JPP JOURNAL OF PROTEINS AND PROTEOMICS
8(3), 2017, pp. 133-148

Research Survey

CRITICAL ASSESSMENT OF CONTRIBUTION FROM INDIAN PUBLICATIONS: THE ROLE OF *IN SILICO* DESIGNING METHODS LEADING TO DRUGS OR DRUG-LIKE COMPOUNDS USING TEXT BASED MINING AND ASSOCIATION

Pawan Kumar, Gourab Das and Indira Ghosh*

School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, India

Abstract: Over the several decades, India is constantly challenged by communicable and non-communicable diseases which are originated either by poor lifestyle or by environmental factors. The pools of diseases are constantly posing serious threats to mankind especially among the poverty-stricken families. Scientific communities across the globe are working continuously to design drug molecules to overcome the burden of these life threaten diseases. In last three decades, many computational algorithms and tools have been developed to identify potential drug targets and their inhibitors. It is believed that computational techniques have reduced the time and money required to develop an inhibitor into drug. However, applicability and deliverability of these *in silico* techniques in rational drug designing are not fully evaluated. In the present study, PubMed/Medline extracted data driven analysis has been performed to highlight the influence and progress of the theoretical methods in the field of drug discovery across India and compared with the world. Drug discovery related keyword dictionary has been built and utilized to select only drug discovery related PubMed abstract. A second keyword set (related to bioinformatics tools) is used for normalized pointwise mutual information (PMI) based association analysis. Observations show that drug discovery has been an interdisciplinary research and used many tools starting with QSAR, docking, pharmacophore, Molecular Simulations etc. The publications contributed from India (2%) are similar as compared to the contribution in total world publications, suggesting large scope in future. Data coverage as represented since 1990-2015 in PubMed as indicated by number of publications associated with drug discovery is almost same in world and India (~75%). Emerging institutes/Universities are contributing since last 10 years as observed from Indian publication list. However, this method has many limitations as discussed.

Keywords: PubMed; Association; nPMI; *In silico* techniques; Drug discovery

Note : Supplementary Information available on Journal Website in "Archives" Section

Introduction

Interdisciplinary topics play an important role in the advancement of research in existing subjects from which they have evolved (Chakraborti *et al.*, 2016). In Medical & life sciences, the same role has been assigned to Bioinformatics & Computational Biology which has evolved since last fifty years (Hogeweg, 2011). Translational biology has become an emerging field of research nowadays. In the context of biological sciences, translation is the phenomenon by which mRNA or transcripts are converted to protein. In general sense, translational medical research can be defined as the process of transforming scientific discoveries from laboratories

Corresponding Author: **Indira Ghosh**
E-mail: indira0654@gmail.com

Received: September 1, 2017
Accepted: September 12, 2017
Published: September 16, 2017

to clinical applications (Contopoulos-Ioannidis *et al.*, 2003). In this paper, we would like to address recent advancement related to Translational Biology and impact of Bioinformatics & Computational Biology on these research topics. From a bioinformatics perspective, translational research integrates information derived from multi-omics applications with the clinical outcomes like disease, symptoms, drugs, vaccines etc. (Altman, 2012). Bioinformatics assists in storing, analyzing and visualizing data from various experiments to understand and diagnose patient's health conditions and finding remedies.

Research has been progressed far in the field of bioinformatics and translational drug discovery. Numerous models, methods and tools have been developed for *in silico* drug designing, identifying potential drug targets, novel inhibitors and for *in vitro* validation. Eventually, these rigorous efforts and attempts have been reflected in terms of exponential increment of publications in the electronic resources like Medline, Google Scholar, Scopus, and Web of Science etc. Amongst these databases, Medline is the oldest one which has been initiated by National Library of Medicine (NLM), the USA in 1971 (National Library of Medicine, 2004). Though, the offline counterpart of Medline i.e., Medlars has been started early in the year of 1964. In 1997, NLM first introduced a collection of selective old Medline citations staring from 1809 to 1965 and all citations from 1965 onward in the form single publicly free repository PubMed (NCBI Resource Coordinators, 2017) which has now evolved to more than 27 million citations. All other databases, above mentioned, have been officially announced later in the year of 2004. Comparison of all of these databases including their advantages, disadvantages, available features has been successfully reviewed by Falagas *et. al.* (Falagas *et al.*, 2007).

Presently, PubMed/Medline has included citations from more than 5600 journals available worldwide in more than 40 different languages. Broadly, the subject areas covered by PubMed/Medline includes life sciences, behavioral sciences, chemical sciences, bioengineering, biology, environmental science, marine biology, plant and animal sciences as well as biophysics (National Library of Medicine, 2004; NCBI Resource Coordinators, 2017). Additionally, a small number of biomedicine related newspapers, newsletters, magazines and book chapters have been added to the database. Though subject coverage in other databases like Google Scholar, Web of Science are not only restricted to life sciences and clinical sciences but also physical sciences, humanities, economics, business, administration related publications have been added too. Advanced search system, flexibility in knowledge extraction and cross linking with other sequence, structure databases have made PubMed a comprehensive, optimal and easy to use resource for biologists and medical practitioners.

PubMed Central (Roberts, 2001) is another initiative from NLM which is launched in 2000 to archive digitally the full-text journal articles on biomedical and life sciences. PMC is a subset of PubMed/Medline but the difference between PubMed and PMC is that PubMed contains only article abstracts and books whereas PMC only includes free available full journal articles. PMC articles are often enriched by enhanced metadata, medical ontologies, fully indexed (PMCID) and XML structured data after submission. Moreover, users can freely read and download full texts of the articles in portable document format (pdf) which can be stored in local libraries for further use. To date, approximately 4.4 million articles have been archived in PMC from different sources including 2024 fully participated, 4331 selective and 328 NIH portfolio journals.

With the help of this study, we are attempting to examine the role and impact of *in silico* drug designing methods in novel drug discovery related publications from India in comparison with world.

Materials and Methods

Advanced query building and data download

PubMed (*http://www.pubmed.com/*) can be searched through a browser or programmatically using NCBI utilities (tools like biopython). Advanced queries have been constructed (Table 1) to collect the title, author, Reference, abstract, keyword set during 1990 to 2015. During the advance query set up, we have only considered PubMed record having abstract information but no book abstract is selected. Downloading big data (~1GB) using biopython program (Cock *et al.*, 2009) took ~85 hours, estimated size of world data (~27GB) will not be achievable with biopython program. Hence, publication data for India and the world are downloaded using PubMed browser. World data

has been downloaded in chunks of 5 years using the queries from Table 1 and kept as Medline formatted flat files. Parsing of the Medline files has been done using custom biopython scripts and PMID, abstract, title, keywords and date of publication has been stored in tab-separated format as one entry per line for performing further keyword-based searching. Additionally, these huge data is also stored in MySQL database for future use. In this paper, we have used World data as all publication records excluding Indian publication record during 1990-2015.

Creation of dictionary of keywords

To find appropriate search operations on the two sets of data sets (India and world), three different keyword files are created namely 1-Keyword.txt and 2-Keyword.txt and 3-Keyword.txt. 1-Keyword set has drug discovery related terms, 2-Keyword set has Computational biology and bioinformatics related terms while 3-Keyword set has only medicinal chemistry related terms.

The first file includes terms extracted of drug discovery dictionary (Ganellin *et al.*, 1998; Davies *et al.*, 2016; Wehr, 2001; Buckle *et al.*, 2013), words from the tables (Khan *et al.*, 2011) and other keywords described/included in the text (Spedding, 2006; Wilson and Lill, 2011; Njogu *et al.*, 2016) after manually going through the papers. In addition, a few missing words (list shown in Supplement 1) are also included. This file also includes redundant names used in literature for comprehensive selections. After properly annotated/formatted for searching literature and

Table 1: Important advanced PubMed queries used in present study

No.	*Search criteria*	*PubMed/Medline Advanced query*	*Total No. of records in PubMed*
1.	All abstracts till 2017 July 15th	All[sb] AND hasabstract[Text] NOT pubmed books [FILTER]	1,74,20,910
2.	All abstracts in between 1990 to 2015	All[sb] AND hasabstract[Text] AND 1990:2015[PDAT] NOT 1800:1989[PDAT] NOT 2016:2018[PDAT] NOT pubmed books[FILTER]	1,33,28,337
3.	All abstracts in between 1990 to 2015 from India	All[sb] AND hasabstract[text] AND 1990:2015 [PDAT] NOT 1800:1989[PDAT] NOT 2016:2018 [PDAT] AND India[AD] NOT pubmed books [FILTER]	2,58,886
4.	All abstracts in between 1990 to 1995 excluding India	All[sb] AND hasabstract[text] AND 1990:1995 [PDAT] NOT 1800:1989[PDAT] NOT 1996:2018 [PDAT] NOT India[AD] NOT pubmed books [FILTER]	18,03,683
5.	All abstracts in between 1996 to 2000 excluding India	All[sb] AND hasabstract[text] AND 1996:2000 [PDAT] NOT 1800:1995[PDAT] NOT 2001:2018 [PDAT] NOT India[AD] NOT pubmed books [FILTER]	17,96,260
6.	All abstracts in between 2001 to 2005 excluding India	All[sb] AND hasabstract[text] AND 2001:2005 [PDAT] NOT 1800:2000[PDAT] NOT 2006:2018 [PDAT] NOT India[AD] NOT pubmed books [FILTER]	23,23,317
7.	All abstracts in between 2006 to 2010 excluding India	All[sb] AND hasabstract[text] AND 2006:2010 [PDAT] NOT 1800:2005[PDAT] NOT 2011:2018 [PDAT] NOT India[AD] NOT pubmed books [FILTER]	30,21,525
8.	All abstracts in between 2011 to 2015 excluding India	All[sb] AND hasabstract[text] AND 2011:2015 [PDAT] NOT 1800:2010[PDAT] NOT 2016:2018 [PDAT] NOT India[AD] NOT pubmed books[FILTER]	39,58,979
9.	All abstracts in between 1990 to 2015 excluding India	All[sb] AND hasabstract[text] AND 1990:2015 [PDAT] NOT 1800:1989[PDAT] NOT 2016:2018 [PDAT] NOT India[AD] NOT pubmed books[FILTER]	129,03,764

clustering all redundant names, a total of 832 entries are placed in the 1-Keyword.txt.

In the same way, 2-Keyword.txt (includes 250 computational bioinformatics related terms) and 3-Keyword.txt (includes 256 medicinal chemistry related terms) files are created (for further details and references see Supplement 1).

As the selection list of articles/publications will depend on these keywords, enough care was taken to include many references in the related field. First set of keywords file will be used to select the PubMed abstract using the content in title, abstracts & keywords in publication list having drug discovery terms, if available. 2-Keyword, specific for tools/methods used in Computational Biology and Bioinformatics, will be used to find the relevance of the presence of such words uniquely & interdependence in pair via Unigram, Bigram and Trigram, their role in finding drug related compound also has been analyzed at the end (Figure 1).

After the generation of in three Keyword files, overlap among terms in the these files has been analyzed including their redundancy (see method in Supplement 1). Venn-diagram plot (Figure 2) shows that 1-Keyword set has nearly 30% overlap with other two keyword sets while 2-Keyword set has ~32% overlap with 3-Keyword set while common most overlap is ranging from 36% for 3-Keyword to 13% for 1-Keyword. This plot indicates that importance of having three Keyword files representing different Keyword space.

Selection of publications related to drug discovery research

Publications related to drug discovery have been searched in the tab separated flat files using dictionary file 1-Keyword.txt (see Material Section). Regular expression based searching method has been opted as it is easy to understand, flexible and can find multiple patterns in a single click search. A regular expression can be defined as a generalized string of normal and special characters that can be arranged as per requirement to match the pattern in the text. Perl regular expression (Regex) (Friedl, 2002) has been used to find the patterns enabling its case-insensitive and exact match mode (more detail is in Supplement file 1). In the extracted publications, for each of the term (unigram), searching is performed in publication titles, abstracts, keywords and respective counts for the unigrams have been normalized by the total number of records in the publication list.

Concept of association mining

It is often useful to look further for the combination of the concepts those are frequently co-occurring in publication texts. Estimating the strengths of the co-occurrences would be beneficial in determining the relationship among the concepts for various scientific interests. Traditionally, there are several popular statistical methods available like Pearson, Spearman and Kendall's correlation for estimating the association between two or more variables. However, each of these methods has its own pros and cons. For example, Pearson's correlation measurement is only applicable to estimate the linear dependence between the variables whereas Spearman's and Kendall's correlation do not assume linearity in the relationship and used for ordinal variables and monotonic relationships (Santos *et al.*, 2013) respectively. On the other hand, an information content estimation based method like pointwise mutual information (PMI) can be applied to any variables with linear or non-linear relations. PMI estimates for two independent discrete variables, the discrepancy between their coincidence when the probability of joint occurrence & their single occurrence are known. It quantifies the amount of information shared by the two or more random variables in terms of joint probability compare to their individual probabilities of occurrence, indicating full co-occurrence have +1, full independence have 0 and never occurred together is -1.

In the present work, it is very difficult to make a prior assumption about the nature of the correlations between the terms. Hence, normalized PMI (nPMI) (Bouma, 2009) based method is used for association analysis. Each of the unigram is paired with other unigrams and associations between all possible pairwise combinations (non-redundant) have been measured separately using the following formula

$$Bigram(x_i, x_j) = \ln \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \ / -\ln\ p(x_i, x_j) \quad (1)$$

$$Trigram\left((x_i, x_j), x_z\right) = \ln \frac{p(x_i, x_j, x_z)}{p(x_i, x_j)p(x_z)} \ / -\ln\ p(x_i, x_j, x_z) \quad (2)$$

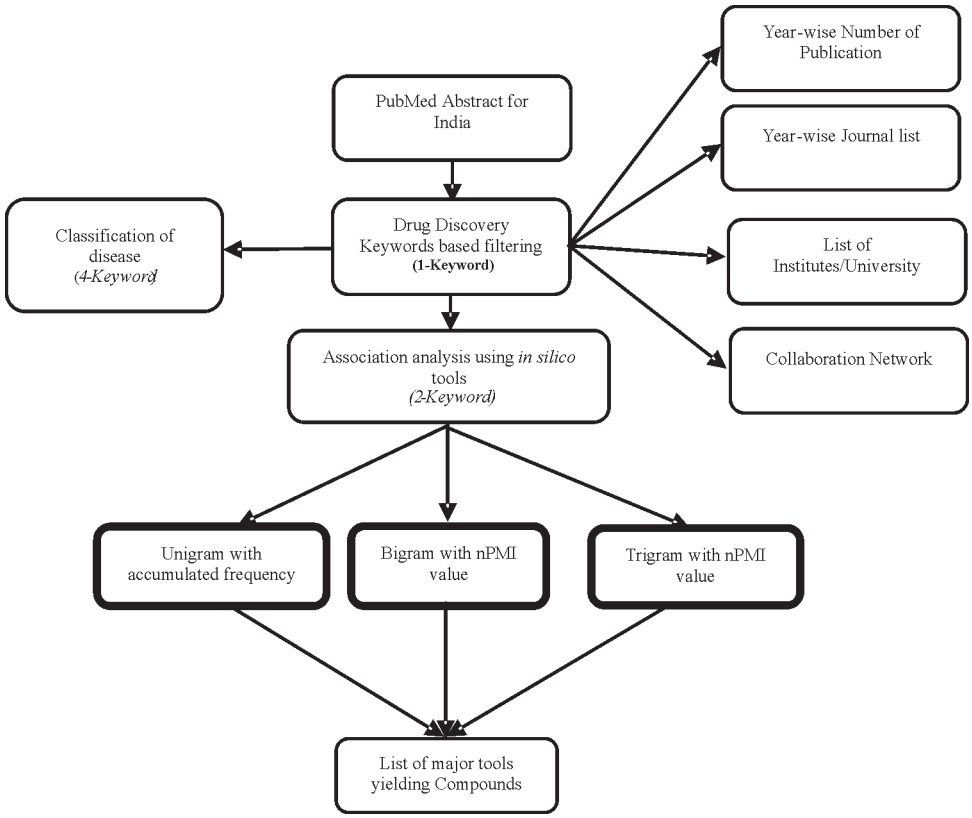

Figure 1: **Workflow for PubMed data mining and association calculations**

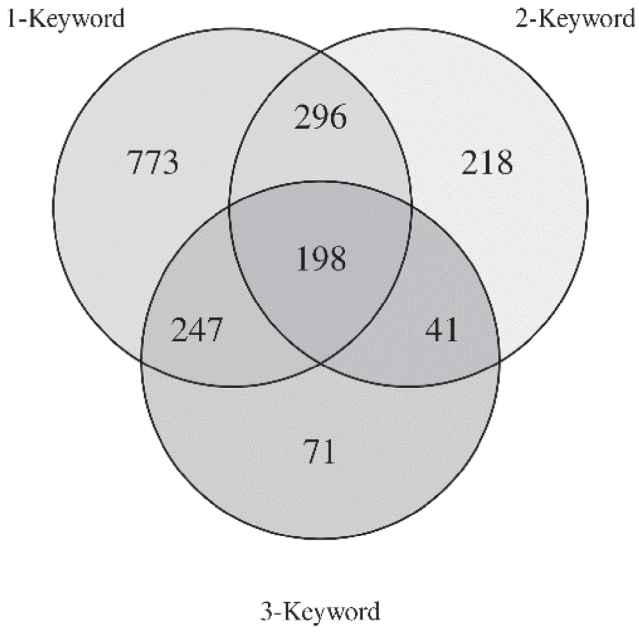

Figure 2: **Venn-diagram plot showing the number of entries belonging to different keyword file. All terms (redundant inclusive) in each keyword are used for the diagram. (1-Keyword is 1514, 2- Keyword is 753& 3-Keyword is 557). The overlaps between 1-Keyword with other keywords are ~30%**

Where, $i_n(x_i, x_j)$ is the nPMI value shared between i^{th} and j^{th} terms in the list of publications in terms of their individual probabilities $p(x_i)$, $p(x_j)$ and their joint probability $p(x_i, x_j)$. Combination of the two terms or unigrams is named as bigrams. Similarly, each bigram (x_i, x_j) is combined with rest of the unigrams (x_z) to construct the trigram and nPMI is calculated. Probability of occurrence and mutual dependency among the concepts beyond trigrams are expected to be small and execution time for computation is in exponential order. As we already have an estimation of terms occurring in bigrams more frequent than others, so the trigrams calculation is done using the selective preferred terms occur in bigrams. Individual counts, joint counts, nPMI value and corresponding unique PMID values for bigrams and trigrams have been listed in Table 1, 2 and 3 of Supplement 1 during the calculation for further analysis.

Drug discovery related abstract analysis

It is necessary to make classifications to study disease associated with the publications in the present analysis. To date numerous diseases have

been emerged causing morbidity and mortality among human populations. World Health Organisation (WHO) has standardized and categorized the diseases, disorders, adverse health conditions into distinct classes (22) and reported in the International classification of diseases and health related problems (ICD 10th revision) (see Supplement 1) (WHO, 2010; Kanehisa *et al.*, 2017). Searching has been performed using Global Regular Expression Print (GREP) (Navarro, 2001) using its case-insensitive and extended regex mode. To ensure exact match word boundaries have been incorporated to each of the word/phrase in a particular group.

Results and Discussion

All data are retrieved using 1-Keyword set, for 26 years (1990-2015) from PubMed publications for India only & World (excluding India). Combining all data i.e., from 1990 till 2015 (five sets in the world data, Table 1) a MYSQL database is made and the same for India (in single set) is developed. All following operations & analysis are done using these two sets of publication records (1990-2015) from World excluding India and India only. The world data being large and to reduce CPU time, it was divided into five sets and accumulated for final searching and calculations. Figure 1 shows the flowchart for the process followed.

Data and keyword space analysis

The uniformity of such space with respect to terms in the keyword set as well as homogeneity in keywords distribution has been analyzed and will be useful to estimate the accuracy of this work. Figure 3A and Figure 3B show the distribution of hits in well separated ten different clusters for India (For world Figure 2 and 3 in Supplement 2). These k-means cluster shows that homogeneity in the used keywords space. We have also plotted the distribution of hits (normalized frequency Figure 4 and 5 in Supplement 2) of the terms in both the keyword sets when search have been done in India and world publication database. Histogram plot shows that hits are distributed in different bins and only one or two bins have extreme values (at the

Table 2: Selection of PubMed record using 1-Keyword based selection

Year range	*No. of Publication Abstracts*	*No. of publications related to 1-Keyword dictionary*	*Number of unigrams with non-zero hits*
India (1990-2015)	258885	195878 (75.66%)	701
World (1990-1995)	1803683	1328173 (73.64%)	699
World(1996-2000)	1796260	1359634 (75.69%)	726
World (2001-2005)	2323317	2322317 (76.14%)	757
World (2006-2010)	3021525	3021525 (75.83%)	770
World (2011-2015)	3958963	3034050 (76.64%)	770

Table 3: Top 10 terms selected from the 3-Keyword set based Unigram analysis in Indian Database

Index	*Term name*	*Count*	*Unigram Score**	*Unigram Score$*
A52	Compound	56096	0.286	0.286
A86	Enzyme	19618	0.1	NA
A238	Target	17326	0.088	0.088
A83	Efficacy	12686	0.065	NA
A17	Assay	12251	0.063	0.063
A131	Lead	12143	0.062	0.041
A144	Metabolite	7282	0.037	NA
A254	V Screening	6830	0.035	0.035
A100	Genome	6696	0.034	NA
A128	LBDD	5849	0.03	0.03

NA means that term is not in the 2-Keyword set; *Normalized term count is used as Unigram Score for 3-Keyword set; $Normalized term count is used as Unigram Score for 2-Keyword set

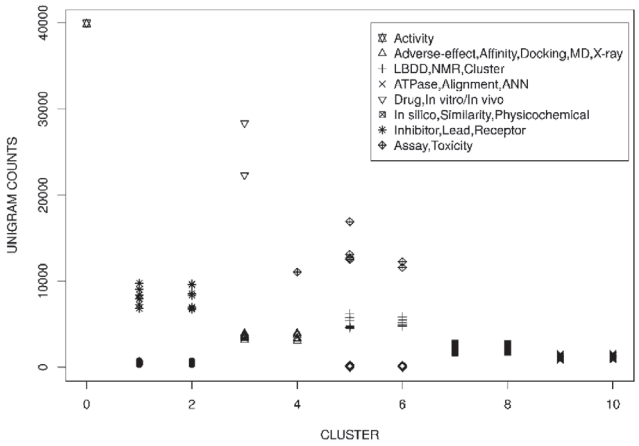

Figure 3A: **K-mean clustering analysis using hit count from the 1-Keyword based selection in Indian database**

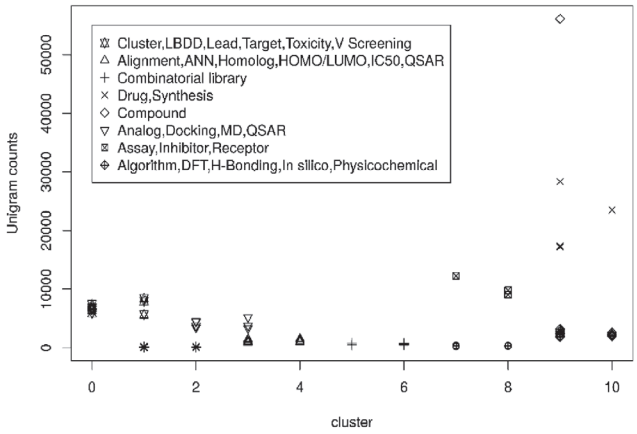

Figure 3B: **K-mean clustering analysis using hit count from the 2-Keyword based selection in Indian database**

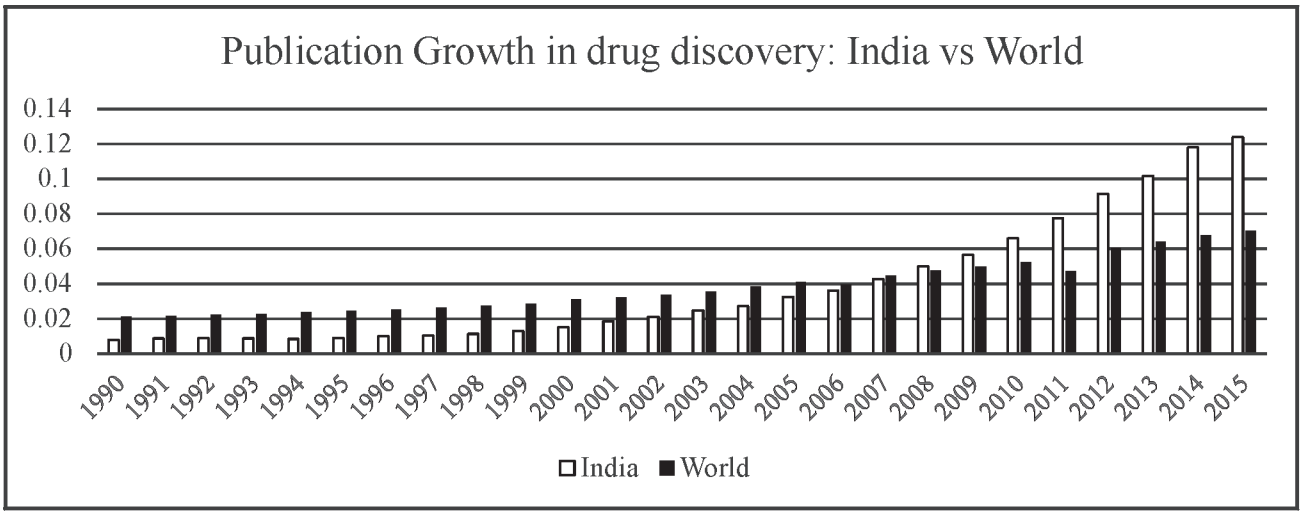

Figure 4: **Plot shows the progress in a number of papers published in drug discovery from India and world during 1990-2015. The normalization is done using total 26 years publications**

right most side). In the world database extracted from main PubMed is sufficient to estimate the occurrence of hits vs. non-hits of terms used in the 1-Keyword. It was observed (Table 2) that it has consistent from 1990 to 2015 almost 75% (also see Figure 1 Supplement 2). These demonstrate that choice of terms associated with drug discovery is appropriate. The present analysis shows that selection of the keywords as well as data space is large enough to derive logical decisions from it.

Journals and Institutes associated with drug discovery research

Total number of Publications with Indian addresses during 1990-2015 in PubMed is almost 2.6 lakh articles, a good enough space to search for ~832 Keywords (0.33%). In PubMed, India's contribution (in all areas) during 1990-2015 is only ~2%; the same is found in publications related to drug discovery field (after 1-Keyword selection). However, in world (excluding Indian Publications) and Indian database, the coverage of keywords is ~75 % (Table 2). After applying 1-Keyword dictionary, 195878 publications from India, i.e., related to drug discovery (~76% of total contributed from India) are extracted. A year wise bar plot for India, normalized by total number of publications, illustrates the progress of publications in compare with the world (Figure 4). As seen from Figure 4 compared to world drug discovery data, Indian publishing is increasing at a higher rate than the world, which shows that there is a growing interest and encouragement in translational research around 2007 in India (Krishnaswamy and Mohan, 2016).

Figure 5A shows that journals in which most often drug discovery papers are published from India are Expt. Biol, Med Chem. and PLoS other than Indian J. of Exp. Med. After 2006, PLoS publication started and it is observed that scientists prefer to publish in the same along with Med. Chem, J Clin.Diagnostic etc. For the world, as shown in Figure 5B, PLoS One, J. Biol. Chem, Proc. Natl. Ac. Sci USA, Biochem. Biophys Res. Comm and J. Immunology are the preferred journals in translational research on average.

Figure 6 shows using only Indian data which University/Institutes are publishing related to drug discovery topics and how the numbers are growing over the year. It is observed that AIIMS, Delhi has maintained its pace and consistent from 1990 to 2015 in publishing papers but University of Delhi, Delhi has emerged in around 2000 with greater role in terms of number of publications and consistently increased its publication numbers. Total contributions of these institute/universities are also shown in Figure 7.

Collaboration network

The list of publications related to drug discovery is also used to find how the collaboration has been achieved in Indian scientific community. To analyze the collaborative nature among the Indian Institute/ University, a network is constructed using Institute/University dictionary (172 entries) which have been extracted (search field: AD) from the present database. A combination of pairs was searched together to find all possible collaboration. Connection lists are prepared by counting the number of time any two institutes/University come together in the affiliation address. Edge weight is calculated normalizing the pair count by the total number of publication from two Institutes/ universities.

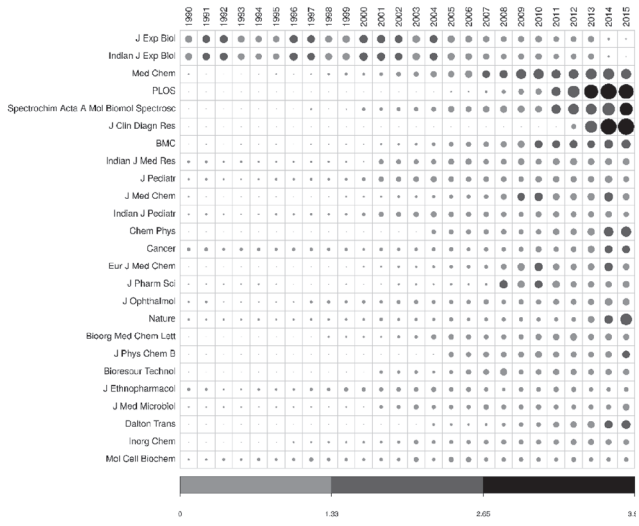

Figure 5A: **Top 25 journals on drug discovery related papers are published from India during 1990-2015. Bubble sizes indicate the 3 level of publications (Normalized by total papers, expressed as x10^3).**

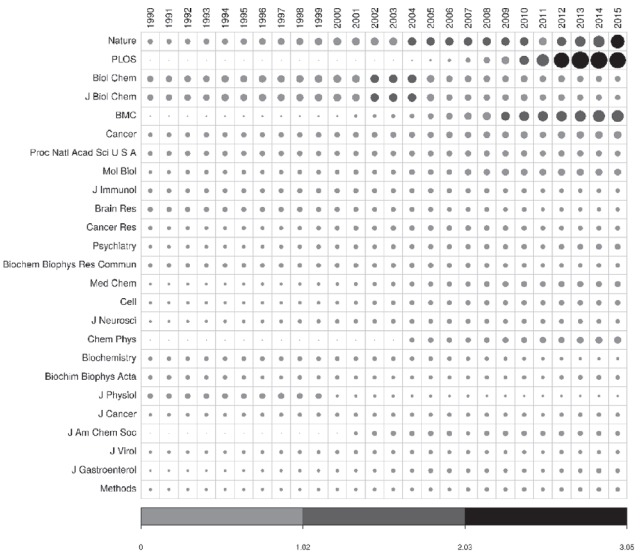

Figure 5B: **will show the top 25 Journals representation for World data (excluding India) after selection using 1-Keyword, (Normalized by total papers, expressed as x103) year-wise evaluation, the range of weight for number of publication will differ from India, bubble size indicates 3 level of publication hits**

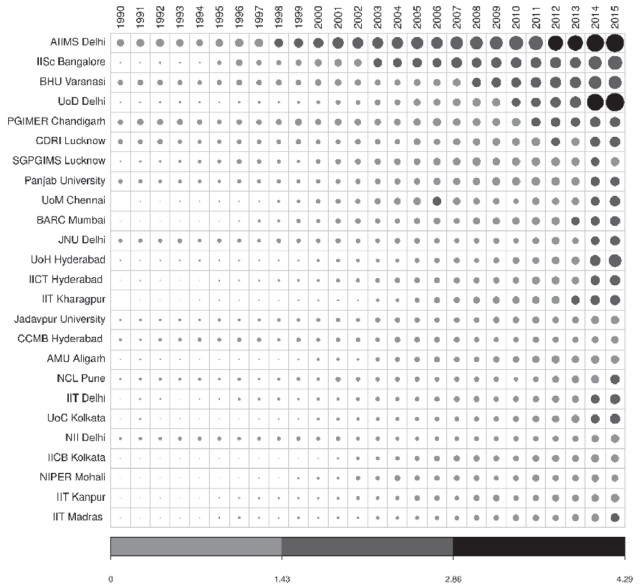

Figure 6: **Top 25 universities/Institutes from India publishing in drug discovery during 1990-2015. Size of the bubble indicates the cluster size. Year wise count is normalized by total papers, expressed as x10^3.**

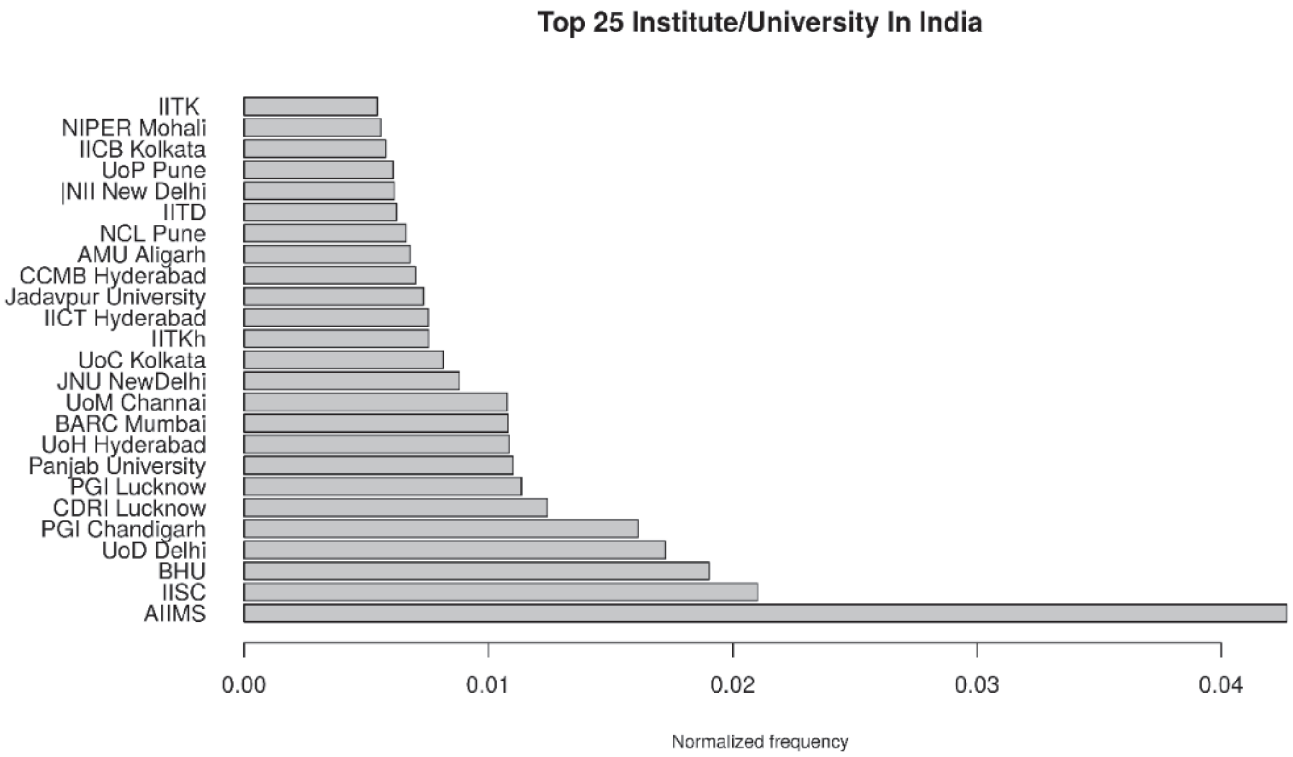

Figure 7: **List of top 25 Institute/ University in India working on Drug Designing**

Edge weightage = $\frac{N_{ij}}{N_i + N_j}$ where N_{ij} is the number of time i and j occur together while N_i and N_j represent the individual count.

Network plot (Figure 8) shows some Indian Universities that are among the most connected University while NCBS and TIFR are interconnected. IIT Delhi, IIT Kanpur, IIT Madras and IIT Kharagpur are among the top publishing institute/University list; however, they are less collaborative institutes.

Classification to study disease association with the publications

Searching Indian and World abstract data using classified disease set, (as mentioned in Methods and Materials) as shown in Figure 9 and Figure 10, that both Indian and world scientific community is working on major life-threaten disease like Cancer and Infectious disease, as reflected from publications.

Unigram, Bigram and Trigram analysis

It is worthwhile to analyze, how the bioinformatics tools have influenced delivering compounds. This

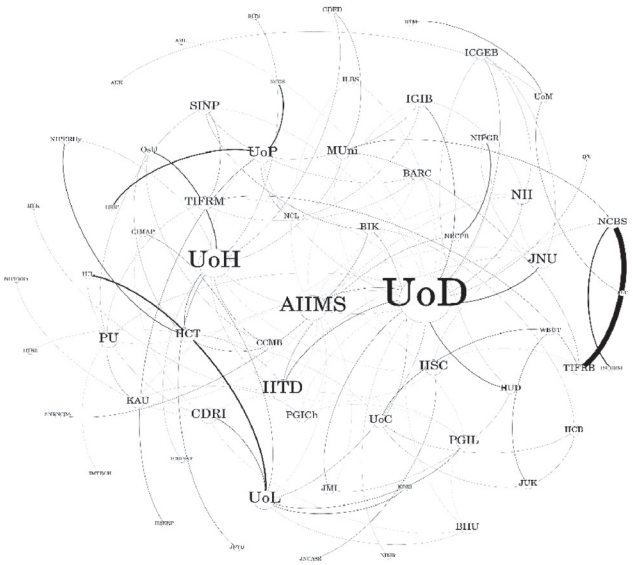

Figure 8: **Network showing the collaboration nature among the Indian Institute/University. Size of each node depicts the degree of the nodes while edge thickness shows the corresponding weightage with each other. The network clearly shows that University of Delhi & University of Hyderabad the highly collaborating institute/university, while NCBS and TIFR Bangalore are interconnected (the same institute). For clarity, collaboration pairs having at least ten papers in common are shown here. Full Network is enclosed in the Supplement file 2 (Figure 6).**

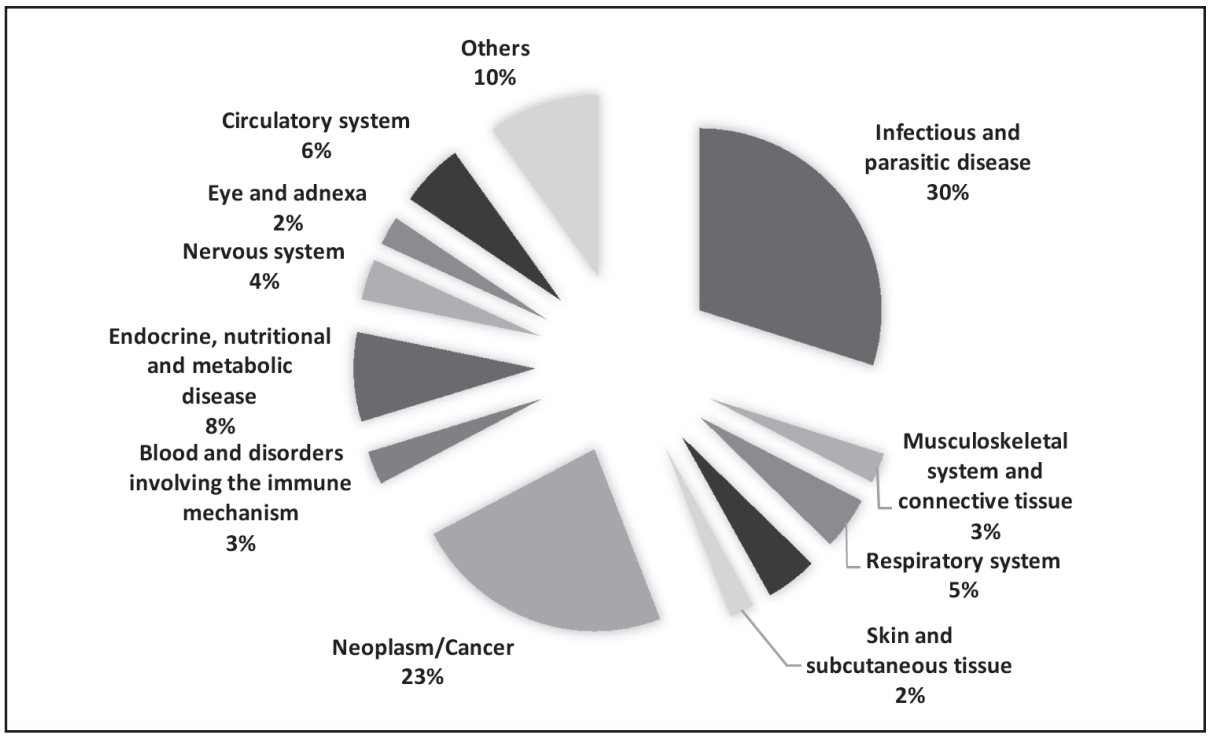

Figure 9: **Major Disease classes extracted from Indian abstract data (1990-2015).Indian Scientific Community has major focus in Infectious disease followed by Cancer.**

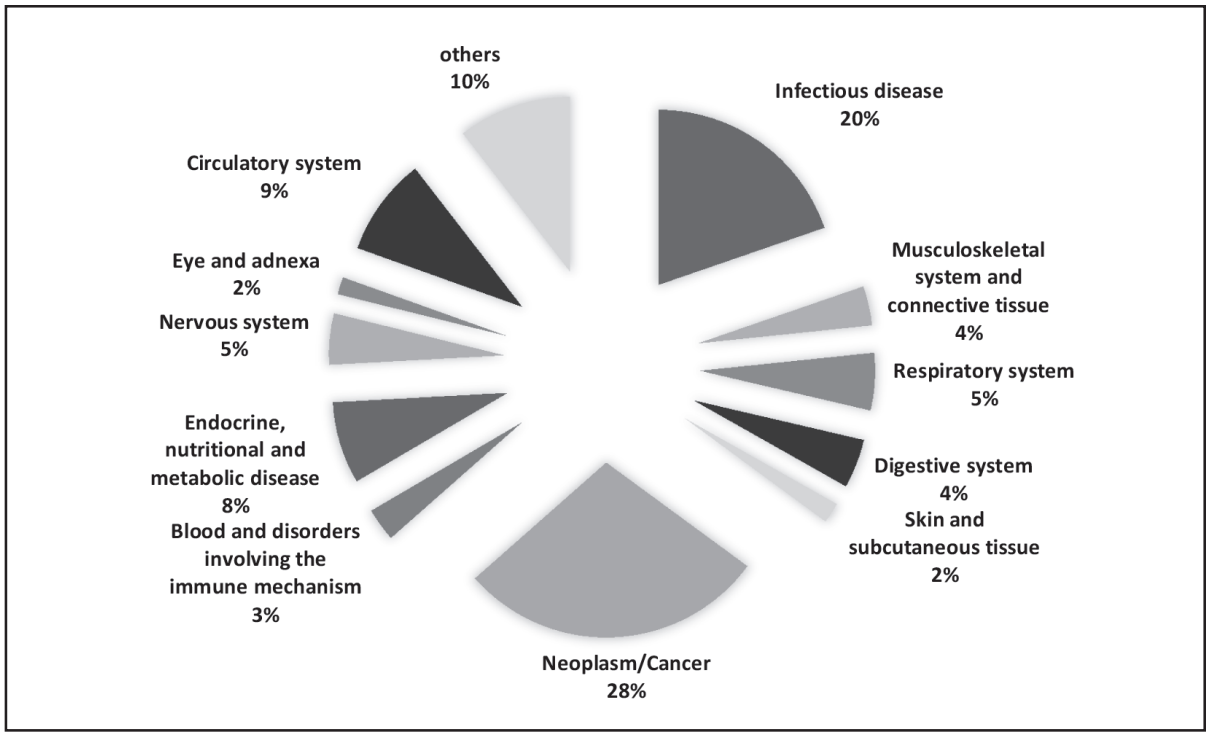

Figure 10: **Major disease classes extracted from the World data (1990-2015).**

Table 4: Top 15 terms selected from the 3-Keyword set form Bigram analysis in Indian Database. Bold rows show the common tools combination with 2-Keyword based Bigram analysis

Term 1	*Term 2*	*nPMI*Score*	*nPMI$^{\$}$Score*
Pharmacophore	**QSAR**	**0.575**	**0.576**
Docking	**Pharmacophore**	**0.468**	**0.467**
Docking	**Homolog**	**0.411**	**0.449**
Docking	**SBDD**	**0.375**	**0.375**
HTS	V Screening	0.375	0.375
Docking	**MD**	**0.363**	**0.363**
Docking	**QSAR**	**0.359**	**0.359**
Pharmacophore	**V Screening**	**0.356**	**0.357**
Docking	**LBDD**	**0.337**	**0.337**
LBDD	Pharmacophore	0.334	0.334
Analog	QSAR	0.318	0.318
Docking	IC50	0.291	0.291
LBDD	SBDD	0.289	0.289
Analog	CoMFA	0.279	0.279

*nPMIScore: Normalized Pointwise Mutual information using 3-Keyword set

$^{\$}$nPMI Score: Normalized Pointwise Mutual information using 2-Keyword set

analysis has been done using Bioinformatics related keyword file (2-Keyword set) and association is calculated using nPMI method (described in Method section) in the subset of PubMed data extracted. Use of such method provides equity between the compared databases as India vs. world which is almost 8 – 15 times larger. Unigram, Bigram & Trigram (s) are used to evaluate the occurrence of computational biology associated tools and their influence in translational research, especially in drug design related publication databases.

Analysis of Unigram or Frequency of occurrence of each computational tool related with drug discovery using the 2-Keyword set for finding an association in the database was followed up. Figure 12 shows that Compound, Synthesis, Target, Assay, Toxicity, Lead, Virtual Screening, Cluster, LBDD, Specificity, NMR, Analog, QSAR, Docking, and Electron Microscopy are among the top preferred terms. Except for some terms like Assay, Toxicity, NMR and Electron Microscopy, all other terms belong to the Computational tools. 2-Keyword terms are then compared with the world in five different year wise sets (1990-1995, 1996-2000, 2001-2005, 2006-2010, 2011-2015) and it is found that trends of preferentially occurrence of these terms are nearly same as Indian publications (Figure 11 & 12).

It is observed that NMR, Electron Microscopy and Crystallography has contributed almost equally to the publication related to drug discovery like bioinformatics tools like Docking or LBDD from India (Table 2 Supplement 1). Most well-known and popular tools in medicinal chemistry like QSAR/QSPR are less preferred in the frequency measure (normalized frequency <0.005) compared to docking/LBDD/MD/Virtual Screening (Table 2 Supplement 1). To compare with medicinal chemistry related keyword (3-Keyword) association calculation is done using the publication data from Indian publication data, later (Table 3 and 4).

All chemical research publications are not included in PubMed, like patent related data; so our dataset will have limited coverage, it will be of future importance to search such databases using same query for comprehensive analysis. A comparative plot (Figure 12) shows that Target based translational research plays more significant role in the world literature than in Indian publication, which provides the scope for future perspective in drug research in India, i.e, target-based concept.

To analyze, how many of these tools uniquely or in combination has influenced to produce chemicals from drug discovery publications in India; we have calculated Bigram and Trigram with

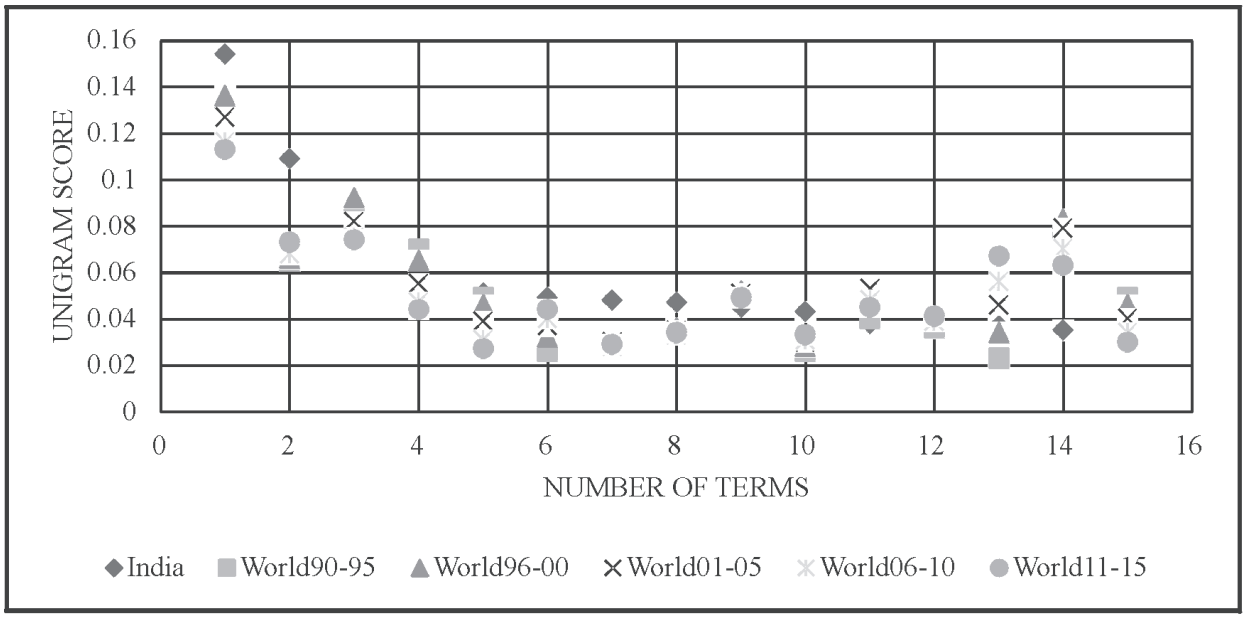

Figure 11: **Top 15 topics from 1-Keyword set in India during 1990-2015. The trend for these topics is analyzed with respect to world data set (1990-1995, 1996-2000, 2001-2005, 2006-2010 and 2011-2015). Top 15 unigrams included here are 1. Activity, 2. Drug, 3. In vitro/In-vitro, 4. Dose, 5. Enzyme, 6. Efficacy, 7. Toxicity, 8. Assay, 9. Acute, 10. Bacteria, 11. Inhibitor, 12. Exposure, 13. Target, 14. Receptor and 15. Membrane.**

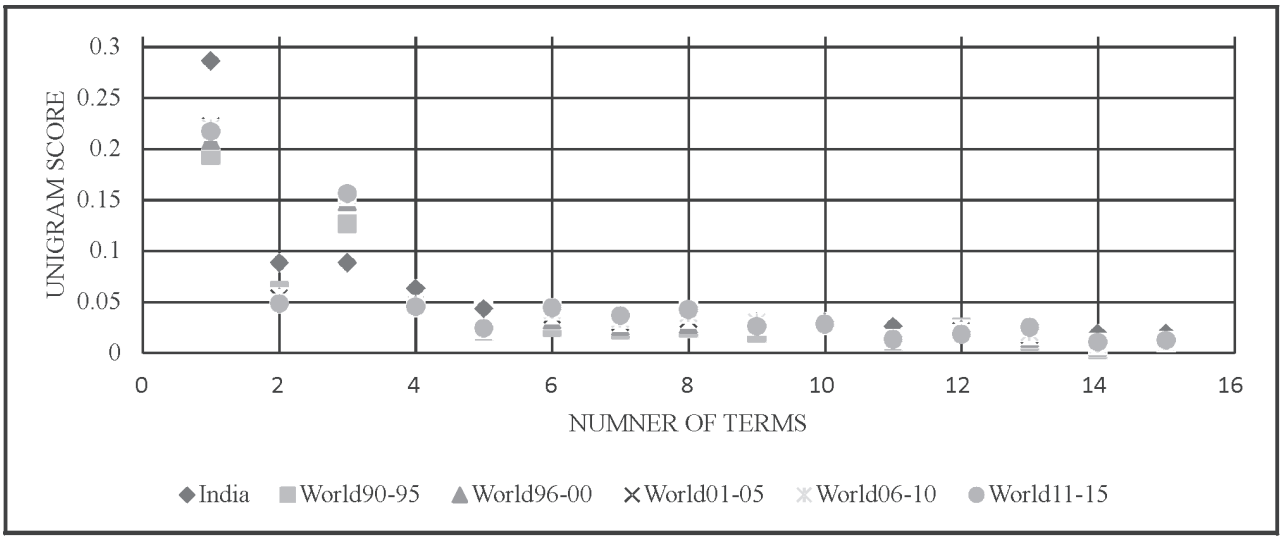

Figure 12: **Top 15 topics from 2-Keyword set in India during 1990-2015. The trend for these topics is analyzed with respect to world data set (1990-1995, 1996-2000, 2001-2005, 2006-2010 and 2011-2015). 1. Compound, 2. synthesis, 3. Target, 4. Assay, 5. Toxicity, 6. Lead, 7. Virtual Screening, 8. Cluster, 9. LBDD, 10. Specificity, 11. NMR, 12. Analog, 13. QSAR, 14. Docking and 15. Electron Microscopy.**

"Compound" and the preferred in Unigrams as mentioned above. Bigram analysis with the top 15 different tools as shown in Figure 13, are used in combination. However, amongst a large number of combinations possible, only those which may be feasible (knowledge driven concept) are shown here and a complete list is presented in Table 3 Supplement 1. The Figure 13 shows that some pairs are highly (0.737 to 0.337) preferred as Bioinformatics tools in Drug discovery publications. Minimum 20 papers criteria (nPMI ~ 0) is used for any combination of tool/method in case of Indian publication and equivalent numbers considered in the world database being larger. Though single occurrence of QSAR/Pharmacophore/ Physicochemical term is less frequent in Unigram,

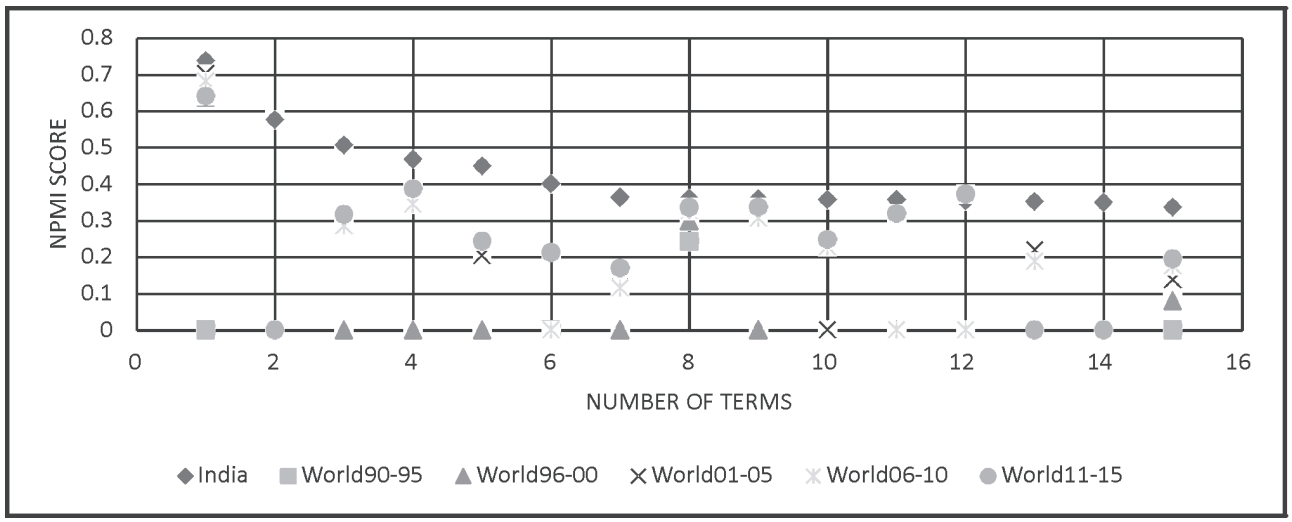

Figure 13: **Combination of 15 different tools 1. (Descriptor &QSAR), 2. (Pharmacophore & QSAR),3. (Docking & In silico),4. (Docking &Pharmacophore), 5.(Docking &Homology),6. (Docking &MD),7. (LBDD &X-ray), 8.(Docking & QSAR),9. (Pharmacophore & Virtual Screening), 10.(Pharmacophore & Scaffold-hopping), 11.(SBDD &X-ray),12. (Descriptor & Regression analysis), 13. (Descriptor & Pharmacophore), 14.(Docking & LBDD)and 15.( )are used for bigram analysis. This figure shows different trends with time in world bigram values. In the world, publications with combination tools emerging after 2000 and most of them are not associated at all.**

in combination tools, they have a greater impact (0.5 to 0.3) in drug discovery, as seen from Bigram analysis, from World & India (Figure 13), emphasizing the fact that they are used more associated with designing chemicals/compounds rather than isolated papers, as expected. For Indian publications, Pharmacophore & QSAR association is high (nPMI = 0.576) while it is below significance for world PubMed data set, this reflects the utility of single or combined usage of tools which produce publications. To evaluate the influence of any tool in drug designing in translational research, Bigram analysis is performed with "Compound". The results shown in Figure 14, it can be inferred that

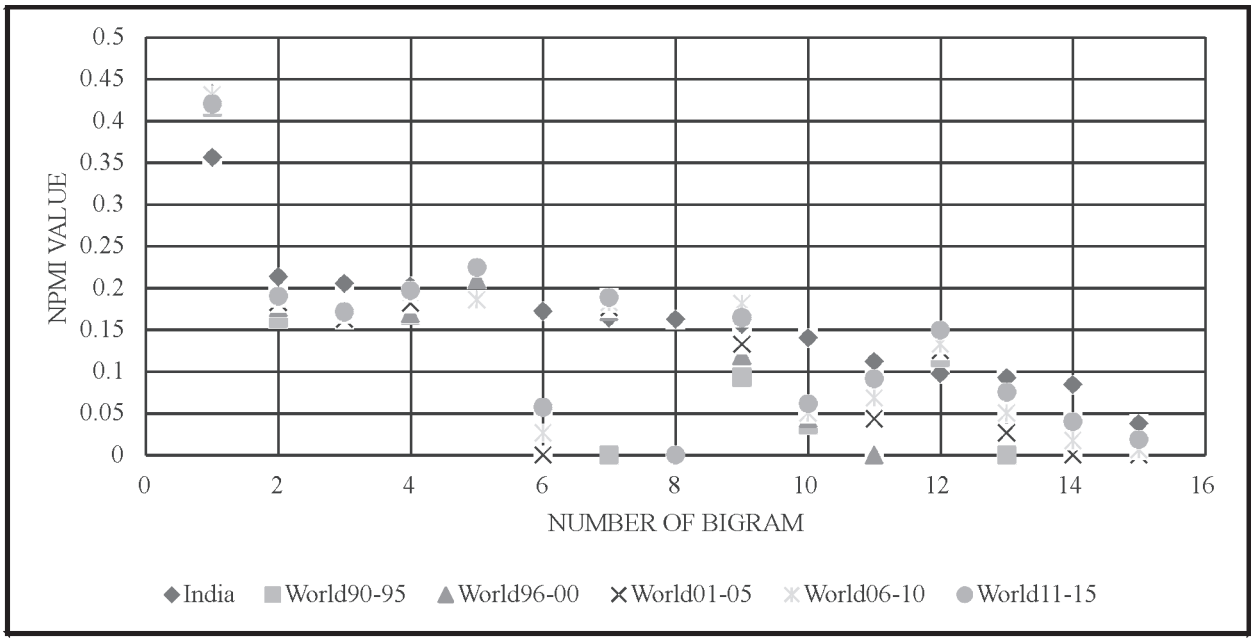

Figure 14: **15 different tools (1.LBDD, 2. QSAR, 3.Physicochemical, 4.Pharmacophore, 5.IC50, 6.Docking, 7.SBDD, 8.CoMSIA, 9.X-ray, 10.HOMO/LUMO, 11.In silico, 12.EC50, 13.Scaffold-hopping, 14.homolog, 15.Virtual screening) are analyzed with association to Compound and nPMI value is calculated. First LBDD is most preferable tool in world community than India and docking and virtual screening terms started emerging after 2010 in world Publications.**

both Indian and world drug discovery community prefer LBDD approach than docking and virtual screening has emerged (significantly higher) after 2010 in world publications. QSAR and pharmacophore modeling tools continued to play important role for drug discovery compared to docking and SBDD.

However, a Trigram analysis will reveal the association for usage in pair in compound discovery, as one can take the three terms together. Selected terms from the experience gained with Unigram and Bigram are used for Trigram analysis. The observation as shown in Figure 15 supports the idea that Drug Discovery is a multi-disciplinary research. High value (~0.35) is listed with Compound & (Pharmacophore & Virtual screening), which signifies the use of such tools to translational research. Other preferable combinations are (Docking & Virtual screening), (Pharmacophore & Docking), (QSAR & Docking) which are supposed to be guiding Compounds too.

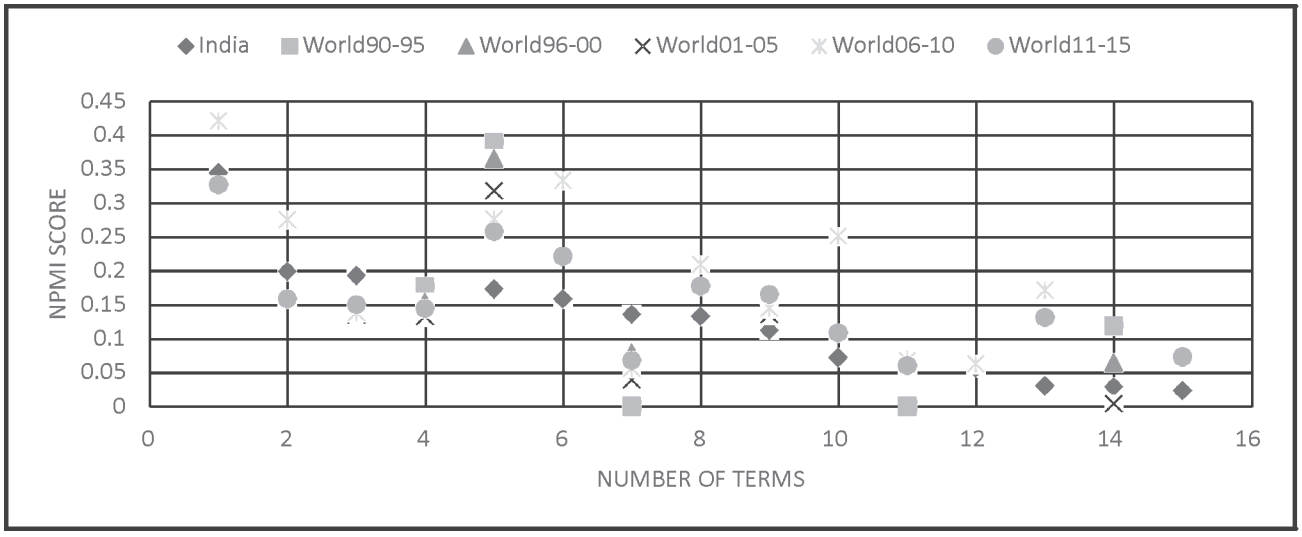

Figure 15: **Compound association with 15 different tool combinations 1.(Pharmacophore & V Screening), 2.(Docking & V Screening), 3.(SBDD & Target), 4.(Pharmacophore & Target), 5.(LBDD & Target), 6.(QSAR & V Screening), 7.(QSAR & Target), 8.(Pharmacophore & Docking), 9.(QSAR & Descriptor), 10.(LBDD & V Screening), 11.(Pharmacophore & Synthesis), 12.(Homolog & Docking), 13.(QSAR & Docking), 14.(LBDD & Assay) and 15.(pharmacophore & Scaffold hopping) are analysed and nPMI value is calculated. First LBDD & target is more preferable tool in world community while Indian community prefer QSAR target combination for novel compound selection. (Pharmacophore & docking) and (Docking & virtual screening) terms started coming after 2010 in world Publications.**

Applicability of the method

Robustness of any method is related to the applicability of that method to perform well in different databases. To check the applicability of the presented method, 3-Keyword (which belong to the medicinal chemistry related terms) is used to calculate unigram, bigram and trigram similar way as done with 2-Keyword (Bioinformatics related). As 3-Keyword set have ~30% common terms with the 2-Keyword set (Figure 2), it is expected that some unigram, bigram and trigram combinations will emerge similar to 2-Keyword (Table 1, 2 and 3 in Supplement 1). Table 3 shows the top 10 unigrams which have both medicinal chemistry terms and bioinformatics terms in common. Bigram calculation with tools (Table 4) further supports this, as 7 out of 15 tools in combination (bold in Table 4) are common with Figure 12. This result also vindicate that the coverage space of the data as well as the choice of keyword used for drug discovery related association is sufficiently large for this analysis.

Data limitation

A recent paper (Westergaard *et al.*, 2017) has shown that using only abstract limits the analysis. It is difficult to convert full text in a common readable format by Artificial Intelligent programs, which restricts the full-text analysis, but the referred paper has shown that for more associations can be traced between two or more terms using the full-text articles than the abstracts only. In future, an attempt will be made to use full text under the PubMed Central corpus using the XML file format for searching and association.

Conclusions

In the present study, a method has been established to find the importance of many *in silico* tools for interdisciplinary research and quantify the implication by text mining the abstracts of published papers. It has provided the role of Computational Biology/Bioinformatics tools in Translational research, mainly in compound design and comparison of Indian publications with the world. Data manifests that volume of publication is increasing faster in India since 2007, may implicate the rise in funding. However, most of the important application tools are similar to the world publications. Comparing preferred terms from Medicinal Chemistry and Computational Biology, it is found that preferred tools combination list is almost similar in both fields. Infectious disease is major application, addressed in Indian publications, nearly similar to world data which prefers Cancer as disease. It also brings out the importance of collaborative research in the translational biology and identifies the benefitting universities/institutes in India, hence scope for more collaboration across the country to accelerate the translational research. Though this is a limited database (only publication) study, it reflects the role of computational analysis of large data and application of appropriate mining tools for analysis of such text data to extract important knowledge for objectively guiding future decision.

Acknowledgement

We sincerely thank the facilities under Center of Excellence (CoE), Department of Biotechnology (DBT), for supporting the computational work. PK is supported by ICMR-SRF fellowship and GD is supported by CSIR-SRF fellowship. IG wishes to thank Department of Biotechnology CCPM project, Government of India for support.

Conflict of interest

None declared.

Abbreviations

nPMI : Normalized Pointwise Mutual Information; KW: Keyword

References

Altman, R. B. (2012). Translational Bioinformatics: Linking the Molecular World to the Clinical World. *Clin Pharmacol Ther*. 91, 994–1000.

Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. *Proc. Ger. Soc. Comput. Linguist. (GSCL 2009)*. pp 31–40.

Buckle, D. R., Erhardt, P. W., Ganellin, C. R., Kobayashi, T., Perun, T. J., Proudfoot, J. and Senn-Bilfinger, J. (2013). Glossary of Terms Used in Medicinal Chemistry Part II. *Pure Appl Chem*. 48, 387–418.

Chen, H. and Boutros, P. C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* 12, 35-41.

Chakraborti, A., Raina, D. and Sharma, K. (2016). Can an interdisciplinary field contribute to one of the parent disciplines from which it. *Eur Phys J Spec Top.* 225(17-18), 3127-3135.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. and De Hoon, M.J. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 25, 1422–1423.

Contopoulos-Ioannidis, D. G., Ntzani, E. E., and Ioannidis, J. P. A. (2003). Translation of highly promising basic science research into clinical applications. Am. J. Med. 114, 477–484.

Davies, S., Kennewell, P., Russell, A., Westwood, R. and Wynne, G. (2016). Drug Discovery Glossary, (January), 1–66. (*http://russell.chem.ox.ac.uk/resources/Drug_Discovery_Glossary.PDF*)

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A. and Pappas, G. (2007). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB J*. 22, 338–342.

Friedl, J. (2002). Mastering regular expressions. 3rd ed. (eds. A. Oram), O'Reilly Media, Inc.. Sebastopol, California, USA.

Ganellin, C. R., Lindberg, P. and Mitscher, L. A. (1998). Glossary of terms used in medicinal chemistry. *Pure Appl Chem*. 70, 1129-1143.

Hogeweg, P. (2011). The Roots of Bioinformatics in Theoretical Biology. *PLOS Comput. Biol.* 7, 1–5.

Kapetanovic, I. M. (2008). Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach. *Chem. Biol. Interact*. 171, 165–176.

Khan, M. O. F., Deimling, M. J. and Philip, A. (2011). Medicinal chemistry and the pharmacy curriculum.*Am J Pharm Educ*. 75, 161.

Krishnaswamy, S. and Mohan, T. M. (2016). The largest distributed network of bioinformatics centres in the world: Biotechnology Information System Network. *Curr. Sci.* 110, 556-561.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45(D1), D353–D361.

Navarro, G. (2001). NR-grep: A fast and flexible pattern-matching tool. *Softw Pract Exp*. 31, 1265–1312.

National Library of Medicine. (2004). Medline (*https://www.nlm.nih.gov/pubs/factsheets/medline.html*).

NCBI Resource Coordinators. (2017). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*45(Database issue), D12–D17.

Roberts, R. J. (2001). PubMed Central: The GenBank of the published literature. ý*Proc. Natl. Acad. Sci.* 98, 381–382.

Spedding, M. (2006). New directions for drug discovery. *Dialogues Clin Neurosci*. 8, 295–301.

Santos, S. de S., Takahashi, D. Y., Nakata, A. and Fujita, A. (2013). A comparative study of statistical methods used to identify dependencies between gene expression signals. *Brief Bioinform*. 15, 906–918.

Wehr, T. (2001). A Glossary of Drug Discovery Terminology. *LCGC N Am. J.* 19, 1066-1075.

WHO. (2010). ICD-10: International Statistical Classification of Diseases and Related Health Problems: tenth revision.

Westergaard, D., Stærfeldt, H. H., Tønsberg, C., Jensen, L. J. and Brunak, S. (2017). Text mining of 15 million full-text scientific articles. *bioRxiv*. 162099.