004, 337-341.

[6]. Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. J.Med.Chem, 41, 1998, 3325-3329.

[7]. Ajay, A; Walters, W.P.; Murcko, M.A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? J.Med.Chem, 41, 1998, 3314-3324.

[8]. Murcia-Soler, M.; Pŕrez-Gimŕnez, F.; Garcɔ´a-March, F.J.; Salabert-Salvador, M.T.; Diaz-Villanueva, W.; Castro-Bleda M.J. Drugs and Nondrugs: An Effective Discrimination with Topological Methods and Artificial Neural Networks. J. Chem. Inf. Comput. Sci., 43, 2003, 1688-1702.

[9]. Oprea, T.I. Property distribution of drug-related chemical databases. J. Comput. Aided Mol. Des. 2000, 14, 251-264.

[10]. Rishton, G.M. Reactive compounds and in vitro false positives in HTS. Drug Discovery Today, 2, 1997, 382-384.

[11]. Hann, M. M.; Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. Curr Opin Chem Biol, 8, 2004, 255-263.

First Author:

1. Subhash Chandra Bose. Kotte obtained his Masters in Chemistry from Jawaharlal Nehru Technological University at Hyderabad. He specializes in Chemical synthesis, Analytics and presently a researcher in the field of Analytical Chemistry with Jawaharlal Nehru Technological University Hyderabad.
2. K.V.T.S.Pavan Kumar obtained his Masters in biochemistry. He specializes in Bio-informatics and is presently a researcher in the field of Plant biotechnology with Dalmia Centre for Research and Development.
3. T. Ravi Kumar obtained his Masters in biochemistry. He is presently a researcher in the field of biochemistry with Acharya Nagarjuna University

Second Author:

Shriram Raghavan obtained his Masters in biochemical engineering and biotechnology from HP University at Shimla. He specializes in Bio-informatics and is presently a researcher in this field with Dalmia Centre for Research and Development.

Third Author:

1. Dr. P.K. Dubey is M.Sc.(OU), Ph.D (O.U), Specialized in Organic Chemistry. His research interests include Synthetic organic chemistry. He guided over 20 Ph.D. Students and presently guiding seven students for Ph.D. Degrees. He published over 150 research papers. He has 32 years of teaching and research experience.

2. Dr. P.M. Murali received a Ph.D. in Microbiology and Microbial technology from Madurai Kamaraj University, having over 22 years of experience in Pharmaceutical & Healthcare R&D, including management of more than 10 clinical trials, in particular in respiratory diseases. He has remained the Founder and Director of Dalmia Centre for Research and Development for 16 years, developing and launching natural product based therapeutics, and Founder and Chairman of MLC & Netpeople group of IT & Telecom companies (Networking solutions, Banking Security and communication services). Dr. Murali is a former Indo-US scientist at Battelle-Kettering, Ohio and fellow of Unilever India. Dr. Murali is the Managing Director and CEO of Evolva India since September 2006 and in addition presently he is President of the Association of Biotechnology Led enterprises (ABLE) in India.