2011 International Conference on Information Communication and Management
IPCSIT vol.16 (2011) © (2011) IACSIT Press, Singapore

Bioinformatics Research – an Informetric View

A.Manoharan[1], B.Kanagavel[2], A.Muthuchidambaram[3], J.P.S.Kumaravel[4]

[1]Associate Professor, Bishop Heber College, Trichy. India. Bharathilib@yahoo.co.in

[2]Technical Officer, School of Biological Sciences, Madurai Kamaraj University, Madurai . India. kanagavelb@gmail.com

[3]Librarian, P.A.C. Ramasamy Raja Polytechnic College. Rajapalayam. muthuchidambaram.a@gmail.com

[4]Deputy Librarian, Madurai Kamaraj Univesity, Madurai 625021. India. jpskumar@yahoo.com

Abstract. Purpose – The purpose of this study is to conduct a scientometric analysis of the body of literature on Bioinformatics covered by Thompson's Web of Science database for a period from 2000 to 2010Design/methodology/approach – A total of 8729 articles were downloaded from Thompson's Web of Science database using the search term Bioinformatics subjected to scientometric data analysis techniques. Findings – A number of research questions pertaining to publication frequency, country, individual productivity and collaborative were proposed and answered. Based on the findings, many implications emerged that improve one's understanding of the identity of Bioinformatics as a distinct biomedical field. Research limitations/implications – The pool of articles are drawn from Thompson's Web of Science database only though there are other databases also.

Keywords: Bioinformatics, Scientometrics

1. Introduction

Biology in the 21st century is being transformed from a purely lab-based science to an information science as well. Biology may be viewed as the study of transmission of information: from mother cell to daughter cell, from one cell or tissue type to another, from one generation to the next, and from one species to another. This informational viewpoint is termed bioinformatics. Bioinformatics is the application of computer technology to the management of biological information. Computers are used to gather, store, analyze and integrate biological and genetic information which can then be applied to gene-based drug discovery and development. The science of bioinformatics has developed in the wake of methods to determine the sequences of the informational macromolecules - DNAs, RNAs and proteins. But in a wider sense, the biological world depends in its every process on the transmission of information, and hence bioinformatics is the fundamental core of biology. Bioinformatics is used for a vast array of important tasks like

- ❖ analysis of genome sequence data,
- ❖ analysis of gene variation and expression,
- ❖ analysis and prediction of gene
- ❖ protein structure and function,
- ❖ prediction and detection of gene regulation networks
- ❖ simulation environments for whole cell modelling,
- ❖ complex modelling of gene regulatory dynamics and networks
- ❖ presentation and analysis of molecular pathways in order to understand gene-disease interactions.
- ❖ designing primers (short oligonucleotide sequences needed for DNA amplification in polymerase chain reaction (PCR) experiments)
- ❖ predicting the function of gene products.

Bioinformatics is a discipline of science that analyses, seeks understanding and models the whole life as an information processing phenomenon over energy with methods from philosophy, mathematics and computer science using biological experimental data. There are several fundamental domains in bioinformatics. The ultimate material bioinformatists are working on is information. A biological function that bioinformatists try to define and analyze from such an information processing system is a relatively distinct layer of information processing over energy-flow in a relatively distinct time period. Many philosophical questions can rise to biology in the above scheme. How is the biological information quantified? Is there different quality of information in life? Is the sum of information in a cell equal to the sum of information in a protein? Is human being a higher quality or quantity information processing entity than a bacterium?

2. Literature Review and Research Questions

Molatudi et al reports on the practices of bioinformatics research in South Africa using bibliometric techniques. The search strategy was designed to cover the common concepts in biological data organization, retrieval and analysis; the development and application of tools and methodologies in biological computation; and related subjects in genomics and structural bioinformatics. The South African literature in bioinformatics has grown by 66.5% between 2001 and 2006. However, its share of world production is not on par with comparator countries, Brazil, India and Australia.

Swapan Kumar Patra and Saroj Mishra analysed the growth of the scientific literature in this area as available from NCBI PubMed using standard bibliometric techniques. Bradford's law of scattering was used to identify core journals and Lotka's law employed to analyze author's productivity pattern. Study also explored publication type, language and the Country of publication. Twenty core journals were identified and the primary mode of dissemination of information was through journal articles. Authors with single publication were more predominant (73.58%) contrary to that predicted by Lotka's law.

The present investigation contributes to overall trend of Bioinformatics research by analyzing the literature available in Thomson's Web of Science database by using various scientometric techniques. It proposes and answers six important research questions.

RQ1. What is the trend of research in the subject Bioinformatics?

RQ2. What is the trend of authorship pattern in the field of Bioinformatics?

RQ3. What is the productivity ranking of various countries in the field of Bioinformatics?

RQ4. What are the more productivity journals in Bioiformatics?

RQ5. What are the approaches of the ranking of the scholars in Bioinformatics? What are the differences in the individual research output calculated by (a) author position, (b) direct count and (c) equal credit methods?

Investigation of individual research productivity is perhaps the most frequent topic of scientometric projects (WrightandCohn,1996; Bapnaand Marsden,2002). The development of a list of key contributors in Bioinformatics may potentially help these scholars and/or practitioners gain reputation. In addition, doctoral students seeking potential supervisors and junior researchers looking for mentors need to know whom to approach.

A critical issue in determining individual faculty productivity involves assigning credit for multi-authored papers. There are many basic approaches to determining authorial credit:

1. author position; 2. direct count; and 3. equal credit (Chua *et al*., 2002; Lowry *et al*., 2007).

The author position method assigns values according to where the author is positioned in the citation (Howard and Day, 1995). Where two authors are listed, the first receives a score of 0.6, while the second receives a score of 0.4. A paper with four authors can generate the scores 0.415, 0.277, 0.185 and 0.123 for the authors in order of their position, in accordance with the formula of Howard et al. (1987). Similarly Dr.S.R.Ranganathan formulates an equal sharing method for a paper with two authors. Many collaborators, however, prefer to list authors in alphabetical order, which results in an unfair advantage to those with names higher in the alphabet. Individual productivity rankings obtained by this method may also lead to a conflict among co-authors who contributed equally to a manuscript. Therefore, this approach was also excluded to report productivity rankings in this study.

The direct count technique assigns a value of 1.0 for each author, regardless of the number of authors, but this approach is seen as having at least two major drawbacks. First, researchers who tend to work independently can potentially receive lower scores than researchers who tend to work collaboratively, since collaborative work can allow for a greater number of publications in any given measurement period. Second, this method inflates the ranking of those who tend to co-author a large number of papers with multiple authors while keeping their contribution to each paper marginal. Therefore, the direct count technique was not employed in the present investigation.

Scoring according to Ranganathan's Canon of Prepotency is believed to be less biased when compared to the other techniques. According to the Canon of Prepotence by Dr.S.R.Ranganathan, *the potency (power or strength) to decide the position of an entry among the various entries in a catalogue should, if possible, be concentrated totally in the leading section* . . . . Applying this canon to the position of authors in the list of authors for a specific publication, weightage can be given to the authors according to their position. If there are n authors for a publication, the weightage (w) of an author in pth postion ($p < n$) for that publication can be calculated as $W = (n – p +1)/ n!$

For example, in a publication by 5 authors, the weightage for authors in various (five) positions can be calculated as 1st Position = (5 -1 +1)/5! = 5/15; 2nd position = (5 -2 +1)/5! = 4/15; 3rd position = (5-3+1)/5! = 3/15

4th position = (5 -4 +1)/5! = 2/15 5th position = (5 -5 +1)/5! = 1/15

There is evidence to suggest that in some cases, the direct count, author position and equal credit methods may produce comparable results (Serenko et al., 2008).

RQ6. Does the frequency of publication by authors in the Bioinformatics field follow Lotka's law?

The research questions presented above concentrate on the distribution of productivity scores among a group of leading countries and individuals. In addition to this, it would be interesting to observe the overall productivity distribution patterns of all Bioinformatics authors. For this, Lotka's law (Lotka, 1926) has been frequently utilized in prior scientometric studies (Chung and Cox, 1990; Nath and Jackson, 1991; Rowlands, 2005; Kuperman, 2006; Cocosila et al., 2009). Lotka deduced a general equation for the relation between the frequency 'y' of persons making 'x' contributions as follows: $x^n y = \text{constant}$

The purpose of Lotka's law is to predict an approximate number of authors who contribute to the academic body of knowledge with a certain frequency of publications. It proposes that the number of individuals publishing a specific number of papers in a certain discipline is a fixed ratio to the number of scholars producing only a single work (Egghe, 2005). For example, within a particular timeframe, there may be one quarter as many authors with two publications as there are single-paper authors, one ninth as many with three, one sixteenth as many with four, etc.

2.1 Data and methods

Data were downloaded from Thomson's ISI Web of Science Database using the keyword Bioinformatics or Bioinformatic in Topic Search for a period from 2000 to 2010. The downloaded data is restricted to journal articles only by eliminating the other formats like editorial, letters, biographies etc. The data downloaded thus in the text format are converted into MS access database for analysis.

2.2 Limitations

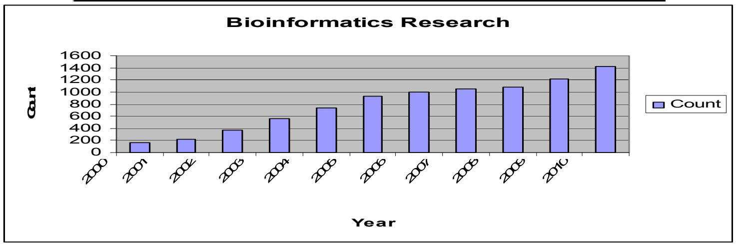
This investigation concentrates on research productivity in terms of the number of publications only with the downloaded data from Thomson's ISI Web of Science Database

3. Analysis and Interpretations

3.1 Overall trend

Table 1 Research productivity Trend

Year	Count	Growth Rate	Year	Count	Growth Rate
2000	163		2006	993	0.07
2001	214	0.31	2007	1056	0.06
2002	376	0.76	2008	1078	0.02
2003	556	0.48	2009	1213	0.13
2004	737	0.33	2010	1416	0.17
2005	927	0.26		8729	2.59



During the period of eleven years from 2000 to 2010 8729 articles were published. The research productivity in Bioinfomatics is on the increase though not uniform. At the start of the new Millennium, the growth is more while in 2010 the growth rate is less. The average growth rate is 0.235 showing that every year the research productivity in Bioinformtics grows by 0.235 per cent.

Table 2 Authorship Pattern

No of Authors	Count	Percent	No of Authors	Count	Percent	No of Authors	Count	Percent
1	727	8.33	18	30	0.34	35	3	0.03
2	1225	14.03	19	15	0.17	36	1	0.01
3	1310	15.01	20	12	0.14	39	1	0.01
4	1223	14.01	21	12	0.14	40	3	0.03
5	1002	11.48	22	4	0.05	41	1	0.01
6	870	9.97	23	8	0.09	42	1	0.01
7	621	7.11	24	7	0.08	43	1	0.01
8	448	5.13	25	8	0.09	44	2	0.02
9	361	4.14	26	6	0.07	45	1	0.01
10	235	2.69	27	3	0.03	48	4	0.05
11	165	1.89	28	4	0.05	50++	12	0.14
12	130	1.49	29	3	0.03		8729	100
13	86	0.99	30	1	0.01			
14	66	0.76	31	1	0.01			
15	48	0.55	32	4	0.05			
16	31	0.36	33	3	0.03			
17	29	0.33	34	1	0.01			

Single authorship papers are less in number than the multi authored papers. At the same time it is found that when the number of authors for a paper increases, the total publication count decreases. The maximum number publications is by three authors. Hence it can be understood that the optimum number of authors in the field of Bioinformatics is 3.

Table 3 Country of Publication

Country	Count	Percent	Country	Count	Percent	Country	Count	Percent
USA	2777	31.81	WALES	39	0.45	Bangladesh	2	0.02
China	842	9.65	Portugal	32	0.37	Cyprus	2	0.02
England	644	7.38	MEXICO	27	0.31	Iceland	2	0.02
GERMANY	516	5.91	NEW ZEALAND	25	0.29	Kuwait	2	0.02
Japan	412	4.72	THAILAND	25	0.29	Lebanon	2	0.02
Canada	297	3.4	Oman	22	0.25	NIGERIA	2	0.02
FRANCE	279	3.2	TURKEY	21	0.24	Uruguay	2	0.02
INDIA	237	2.72	COLOMBIA	19	0.22	Algeria	1	0.01
ITALY	230	2.63	CHILE	18	0.21	Bosnia	1	0.01
AUSTRALIA	199	2.28	CZECH REPUBLIC	17	0.19	Cameroon	1	0.01
SPAIN	184	2.11	ARGENTINA	15	0.17	Guadeloupe	1	0.01
TAIWAN	154	1.76	Hungary	15	0.17	Kazakhstan	1	0.01
BRAZIL	148	1.7	SLOVAKIA	12	0.14	Malta	1	0.01
SWEDEN	143	1.64	MALAYSIA	11	0.13	Philippines	1	0.01
SOUTH KOREA	128	1.47	Cuba	10	0.11	Sri Lanka	1	0.01
NETHERLANDS	123	1.41	CROATIA	9	0.1	U ARAB EMIRATES	1	0.01
ISRAEL	113	1.29	LITHUANIA	9	0.1	Uganda	1	0.01
SWITZERLAND	94	1.08	SLOVENIA	9	0.1	Ukraine	1	0.01
SINGAPORE	81	0.93	South Africa	9	0.1	VENEZUELA	1	0.01
SCOTLAND	76	0.87	Pakistan	7	0.08	Vietnam	1	0.01
DENMARK	74	0.85	SERBIA	7	0.08		8729	100
BELGIUM	72	0.82	Peru	6	0.07			
POLAND	72	0.82	SAUDI ARABIA	6	0.07			
GREECE	52	0.6	Tunisia	6	0.07			
IRELAND	51	0.58	Estonia	5	0.06			
FINLAND	50	0.57	JORDAN	5	0.06			
AUSTRIA	48	0.55	BULGARIA	4	0.05			
IRAN	44	0.5	Egypt	4	0.05			

Scholars from 80 countries have contributed 8729 research papers in Bioinformatics. USA has the maximum productivity with nearly one third of the total output. China has nearly 10 per cent of the total world output in the field Bioinformatics while India is in the eight position in the research productivity.

3.2 Ranked journals in bioinformatics

Table 4 Core Journals

Journal Name	Count
BIOINFORMATICS	436
BMC BIOINFORMATICS	436
NUCLEIC ACIDS RESEARCH	401
PROTEOMICS	175
JOURNAL OF PROTEOME RESEARCH	137
PLOS ONE	133
BMC GENOMICS	130
PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA	113
PROTEINS-STRUCTURE FUNCTION AND BIOINFORMATICS	109
INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE	96
JOURNAL OF BIOLOGICAL CHEMISTRY	87
BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS	86
INTERNATIONAL JOURNAL OF ONCOLOGY	79
BRIEFINGS IN BIOINFORMATICS	74
JOURNAL OF MOLECULAR BIOLOGY	71
INTERNATIONAL JOURNAL OF DATA MINING AND BIOINFORMATICS	63
GENE	55
PROGRESS IN BIOCHEMISTRY AND BIOPHYSICS	48
GENOME RESEARCH	46
METHODS OF INFORMATION IN MEDICINE	45
JOURNAL OF BIOMEDICAL INFORMATICS	44
BIOCHEMISTRY AND MOLECULAR BIOLOGY EDUCATION	42

The total number of journals that have conftibuted to Bioinformatics is 1860, of which the leading journal is Bioinformatics. The first three leading journals are from England. This shows that scholars from various countries prefer to publish their research findings in journals from England with respect to the subject Bioinformatics.

3.3 Ranked authors according to potency/position

The total number of authors who have contributed to Bioinformatics research during the period from 2000 to 2010 is 36667. The authors are ranked according to the weightage or credit given for their position.

Table 5 Ranked authors

Author	Count	Potency	Rank	Author	Count	Potency	Rank
Katoh, M	153	117.43	1	Maojo, V	22	4.29	6
Katoh, Y	34	22.67	2	Mohabatkar, Hassan	7	3.83	95
Wiwanitkit, Viroj	13	13	16	Lee, C	14	3.55	11
Chou, KC	32	12.68	3	Martens, Lennart	24	3.24	4
Cai, YD	23	9.43	5	Bujnicki, JM	12	3	23
Katoh, Masuko	12	8	19	Stevens, R	8	2.98	73
Kirikoshi, H	10	5.97	31	Canduri, F	16	2.93	8
Katoh, Masaru	15	5	9				

The first two ranked authors Katon M and Katoh Y are from Japan while the third ranked author Chou, KC who has maximum output is from China. Wiwanitkit, Viroj from Thailand has only 13 papers and all of them are the results of solo research. Hence he is placed in the third rank among the list of authors.

Table 6 Lotkas law of Author Productivity

No of Papers (X)	No of Authors(Y)	X^n*Y (n =3.2)	No of Papers (X)	No of Authors(Y)	X^n*Y (n =3.2)
1 Paper	31444	31444	6 Papers	92	28436.22
2 Papers	3595	33036.56	7 Papers	55	27840.46
3 Papers	906	30473.07	8 Papers	30	23281.41
4 Papers	329	27783.56	9 Papers	18	20363.32
5 Papers	160	27594.59	10 Papers	9	14264.04

The values in the last column are not constant disproving the Lotka's Law.

4. Discussion and Conclusions

The purpose of this project was to conduct a scientometric analysis of Bioinformtics research in order to understand the discipline's identity. For this, 8729 articles published in 1860 major peer-reviewed journals were analyzed. It is identified that there is a growth in the trend of research in Bioinformatics. During the project, 36667 unique were identified. Despite its relatively short history, Bioinformtics already boasts a continuously growing body of knowledge. The discipline has attracted the attention of a tremendous number of individual contributors from a variety of both academic and non-academic institutions. The number of very productive individuals were identified to be 36667. Among the 81 countries that have contributed to

Bioinformatics research USA takes the majour share. In terms of application of Lotkas law or author productivity, the data does not confine to the law.

Overall, there is a great danger that Bioinformatics may lose its practical side and become a pure scholarly discipline. Developing countries like India and China are generating the most research output. In this project, 81 contributing countries were identified among whom USA takes the lead. Nearly one third of all research was generated by the USA alone. This suggests that the production of scholarly Bioinformatics research is not distributed equally among the nations. Instead, a handful of countries accounts for the majority of publications. A related phenomenon, referred to as the Matthew effect for countries (Bonitz et al., 1997), has been observed in virtually all scientific fields. The Matthew effect, introduced in the seminal works by Merton (1968, 1988) refers to the situation when an initial advantage gained by an individual scholar, institution or country leads to further advantage, whereas their less fortunate counterparts receive little or no gain. It is likely that wealthy countries were able to initially invest heavily in research institutions, attract top faculty, and provide research grants. This in turn facilitates the production of more research in those select countries.

5. References

[1] BANSARD, J. Y et al. (2007), Medical informatics and bioinformatics: A bibliometric study, *Transactions on Information Technology and Biomedicine,* 11 (3) : 237–243.

[2] BENTON, D. (1996), Bioinformatics – principles and potential for a new multidisciplinary tool, *Trends in Biotechnology*, 14 : 261–276.

[3] Merton, R. K. (1968). The Matthew effect in science. *Science, 159*, 56–63.

[4] Achuthsankar S Nair Computational Biology & Bioinformatics – A gentle Overview, Communications of Computer Society of India, January 2007.