International Journal of Research in Library Science (IJRLS)

ISSN: 2455-104X

DOI: 10.26761/IJRLS.8.1.2022.1496

Volume 8, Issue 1 (January-March) 2022, 83-106, Paper ID: IJRLS-1496

Received: 12 January. 2022 ; Accepted: 25 January. 2022 ; Published: 31 January. 2022

A Bibliometric Study on Bioinformatics: An Analytical Study

Dr. M. Krishnappa[1]; Jaishree Khandelwal[2]

Librarian, Nrupathunga University (Formerly Government Science College) Nrupathunga Road, Bangalore, Karnataka[1]; Assistant Librarian, Central Power Research Institute (CPRI), Sadashiva Nagar Bangalore, Karnataka India[2]

mkrishnappa.m@gmail.com, jaishree@drtc.isibang.ac.in

ABSTRACT

Bibliometrics is a relatively new science. It started out as a statistical tool for analysing bibliographic data made necessary by the large increase in the number of journals and scientific papers. Only lately, due to a change of perspective, bibliometrics has become a common tool for the quantitative evaluation of scientific research. Bibliometrics is a systematic method for identifying the research trends in literature of various disciplines using quantitative measures. Bibliometric studies incorporate studies of literature growth of particular subjects, how much literature is contributed by various individual researchers, groups, institutions or countries. It is also used as a technique which is utilized to identify the various scientific indicators such as productivity of a researcher, research output, connections between publications, authors, and areas of research and citation data.

KEYWORDS: Scientometrics, Bibliometrics, Bibliographic, Bibliometrix.

1. INTRODUCTION

Bibliometrics study is both interdisciplinary and multidisciplinary because it studies the different field in the discipline of library and information science and multidisciplinary because it adopts various tools and methods from various other disciplines like statistics for analysis of data, information technology for data processing etc. Bibliometric studies exploit the various trends and patterns in bibliographic data deposited in bibliographical databases such as Scopus, Web of Science, PubMed, Google Scholar etc., using statistical and mathematical methods.

In recent years, bibliometrics has become a primary tool for the evaluation of scientific research. Such an evaluation is commonly used to support promotion decisions, or to allocate grants. Researchers may move from being the subjects of evaluation to those doing it, and then back and forth between the two roles. However, while there is a large literature on research performance indicators, there is little consensus on the best methods of bibliometric measurements.

2. AN OVERVIEW OF BIOINFORMATICS

Bioinformatics is an interdisciplinary field and comparatively new area of science that has made a significant impact

within a short span of time. It is a combination of microbiology and computational technologies. It is concerned with developing software tools for interpreting the biological data and to find patterns in it. Bioinformatics attracts researchers from diverse array of fields incorporating biological science, computer science, information engineering, mathematics and statistics to analyse and interpret the biological data. Bioinformatics uses the computer programming to analyse the data in the field of biological science. Bioinformatics is relatively new and one of the most demanded fields among researchers due to the recent advancement in computational and informatics, which are the pillars of bioinformatics when it comes to analyse Big Biological data. The field has grown exponentially since 2000 and it becomes important to understand not only the literature of the field but also how there has been a shift in topics and contribution of various fields over the years. Because of the complex and broad nature of the field, bibliometric analysis is often considered as an appropriate technique to access the current knowledge structure of a particular field, identify the current research themes, and to find the crucial literature in that subject area. A bibliometric study aims to find the exponential growth of the research over the years and the contribution of other fields such as developing the databases, software tools to store and retrieve the big biological data, to interpret the biological data using the statistical and visualizing tools.

3. STATEMENT OF THE PROBLEM

The statement of the **"A Bibliometric Study on Bioinformatics: An Analytical Study**" The exponential growth in publications of Bioinformatics field led to generation of tremendous amounts of data. Bibliometric Analysis is one of the best way to find the current knowledge and research contribution by the authors. It is essential for the researchers to recognize the core journals and the contribution and involvement of other disciplines in the Bioinformatics domain.

Science mapping technique is becoming significant task for researchers of all scientific community. As the scientific literature continues to increase at an exponential rate and literature grows fragmentarily, the task of agglomerating knowledge becomes further more complex. The discovery of intellectual literature and the research-front of scientific area are essential not only for scholars but also for making policies and its proper execution.

4. OBJECTIVES OF THE STUDY

The main objective of the study is to evaluate its current research performance and landscape, so as to facilitate potential interdisciplinary collaboration in the future. The study includes the following objectives.

1. To investigate the current trend in bioinformatics field,
2. To identifying the most productive and highly cited authors etc.,
3. To Find out historical evolution and elucidate its future direction.
4. To Evaluate innovations across the fields of genomics, computational biology, and bio-imaging.
5. To evaluate its current research performance and also to provide potential collaboration to multidisciplinary field in the future.
6. To analysis subject area showing greatest interest in the field of Bioinformatics.

5. SCOPE

The study would highlight the latest research trends in the topics, involvement of other disciplines in the field of Bioinformatics. The scope of the study is limited to the data collected from Scopus database using the search phrase "Bioinformatics and computational biology". Time series analysis from the period 2014-2018 is to be conducted to examine the interdisciplinary fields of Bioinformatics. The language chosen for conducting the analysis is limited to

only English language. The descriptive analysis of the topic is to be conducted for better understanding of the bioinformatics field. The bibliometric tool used in the study is Bibliometrix R package for analysing the data. It is an open-source tool for science mapping technique to analyse the scientific literature. The Bibliometrix package is flexible to integrate with other statistical tools and packages.

6 RESEARCH METHODOLOGY: SAMPLE, TESTS/TOOLS, STATISTICS TO BE USED

Looking at the objectives and nature of the study, to perform bibliometric analysis the bibliographic data is extracted from Scopus database using the search phrase "Bioinformatics" AND "Computational Biology". The search phrase is extracted which is matching the text in the article, title, keywords and abstract. To extract the data from the Scopus database the query is framed using the Boolean operator AND The query is the combination of search terms linked by Boolean operator. The query defines the search strategy by using the search fields in the title, keyword, abstract and full text articles, journals, affiliations, authors etc.

The bibliometric tool used in the study is Bibliometrix R package. Bibliometrix is an open-source tool for performing comprehensive science mapping analysis. Bibliometrix supports a recommended workflow to perform bibliometric analyses. It provides a set of tools for quantitative research for bibliometric analyses.

Bibliometrix package provides a combination of various tools at one place for quantitative research in bibliometric and scientometrics. Bibliometrics turns the main tool of science, quantitative analysis, on itself. Essentially, bibliometric is the application of quantitative analysis and statistics for literature like journals, articles followed by their citation counts. Not only qualitative but also quantitative analysis of publication and citation data is now used in nearly all scientific areas to evaluate growth, maturity, leading authors, conceptual and intellectual maps, trends of a scientific community.

To analyse bibliographic data, Bibliometrix Package 'R' is used.

I. The data is imported and converted to R format
II. Bibliometric analysis of the dataset can be done
III. Use of bibliometric laws to understand the objectives of the study
IV. Visualizing and analysing the co-citation, bibliographic coupling, collaboration, and co-word analysis performing network analysis, multiple correspondence analysis, and other data reduction techniques.

6.1 Thomson Reuters' Web of Knowledge

The Web of Science (WoS) is the oldest subscription-based citation index for more than 250 disciplines, and is provided by Thomson Reuters. It includes more than 12,000 journals and 150,000 conference proceedings. It is the most prestigious database, and the world's top academic and research institutions have strongly encouraged publications in WoS-indexed journals, influencing institutions' research productivity indicators and their place in global ranking systems such as Times Higher Education World University Ranking. More than 5,600 academic institutions in over 100 countries are now subscribing to WoS and other services that are available through the platform of knowledge.

6.2 SciVerse SCOPUS

SciVerse Scopus is a relatively new subscription database of abstracts and citations, which was launched in 2004 as a service of Elsevier. It is the most comprehensive and well-organised database, with more than 19,500 peer-reviewed journals across various disciplines being indexed. The coverage also includes conference proceedings,

patents, book series, and scholarly web pages. The latter is facilitated by Scirus, a search engine owned by Elsevier. Importantly, all MEDLINE-indexed journals are subject to coverage by Scopus.

6.3 Google Scholar

Google Scholar is a multidisciplinary search engine which shares common features with other search engines such as Elsevier's Scirus and bibliographic databases such as WoS and Scopus (Table 1). It was launched in 2004 by Google as a free web-based search engine. Over the past few years, it has substantially expanded its indexing of full texts of journal articles and books due to the agreements with Elsevier and other big and small publishers, online libraries, and repositories (for example, Ingenta Connect). The search engine also covers patents, conference proceedings, theses, presentations, web-pages, newspapers, and other non-peer-reviewed sources.

Google Scholar has gained a place in basic or back-up literature search algorithms for its comprehensive coverage of information across multiple disciplines, publishing formats and languages, as well for its simplistic approach to literature searches. Searches through Google Scholar are not linked to an organised vocabulary of scholarly keywords, and therefore do not require expert searching skills. The indexed sources, including web pages, are tagged with web-based keywords, which are found in the titles, abstracts, or full-texts of journal, book, and website articles.

Table 1: Main features of Web of Science, Scopus and Google Scholar:

	Web of Science	Scopus	Google Scholar
Owner	Thomson Reuters	Elsevier	Google
Year of launching	1960	2004	2004
Subscription	Yes	Yes	No
Covered items	Peer-reviewed sources, conference proceedings, book series	Peer-reviewed sources, patents, conference proceedings, book series, articles in press	Preprint articles (eg from arXiv [physics], journal articles, books, patents, conference proceedings, theses, presentations, web-pages, non-peer-reviewed sources
Coverage timespan	1900-present (Science), 1956-present (Social Science), 1975-present (Arts & Humanities)	1996-present (for most journals), 1966-present (for some journals)	Uncertain
Citation tracking	1900-present	1996-present	Inconsistent
Citations per article	Retrieves less citations than Scopus or Google Scholar	Retrieves more citations than WoS, particularly from non-English sources	Retrieves more citations than WoS or Scopus
Number of indexed peer-reviewed journals	> 12,000	~19,500	Uncertain
Journal impact indicators	Available	Available	Not available
Individual performance indicators (the *h* index)	Available	Available	Available
Main feature/motto	Selective coverage of journals based on the scientific quality and impact	Comprehensive coverage of journals	"Stand on the shoulders of giants"

6.4 Biomedical Databases

6.4.1 PubMed/MEDLINE

PubMed is a freely accessible search platform of the US NLM at the National Institutes of Health, which was first released in 1996. It employs the Entrez search engine, which interlinks all the databases of the National Centre for Biotechnology Information (NCBI) at the NLM, including PubMed, PubMed Central, and MEDLINE. PubMed is the largest and most well-organised abstract database, which is often accessed by biomedical and other specialists. As of 24 March 2013, it contains over 22.6 million records of journal articles and books indexed by MEDLINE, Index Medicus, and PubMed, going back to 1966 and selectively to 1809. Some of the old journals have full citation

records in this database. For example, over 171,130 articles of the *BMJ* are indexed from the first issue in 1857, with over 155,900 items being linked to the related full-text articles in PubMed Central. With over 162,700 indexed items, complete PubMed coverage has also been achieved for the top journal *Science*. PubMed is also linked to the NCBI Bookshelf, which is an increasingly popular database of selected online books in the life and health sciences.

Rapid updates, ease of access, diverse functionality, and retrieval of relevant information make PubMed the primary biomedical search platform. Although individual and journal impact factors are not calculated by PubMed, it is still widely searched by editors and publishers looking for editorial team members and reviewers with current and most relevant research activity. Searches through PubMed also form the basis for systematic literature reviews

6.4.2 EMBASE

EMBASE is the largest subscription-based biomedical and pharmacological abstracts database. EMBASE, an Elsevier product, contains over 25 million records from 1947 to the present. It indexes over 7,600 journals. Similar to Scopus, EMBASE covers all items indexed by MEDLINE. However, EMBASE contains 5 million more records than MEDLINE, including many European and non-English sources. The distinctive features of EMBASE are its focus on drug-related sources and reliance on the EMTREE thesaurus, an Elsevier product which lists over 56,000 drug and medical terms for EMBASE and EM Biology (a specialised database launched by Elsevier in 2005).

7. DATA ANALYSIS AND INTERPRETATION

This chapter presents the analysis and interpretation of the data collected for the study. To perform the bibliometric analysis, the bibliographic data is collected from the Scopus database. Scopus is an abstract and citation database covering large amount of peer reviewed journals in top level subjects like life-science, social science, health science, physical science.

To extract the data from the Scopus database the query is framed using the Boolean operator AND. The query is the combination of search terms linked by Boolean operator. The query defines the search strategy by using the search fields in the title, keyword, abstract and full text articles, journals, affiliations, authors etc.

The bibliographic data is extracted from Scopus database using the search phrase “Bioinformatics” AND “Computational Biology”. The search phrase is extracted which is matching the text in the article, title, keywords and abstract. The document type is limited to article. The number of articles extracted is 3872. The time span of the data is from the year 2014-2018. The language of the articles is limited to English, as much of the scientific publications are written in English. The data is exported from the Scopus database in the CSV, Bibtex format and plain text.

The bibliometric tool used is the Bibliometrix R package. Bibliometrix is an open source tool for performing comprehensive science mapping analysis. Bibliometrix supports a recommended workflow to perform bibliometric analyses. It provides a set of tools for quantitative research for bibliometric analyses. As it is programmed in R language, this tool is versatile and is easily upgraded and combined with other statistical R-packages. It is therefore

The bibliometrix R-package provides a set of tools for quantitative research in bibliometrics and scientometrics. It is written in the R language, which is an open-source environment and ecosystem. The existence of substantial, effective statistical algorithms, access to high-quality numerical routines, and integrated data visualization tools are perhaps the strongest qualities to prefer R to other languages for scientific computation.

The following are the steps followed in the Bibliometrix work flow process:

1. **Data collection. bibliometrix supports the following sub-stage:**
   - **a.** Data loading and conversion to R data frame
2. **Data Analysis, articulated in three sub-stages:**
   - a. Descriptive analysis of a bibliographic data frame
   - b. Network creation for bibliographic coupling, co-citation, collaboration, and co-occurrence analyses
   - c. Normalization
3. **Data visualization:**
   - a. Conceptual structure mapping
   - b. Network mapping
   - c. To describe the main functions of Bibliometrix, the bibliometric functions are defined in Table: 1, we analysed articles on Bioinformatics from 2014 to 2018. The data is collected from Scopus database.

a. Data loading and converting to R data frame

Data collection is a task composed of different subtasks as follows:

- **Data retrieval:**

Bibliometrix functions with data extracted from the bibliographic databases, Scopus. Bibliometrix connects with the Scopus API to automatically collect metadata regarding the complete scientific production The following are the searching criteria used for data extraction from Scopus data base:

1. The generic keyword "BIOINFORMATICS" AND "COMPUTATATIONAL BIOLOGY" as the topic, only articles written in English for the document type
2. "Biology", "Computer science" and "Statistics" and as subject categories, and
3. The timespan 2014–2018.

- **Data loading and converting:**

(hereinafter, square brackets denote the R syntax for commands). The export files are read by R using the read Files function [D<-read Files that creates a large character object called D. The function supports plain text (for Clarivate Analytics database) and BibTex (for both Clarivate Analytics and Scopus databases) formats and allows importing simultaneously multiple export files. These can be converted into a data frame using the convert2df function [M <-convert2df(D, dbsource = "scopus", format = "plaintext")]. convert2df creates a bibliographic data frame with cases corresponding to documents and variables to field tags in the original export file. Each document contains several elements such as authors' names, title, keywords and other information. These elements constitute the bibliographic attributes of a document, also called the metadata. We have chosen to use standard column names for the bibliographic data frame adopting the field tags proposed for Scopus collections. This facilitates merging different sources and applying R routines. Table 2 contains the structure of the bibliometrix data frame considering the main field tags. The column "class" reports the data type of each data frame column.

- **Data cleaning.**

Bibliometrix does not have specific routines dedicated to data cleaning. It does include in its main functions (e.g., loading and converting, citation analysis) a set of cleaning rules such as:

(i) transform full text into uppercase,

(ii)remove non-alphanumeric characters,

(iii) remove punctuation symbols and extra spaces, and

(iv) truncate author's first and middle names to the initials.

Table 2: Main Bibliometric Functions

Software assisted work flow steps	**Bibliometric Function**	**Description**	**Output**
Data loading and converting	• readFiles() • Convert2df() • retrievalByAuthorID()	Loads a sequence of Scopus and Clarivate Analytics WoS export files into R Creates a bibliographic data frame Uses Scopus API search to obtain information regarding documents on a set of authors using Scopus ID	Bibliographic data frame
Descriptive bibliometric analysis	• biblioAnalysis() • Summary() and plot() • citations() • localCitations() • dominance() • Hindex() • lotka() • keywordGrowth() • keywordAssociation()	Returns an object of class bibliometrix Summarize the main results of the bibliometric analysis Identifies the most cited references or authors Identifies the most cited local authors Calculates the authors' dominance ranking Measures productivity and citation impact of a scholar Estimates Lotka's law coefficients for scientific productivity Calculates yearly cumulative occurrences of top keywords/ terms Associates authors' keywords to keywords plus	Tables of results
Document x Attribute Matrix creation	• metaTagExtraction() • termExtraction() • cocMatrix()	Extracts other field tags, different from the standard WoS/Scopus codify Extracts and stems terms from textual fields (abstract, title, author's keywords, and others) of a bibliographic data frame Computes a Document x Attribute matrix Computes a Document x Attribute matrix	

The above table describes the main bibliometric functions which are used in Bibliometrix package along with the assisted work flow and the output of the function**.**

Table 3: Bibliometric data frame structure

Field Tag	Frame	Description
UT	Character	Unique Article Identifier
AU	Character	Authors
TI	Character	Document Title
SO	Character	Publication Name (or Source)
JI	Character	ISO Source Abbreviation
DT	Character	Document Type
DE	Character	Authors' Keywords
ID	Character	Keywords associated by WoS or Scopus database
AB	Large Character	Abstract
CI	Character	Author Address
RP	Character	Reprint Address
CR	Large Character	Cited References
TC	Numeric	Times Cited
PY	Numeric	Year
SC	Character	Subject Category
DB	Character	Bibliographic Database

It above table describes the various field tag with its frame type.

8. BIBLIOMETRIC ANALYSIS

The first step is to perform a descriptive analysis of the bibliographic data frame. The biblio Analysis function calculates the main bibliometric measures using this syntax:

[results <- biblioAnalysis(M,sep =¨;¨)].

The biblio Analysis function returns an object of class "bibliometrix". An object of class "bibliometrix" is a list containing the following components, shown in Table 4.

Table 4: List of Components in Bibliometrix

List Elements	Descriptions
Articles	Total number of documents
Authors	Authors' frequency distribution
Authors Frac	Authors' frequency distribution (fractionalized
First authors	First author of each document
nAUperPaper	Number of authors per document
Appearances	Number of author appearances
nAuthors	Total number of authors
AuMultiAuthoredArt	Number of authors of multi-authored articles
Years	Publication year of each document
FirstAffiliation	Affiliation of the first author for each document

Affiliations	Frequency distribution of affiliations (of all co-authors for each document)
Aff frac	Fractionalized frequency distribution of affiliations (of all co-authors for each paper)
CO	Affiliation country of first author
Countries	Affiliation countries' frequency distribution
TotalCitation	Number of times each document has been cited
TCperYear	Yearly average number of times each document has been cited
Sources	Frequency distribution of the sources (journals, books, others)
DE	Frequency distribution of the authors' keywords
ID	Frequency distribution of keywords associated to the document by Clarivate Analytics Web of Science and Scopus databases

8.1 Functions Summary and plot

To summarize the main results of the bibliometric analysis, use the generic function *summary*. It displays main information about the bibliographic data frame and several tables, such as annual scientific production, top manuscripts per number of citations, most productive authors, most productive countries, total citation per country, most relevant sources (journals) and most relevant keywords.

S<- summary (object = results, k=10, pause=FALSE)

K= formatting value that indicates the number of rows of each table.

Pause= logical value (TRUE or FALSE) used to allow (or not) pause in screen scrolling.

Here we have chosen k=10, we decide to see fist 10 authors. First 10 source, etc.

8.2 Main information regarding the collection.

The below table describes the collection size in terms of number of documents, number of authors, number of sources, number of keywords, timespan, and average number of citations.

Table 5: Main information regarding the data collection

Description	**Results**
Documents	3872
Sources (Journals)	864
Keywords Plus (ID)	28648
Author's Keywords (DE)	8883
Period	2014 - 2018
Average citations per documents	11.82
Authors	19595
Author Appearances	27656
Authors of single-authored documents	86
Authors of multi-authored documents	19509
Single-authored documents	96
Documents per Author	0.198
Authors per Document	5.06

Co-Authors per Documents	7.14
Collaboration Index	5.17
Document types	Articles
ARTICLE	3872

8.3 Descriptive analysis

Table 6; Top 10–Most productive authors

Sl.No	Authors	Articles	Authors	Articles Fractionalized
1.	ZHANG Y	92	WANG Y	13.26
2.	WANG Y	91	ZHANG Y	12.85
3.	WANG J	80	CHRISTIE AE	11.7
4.	WANG X	79	WANG J	11.59
5.	LIU Y	77	LI Y	10.88
6.	LI Y	71	WANG X	10.69
7.	ZHANG X	62	ZHANG J	10.06
8.	ZHANG J	61	LIU Y	9.94
9.	LIU X	58	ZHANG X	9.4
10.	LI X	56	LIU X	7.84

Table 7: Top 10–Most cited papers

Sl. No	Paper	TC (Total Citation)	Total Citation Per Year
1.	KELLEY LA, 2015, NAT PROTOC	2368	592
2.	YOON SH, 2017, INT J SYST EVOL MICROBIOL	986	493
3.	LI W, 2015, NUCLEIC ACIDS RES	478	119.5
4.	GUPTA SK, 2014, ANTIMICROB AGENTS CHEMOTHER	314	62.8
5.	PATHAN M, 2015, PROTEOMICS	310	77.5
6.	NI Y, 2014, GASTROENTEROLOGY	294	58.8
7.	PABINGER S, 2014, BRIEF BIOINFORM	285	57
8.	JANDHYALA SM, 2015, WORLD J GASTROENTEROL	269	67.2
9.	QUINLAN AR, 2014, CURR PROTOC BIOINFORM	261	52.2
10.	MICALLEF L, 2014, PLOS ONE	231	46.2

Table 8 Top 10–Most productive countries (based on first author's affiliation)

Sl.No	Country	Articles	Freq	SCP	MCP	MCP_Ratio
1.	CHINA	692	0.3185	625	67	0.0968
2.	USA	445	0.2048	351	94	0.2112
3.	GERMANY	104	0.0479	59	45	0.4327
4.	UNITED KINGDOM	94	0.0433	54	40	0.4255

5.	ITALY	73	0.0336	45	28	0.3836
6.	CANADA	61	0.0281	31	30	0.4918
7.	INDIA	61	0.0281	52	9	0.1475
8.	SPAIN	51	0.0235	29	22	0.4314
9.	KOREA	50	0.023	38	12	0.24
10.	AUSTRALIA	44	0.0202	22	22	0.5

SCP: Single country publications *MCP: Multiple country publications*

Table 9: Top 10–Most frequent journals

Sl. No	Sources	Articles
1.	PLOS ONE	287
2.	BMC BIOINFORMATICS	221
3.	BMC GENOMICS	139
4.	MOLECULAR MEDICINE REPORTS	130
5.	IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS	104
6.	INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES	72
7.	PLOS COMPUTATIONAL BIOLOGY	57
8.	GENE	56
9.	JOURNAL OF PROTEOME RESEARCH	50
10.	BIOMED RESEARCH INTERNATIONAL	49

Table 10: Top 10–Most frequent keywords

Sl. No	Author Keywords (DE)	Articles	Keywords-Plus (ID)	Articles
1.	BIOINFORMATICS	739	COMPUTATIONAL BIOLOGY	3869
2.	DIFFERENTIALLY EXPRESSED GENES	120	BIOLOGY	3685
3.	COMPUTATIONAL BIOLOGY	115	BIOINFORMATICS	3539
4.	MICRORNA	110	HUMAN	3513
5.	BIOINFORMATICS ANALYSIS	109	ARTICLE	3240
6.	MICROARRAY	89	GENETICS	2554
7.	NEXT GENERATION SEQUENCING	88	HUMANS	2328
8.	PROTEOMICS	83	PROCEDURES	1938
9.	GENE EXPRESSION	72	GENE EXPRESSION REGULATION	1869
10.	GENOMICS	58	CONTROLLED STUDY	1865

Author Dominance Ranking: The function dominance calculates the authors dominance ranking.

DF <- dominance (results, k=10)

Table 11: Top 10 Author Dominance Ranking

Sl.No	Author	Dominance Factor	Total Articles	Single-Authored	Multi-Authored	First-Authored	Rank by Articles	Rank by DF
1	ZHANG Y	0.217391304	92	0	92	20	10	1
2	LI Y	0.183098592	71	0	71	13	5	2
3	ZHANG X	0.177419355	62	0	62	11	4	3
4	LIU X	0.172413793	58	0	58	10	2	4
5	LIU Y	0.142857143	77	0	77	11	6	5
6	WANG J	0.1375	80	0	80	11	8	6
7	ZHANG J	0.133333333	61	1	60	8	3	7
8	WANG Y	0.131868132	91	0	91	12	9	8
9	WANG X	0.126582278	79	0	79	10	7	9
10	LI X	0.035714286	56	0	56	2	1	10

Table 12: Authors Productivity Per year

Sl.No	Author	year	freq	Total Citation	Total Citation per Year
1	LI X	2014	11	262	43.66667
2	LI X	2015	12	112	22.4
3	LI X	2016	8	30	7.5
4	LI X	2017	19	174	58
5	LI X	2018	19	47	23.5
6	LI Y	2014	14	272	45.33333
7	LI Y	2015	8	119	23.8
8	LI Y	2016	22	248	62
9	LI Y	2017	21	152	50.66667
10	LI Y	2018	27	45	22.5
11	LIU X	2014	8	132	22
12	LIU X	2015	15	143	28.6
13	LIU X	2016	10	169	42.25
14	LIU X	2017	17	121	40.33333
15	LIU X	2018	15	42	21
16	LIU Y	2014	11	161	26.83333

17	LIU Y	2015	14	152	30.4
18	LIU Y	2016	19	208	52
19	LIU Y	2017	13	59	19.66667
20	LIU Y	2018	24	121	60.5
21	WANG J	2014	18	195	32.5
22	WANG J	2015	16	340	68
23	WANG J	2016	16	174	43.5
24	WANG J	2017	21	198	66
25	WANG J	2018	22	42	21
26	WANG X	2014	14	193	32.16667
27	WANG X	2015	15	182	36.4
28	WANG X	2016	20	227	56.75
29	WANG X	2017	15	105	35
30	WANG X	2018	26	43	21.5
31	WANG Y	2014	17	242	40.33333
32	WANG Y	2015	22	232	46.4
33	WANG Y	2016	13	142	35.5
34	WANG Y	2017	30	226	75.33333
35	WANG Y	2018	31	51	25.5
36	ZHANG J	2014	11	144	24
37	ZHANG J	2015	13	255	51
38	ZHANG J	2016	5	84	21
39	ZHANG J	2017	20	174	58
40	ZHANG J	2018	25	37	18.5
41	ZHANG X	2014	12	162	27
42	ZHANG X	2015	13	152	30.4
43	ZHANG X	2016	12	165	41.25
44	ZHANG X	2017	17	87	29
45	ZHANG X	2018	17	49	24.5
46	ZHANG Y	2014	19	401	66.83333
47	ZHANG Y	2015	14	216	43.2
48	ZHANG Y	2016	18	233	58.25
49	ZHANG Y	2017	30	293	97.66667
50	ZHANG Y	2018	30	61	30.5

8.4 Most Frequently cited Articles

The function citations generate the frequency table of the most cited references or the most cited first authors (of references)

Table 13: Most frequently cited articles

Sl. No	Article Name	V1
1.	LI, H., DURBIN, R., FAST AND ACCURATE SHORT READ ALIGNMENT WITH BURROWS-WHEELER TRANSFORM (2009) BIOINFORMATICS, 25, PP. 1754-1760	56
2.	BARTEL, D.P., MICRORNAS: GENOMICS, BIOGENESIS, MECHANISM, AND FUNCTION (2004) CELL, 116, PP. 281-297	54
3.	SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N.S., WANG, J.T., RAMAGE, D., AMIN, N., IDEKER, T., CYTOSCAPE: A SOFTWARE ENVIRONMENT FOR INTEGRATED MODELS OF BIOMOLECULAR INTERACTION NETWORKS (2003) GENOME RES, 13, PP. 2498-2504	53
4.	KANEHISA, M., GOTO, S., KEGG: KYOTO ENCYCLOPEDIA OF GENES AND GENOMES (2000) NUCLEIC ACIDS RES, 28, PP. 27-30	50
5.	HUANG DA, W., SHERMAN, B.T., LEMPICKI, R.A., SYSTEMATIC AND INTEGRATIVE ANALYSIS OF LARGE GENE LISTS USING DAVID BIOINFORMATICS RESOURCES (2009) NAT PROTOC, 4, PP. 44-57	49
6.	BARTEL, D.P., MICRORNAS: TARGET RECOGNITION AND REGULATORY FUNCTIONS (2009) CELL, 136, PP. 215-233	34
7.	LIVAK, K.J., SCHMITTGEN, T.D., ANALYSIS OF RELATIVE GENE EXPRESSION DATA USING REAL-TIME QUANTITATIVE PCR AND THE 2(-DELTA DELTA C(T)) METHOD (2001) METHODS, 25, PP. 402-408	34
8.	ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W., LIPMAN, D.J., BASIC LOCAL ALIGNMENT SEARCH TOOL (1990) J MOL BIOL, 215, PP. 403-410	33
9.	ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W., LIPMAN, D.J., BASIC LOCAL ALIGNMENT SEARCH TOOL (1990) J. MOL. BIOL., 215, PP. 403-410	31
10.	GAUTIER, L., COPE, L., BOLSTAD, B.M., IRIZARRY, R.A., AFFY-ANALYSIS OF AFFYMETRIX GENECHIP DATA AT THE PROBE LEVEL (2004) BIOINFORMATICS, 20, PP. 307-315	31

Table 14: Most cited Authors

Sl.No	Author Name	V1
1.	ZHANG Y	1330
2.	WANG J	1182
3.	WANG Y	1143
4.	LI Y	1021
5.	WANG X	931
6.	ZHANG J	906
7.	LI H	853
8.	LI J	796
9.	LIU Y	774
10.	ZHANG X	702

8. 5 Top Authors H Index

The h-index is an author level metric that attempts to measure both the productivity and citation impact of the publications of a scientist or scholar. The index is based on the set of of the scientists most cited papers and the number of citations that they have received in other publications.

Table 15: Top 10 authors H index

Sl.No	Author	h_index	g_index	m_index	TC	NP	PY_start
1	ZHANG Y	18	27	3	1014	91	2014
2	WANG Y	15	21	2.5	691	88	2014
3	WANG J	14	25	2.333333	818	79	2014
4	WANG X	15	22	2.5	719	78	2014
5	LIU Y	11	19	1.833333	555	75	2014
6	LI Y	13	19	2.166667	539	70	2014
7	ZHANG X	12	19	2	543	62	2014
8	ZHANG J	13	22	2.166667	585	60	2014
9	LIU X	13	21	2.166667	562	55	2014
10	LI X	10	17	1.666667	388	55	2014

8.6 Some basic plots

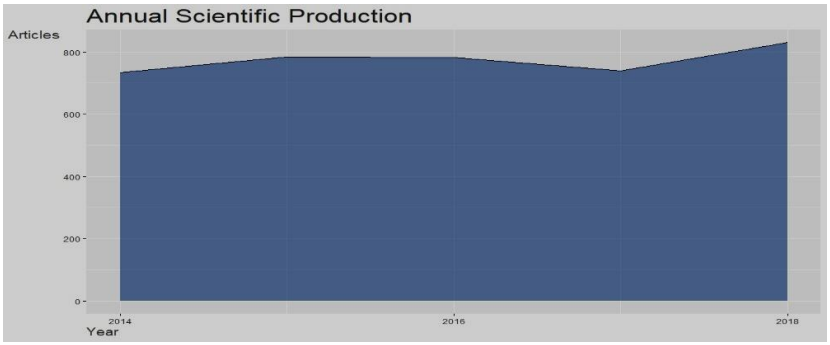

Figure 1: Annual Scientific Production

In fig. 1 Annual Scientific Production is shown i.e. no. of articles published over the years starting from 2014 to 2018. This basic plot gives an overall idea how the research in subject
is growing over the years as it is clearly shown in the plot that research in the bioinformatics
is steady with slight increment in the year 2018.

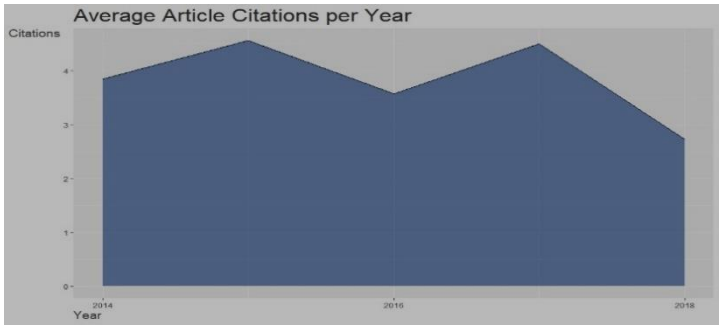

Figure 2: Average Article Citation Per Year

In fig. 2 Average article citation per year is shown i.e. average no. of citations for each article over the years starting from 2014 to 2018.

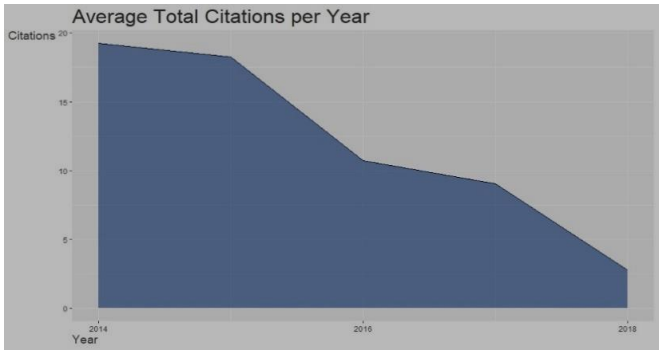

Figure 3: Average Total Citation Per Year

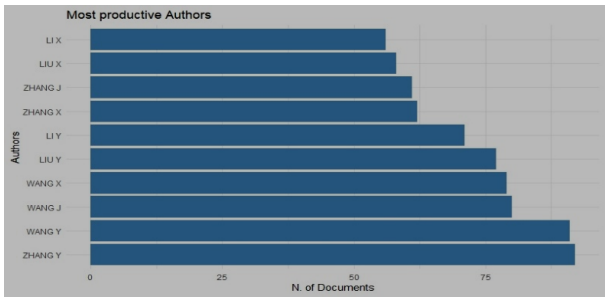

Figure 4: Most Productive Authors

In fig. 4 Most Productive Authors is shown i.e. total no. of documents published by an author from 2014 to 2018. This basic plot gives an overall idea that who is the most prominent author in this "**topic name** "within given time frame.

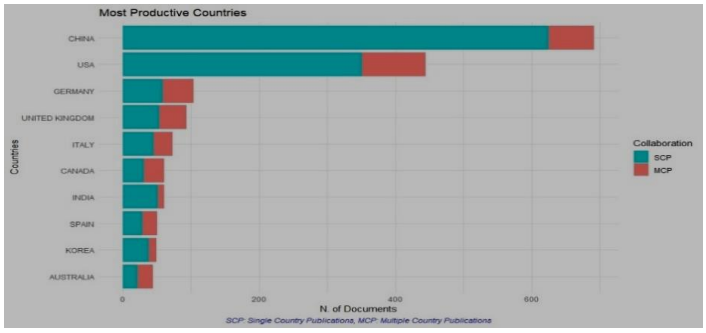

Figure 5: Most Productive Countries

In fig. 5 Most Productive Countries is shown i.e. total no. of documents contributed by an country from 2014 to 2018. This basic plot gives an overall idea that who is the most prominent country in this "**topic name** "within given time frame.

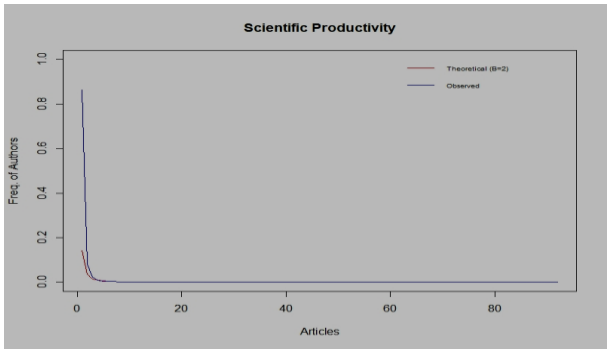

Figure 6: Scientific Productivity

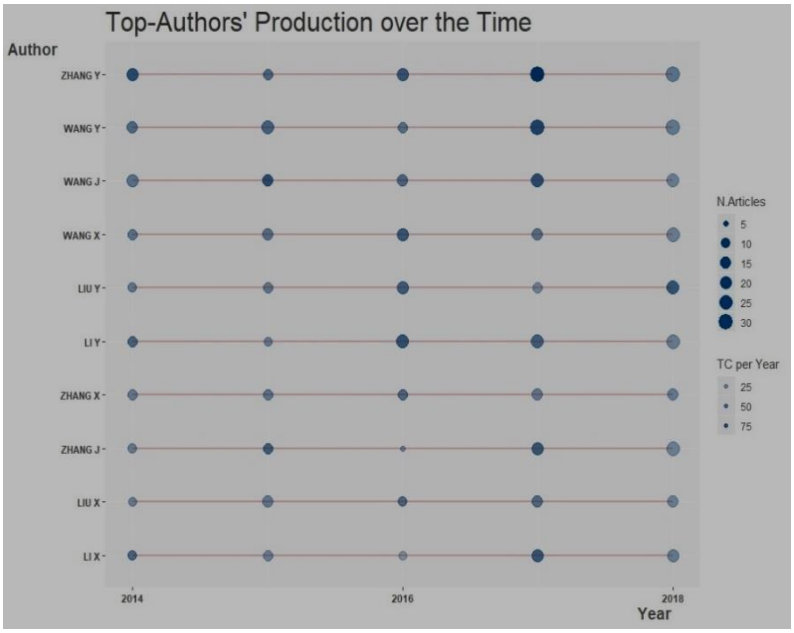

Figure 7: Top Authors Productivity

The function AuthorProdOverTime calculates and plots the authors production (in terms of number of publications, and total citations per year) over time.

8.7 Visualizing bibliographic networks

All bibliographic networks can be graphically visualized.

Here we visualize networks using *networkplot.* Using the networkplot we can plot the biblionetwork using R routines.

(i) Bibliographic network matrices

The various bibliographic networks have been described and we have shown that various bibliographic data can be transformed into the group of compatible networks. As we know that the manuscripts attributes are connected to each other through the document itself for e.g. Author(s) to journal, keywords to publication date, etc.

These connections of different attributes generate bipartite networks that can be represented as rectangular matrices (Manuscripts x Attributes). Furthermore, scientific publications regularly contain references to other scientific works. This generates a further network, namely, co-citation or coupling network. These networks are analysed in order to capture meaningful properties of the underlying research system, and to determine the influence of bibliometric units such as scholars and journals.

(ii) Bipartite networks

cocMatrix is a general function to compute a bipartite network selecting one of the metadata attributes.

For example, to create a network *Manuscript x Publication Source* we have to use the field tag "SO":

A <- cocMatrix(M, Field = "SO", sep = ";")

A is a rectangular binary matrix, representing a bipartite network where rows and columns are manuscripts and sources respectively.

Following this approach, we can compute several bipartite networks:

Citation network

A <- cocMatrix(M, Field = "CR", sep = ";")

Author network

A <- cocMatrix(M, Field = "AU", sep = ";")

Country network

Authors' Countries is not a standard attribute of the bibliographic data frame. We need to extract this information from affiliation attribute using the function *metaTagExtraction*.

M <- metaTagExtraction(M, Field = "AU_CO", sep = ";")

A <- cocMatrix(M, Field = "AU_CO", sep = ";")

metaTagExtraction allows to extract the following additional field tags: *Authors' countries* (Field = "AU_CO"); *First Author's countries* (Field = "AU_CO"); *First author of each cited reference* (Field = "CR_AU"); *Publication source of each cited reference* (Field = "CR_SO"); and *Authors' affiliations* (Field = "AU_UN").

Author keyword network

A <- cocMatrix(M, Field = "DE", sep = ";")

Keyword Plus network

A <- cocMatrix(M, Field = "ID", sep = ";")

(iii) Bibliographic coupling

Two articles are said to be bibliographically coupled if at least one cited source appears in the bibliographies or reference lists of both articles (Kessler, 1963).

A coupling network can be obtained using the general formulation:

$B = A\ X\ A^T$

where A is a bipartite network.

The function ***biblioNetwork*** calculates, starting from a bibliographic data frame, the most frequently used coupling networks: Authors, Sources, and Countries.

biblioNetwork uses two arguments to define the network to compute:

- analysis argument can be "co-citation", "coupling", "collaboration", or "co-occurrences".
- network argument can be "authors", "references", "sources", "countries", "universities", "keywords", "author_keywords", "titles" and "abstracts".

The following code calculates a classical article coupling network:

NetMatrix <- biblioNetwork(M, analysis = "coupling", network = "references", sep = ")

Articles with only a few references, therefore, would tend to be more weakly bibliographically coupled, if coupling strength is measured simply according to the number of references that articles contain in common.

This suggests that it might be more practical to switch to a relative measure of bibliographic coupling.

normalize Similarity function calculates Association strength, Inclusion, Jaccard or Salton similarity among vertices of a network. *Normalize Similarity* can be recalled directly from *network Plot()* function using the argument *normalize*.

NetMatrix <- biblioNetwork(M, analysis = "coupling", network = "authors", sep = ";")

net=networkPlot(NetMatrix, normalize = "salton", weighted=NULL, n = 100, Title = "Authors' Coupling")

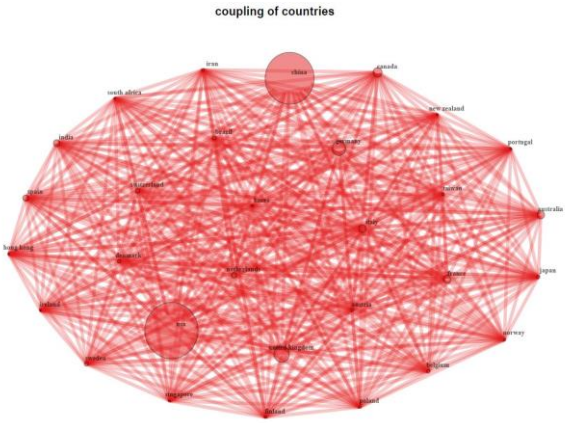

Figure 8: Coupling of Countries

In fig. 8 coupling of different countries is shown. We can clearly see that China and USA are most prominent countries when it comes to bioinformatics related research publication.

Bibliographic coupling of countries is extended from the bibliographic coupling concept and holds the view that two countries with more common references are more related and have more similar research interests.

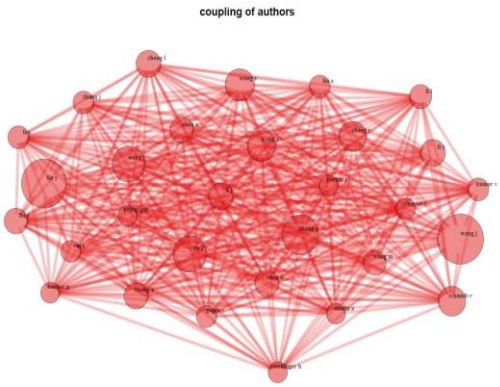

Figure 9: Coupling of Authors

In fig. 9 network for coupling of authors is shown. Bibliographic coupling of authors is extended from the bibliographic coupling concept and holds the view that two authors with more common references are more related and have more similar research interests.

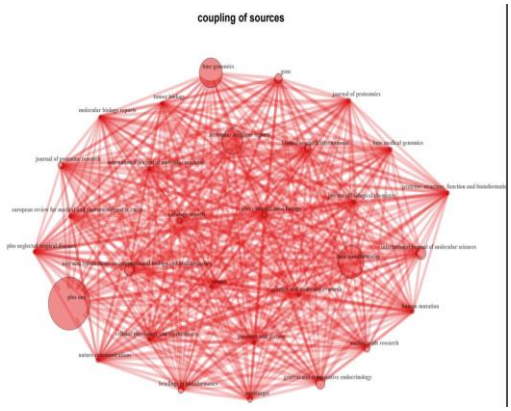

Figure 10: Coupling of Sources

In fig. 10 Coupling of sources network is shown. Bibliographic coupling of sources is extended from the bibliographic coupling concept and holds the view that two sources with more common references are more related and have more similar research interests.

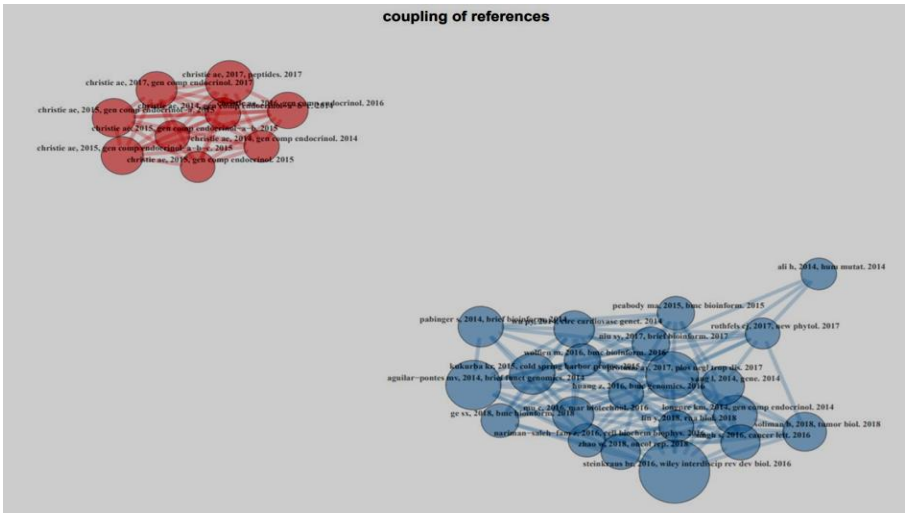

Figure 11 Coupling of References

In Fig. 11 Coupling of references network is shown. Bibliographic coupling of references is extended from the bibliographic coupling concept and holds the view that two articles with more common references are more related and have more similar research interests.

(iv) Bibliographic co-citation

We talk about co-citation of two articles when both are cited in a third article. Thus, co-citation can be seen as the counterpart of bibliographic coupling.

A co-citation network can be obtained using the general formulation:

$$C = A\,X\,A^T$$

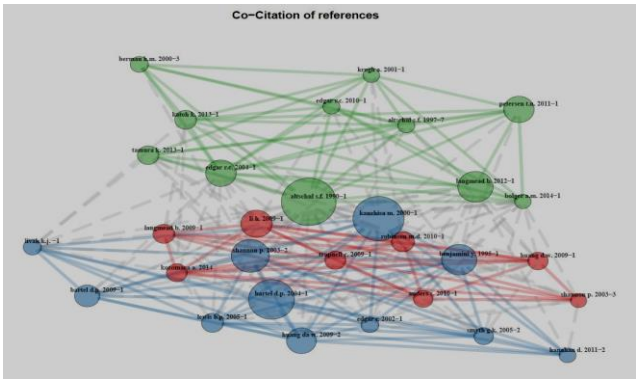

Figure 12: Co-citation of References

In fig. 12 Co-citation of references network is shown. It describes the co-citation between two articles if both of them are cited in third article.

(v) Co-occurrence's network: co-occurrence networks are the collective interconnection of terms based on their paired presence within a specified unit of text. Networks are generated by connecting pairs of terms using a set of criteria defining co-occurrence. For example, terms A and B may be said to "co-occur" if they both appear in a particular article. Another article may contain terms B and C. Linking A to B and B to C creates a co-occurrence network of these three terms. Rules to define co-occurrence within a text corpus can be set according to desired

criteria. For example, a more stringent criteria for co-occurrence may require a pair of terms to appear in the same sentence.

 # Create keyword co-occurrences network

NetMatrix <- biblioNetwork(M, analysis = "co-occurrences", network = "keywords", sep = ";")

 # Plot the network

net=networkPlot(NetMatrix, normalize="association", weighted=T, n = 30, Title = "Keyword Co-occurrences", type = "fruchterman", size=T,edgesize = 5,labelsize=0.7)

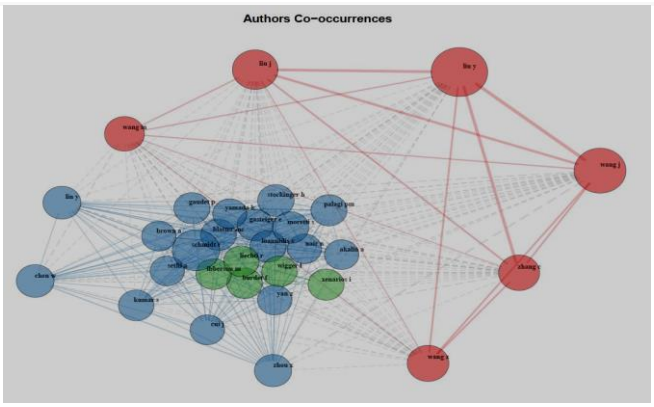

Figure 13: Authors Co-occurrences

In fig. 13 Authors co-occurrence network is shown in which we can associate which are all the authors that contributes together in an article and can find the interesting insights from the same.

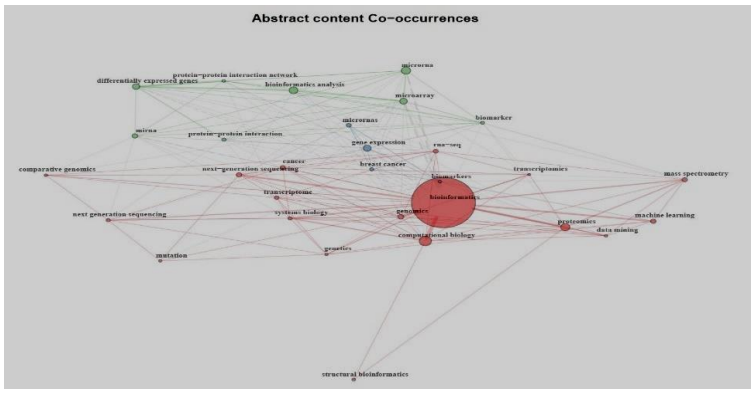

Figure 14: Abstract Content Co-occurrence

In fig. 14 Abstract content Co- occurrence network is shown which describes what are the words which are very common across the different articles abstract content. Bioinformatics is most common followed by computation biology, proteomics and so on.

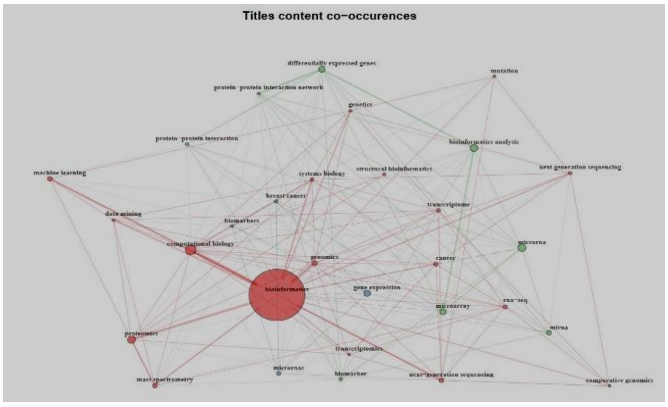

Figure 15: Title Content Co-occurrences

In Fig. 15 Title content co- occurrence is shown in which it is described that what are all the common key words associated across all the documents title content.

(vi) Bibliographic collaboration

Scientific collaboration network is a network where nodes are authors and links are co-authorships as the latter is one of the most well-documented forms of scientific collaboration (Glanzel, 2004).

An author collaboration network can be obtained using the general formulation:

$$AC = A\ X\ A^T$$

where A is a bipartite network *Manuscripts x Authors*.

Using the function *biblioNetwork*, you can calculate :

authors' collaboration network:

NetMatrix <- biblioNetwork(M, analysis = "collaboration", network = "authors", sep = ";")

country collaboration network:

NetMatrix <- biblioNetwork(M, analysis = "collaboration", network = "countries", sep = ";")

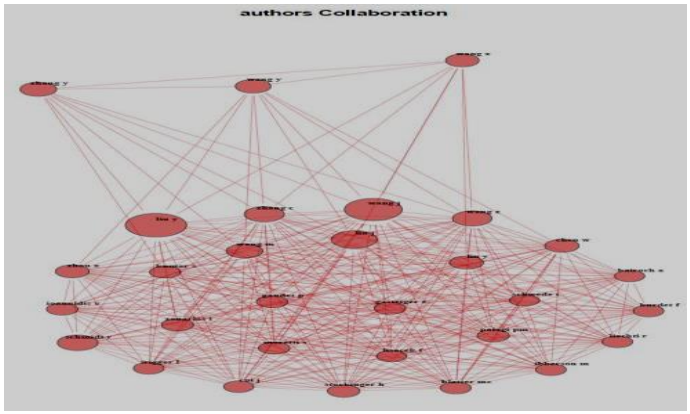

Figure 16: Authors Collaboration

In fig.16 Authors collaboration network is shown in which how the authors are collaborating with each other is explained, which further explains which two authors have similar research interest across the domain.

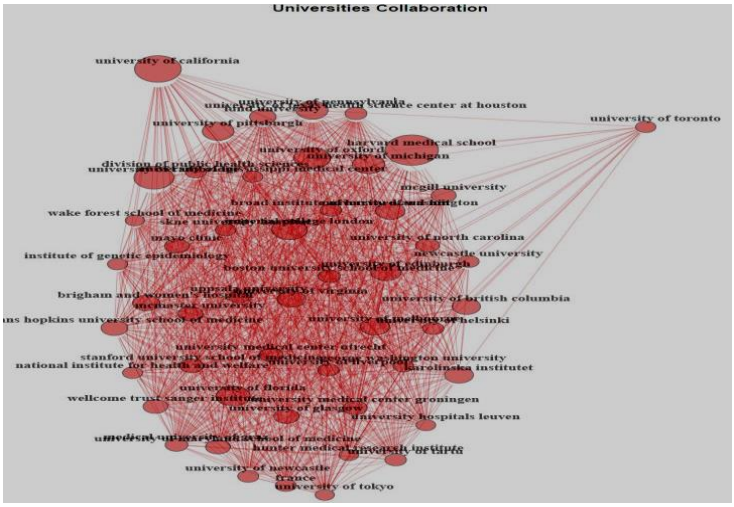

Figure 17: University Collaboration

In fig. 17 University Collaboration network is shown which explains the which are all the universities are collaborating with each other.

9. MAJOR FINDINGS

Main Information Table

- It is found that the total documents collected are 3872. The source (Journals) is 864 numbers. The average citations per documents is 11.82. The total number of authors in the documents collected form Scopus database are 19595.

From the Top 10 most productive authors

- ZHANG Y is in first place with 92 publications and LI X is at the last place with 56 publications.

Top 10 most cited papers (**Table 7**)

- KELLEY LA, 2015, NAT PROTOC is in first position having 2368 total citations and 592 total citations per year. The MICALLEF L, 2014, PLOS ONE is in 10th place with 231 total citations and 46.2 total citations per year.

Most productive countries (based on first author's affiliation) (Table 8)

- It is found that China is in first place with 692 articles having single country publications with 625 articles and 67 articles with multiple country publications. China stands first in single country publications.
- Australia is in 10th position with 44 articles published. The Single country publications are 22 and multiple country publications are also 22 articles.
- India is in last position in Multiple country publication with 9 articles.

Top 10–Most frequent journals. (Table 9)

- It is found that PLOS ONE is the most frequently articles are published in this journal. The total of 287 articles are published in the time span of 2014-2018. BIOMED RESEARCH INTERNATIONAL is at the 10th position having articles.

Top 10–Most frequent keywords (Table 10)

- It is found that 739 articles use author keyword (DE) BIOINFORMATICS as a most frequent keyword and 3869 articles uses COMPUTATIONAL BIOLOGY as a Keywords-Plus (ID) keyword.
- GENOMICS keyword is used in 58 articles as author keyword (DE) and CONTROLLED STUDY keyword is used in 1865 articles as a keyword plus (ID)

Author Dominance Ranking: (Table 11)

- In author dominance ranking ZHANG Y is in first place having 92 articles with dominance factor 0.217391304.
- LI X is in last position having 56 articles with dominance factor 0.035714286.

Most Frequently cited Articles: (Table 12)

- It is found that the most frequently cited article is **(** LI, H., DURBIN, R., FAST AND ACCURATE SHORT READ ALIGNMENT WITH BURROWS-WHEELER TRANSFORM (2009) BIOINFORMATICS, 25, PP. 1754-1760)

Most Cited Authors: (Table 14)

- It is found that ZHANG Y is the most cited author having 1330 citations.

Annual Scientific Production

- It is found that Annual Scientific Production is highest in the year 2018 with 850 articles published.

Average Article Citation Per Year:

- It is found that the Average article citation per year is highest in the year 2015 and 2017 with maximum average citations. And lowest in the year 2018.

Average Total Citation Per Year:

- It is found that the Average Total Citation Per Year is highest in the year 2014 and lowest in the year 2018.

Most Productive Authors:

- ZHANG Y is the most product author and LI X is the least productive author.

CONCLUSION

Bibliometric analysis is becoming an essential activity for scholars of all scientific disciplines. As the number of publications continues to expand at increasing rates and publications develop fragmentarily, the task of accumulating knowledge becomes more complicated. The determination of intellectual structure and the research-front of scientific domains are important not only for the research but also for the policy-making and practice.

The results of the study show that overtime the field of bioinformatics becomes further more multidisciplinary and also there is an invariable raise in peripheral fields like computational, mathematical, and system biology. The annual scientific productivity is increasing over the years. These results are confirmed by analysis of subject distribution and also by finding the core journals, articles, top ranked keywords and authors. By this it can be understood the historical evolution and future direction of the field. More over bioinformatics has stimulated many new innovations across the subfields of genomics, computational biology and system biology. Thus, there is a need to evaluate the current research performance and also to facilitate multidisciplinary collaboration in the future.

REFERENCE

[1] Aria, M., & Cuccurullo, C. (2017).bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics, 11(4), 959-975.

[2] Bansard, J. Y., Rebholz-Schuhmann, D., Cameron, G., Clark, D., Van Mulligen, E.,Beltrame, F., ... & Van Der Lei, J. (2007). Medical informatics and bioinformatics: a bibliometric study. IEEE Transactions on Information Technology inBiomedicine, 11(3), 237-243.

[3] Glänzel, W., Janssens, F., & Thijs, B. (2009). A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics. Scientometrics, 79(1), 109-129.

[4] Heo, G. E., Kang, K. Y., Song, M., & Lee, J. H. (2017). Analyzing the field of bioinformatics with the multi-faceted topic modeling technique. BMC bioinformatics, 18(Suppl 7), 251. doi:10.1186/s12859-017-1640-x

[5] Molatudi, M., Molotja, N., & Pouris, A. (2009). A bibliometric study of bioinformatics research in South Africa. Scientometrics, 81(1), 47-59.

[6] Patra, S. K., & Mishra, S. (2006). Bibliometric study of bioinformatics literature. Scientometrics, 67(3), 477-489.

[7] Song, M., Kim, S., Zhang, G., Ding, Y., & Chambers, T. (2014). Productivity and influence in bioinformatics: A bibliometric analysis using PubMed central. Journal of the Association for Information Science and Technology, 65(2), 352-371.

[8] Youngblood, M., & Amp; Lahti, D. (2018). A bibliometric analysis of the interdisciplinary field of cultural evolution. Palgrave Communications, 4(1), 120.