Visualization of a Scientific Community of Indian origin in the US: A case study of Bioinformatics and Genomics

APARNA BASU[1]
GRANT LEWISON[2]

[1] Institute of Genomics and Integrative Biology
254 Okhla Industrial Estate – Phase 3, New Delhi 110020, INDIA
[2]Evaluametrics Ltd
50 Marksbury Avenue, Kew, Richmond, Surrey, TW9 4JF

Abstract

We look at the publication output of researchers of Indian origin, currently publishing from the USA in the area of Bioinformatics and Genomics and use the bibliographic details to construct a typology of this community. We find that there is a large output, almost an order of magnitude more than the output from India in the same set of specialized journals in this area. There are several highly productive authors who have contributed 15-20 papers annually. The papers are published in standard journals, including 10-20 papers annually in *Nature* and *Science* and 40 or more papers in *PNAS* and *Bioinformatics*. Inter community co-authorship links suggest that there are some intense links between individuals and well developed groups around the more active researchers.

1. Introduction

India has a large scientific establishment and publishes on an average 13 thousand research papers in a year, according to the Science Citation Index. However, there has been extensive discussion of the fact that the national publication output has been stagnating. It is also known that a large number of Indian students leave for countries in the west every year, for better opportunities thereby depleting the country's scientific talent pool. This 'brain drain' has resulted in sifting policy options that could help retain scientific talent in the country or attract others to return after a few years in foreign countries. While there have been several studies that measure the extent of the brain drain, primarily indicating numbers that have left after graduating from premier institutions in India [1], there has been very little assessment of the actual scientific expertise and potential of the pool of researchers of Indian origin in the western countries. In the US, Stephan[2] has found a that significant contribution to science is made by migrant scientists of foreign origin. In this study, the first of it kind in India, we have looked at the publication output of scientists of Indian origin located in the US in the emerging areas of Bioinformatics and Genomics. These areas are of relatively recent origin and of importance to India's future scientific agenda[3]. Through a PubMed search, we have found more than 10,000 papers authored or co-authored by Indian scientists in the US in the last 5 years in this area with some overlaps with adjoining life science areas. Indian author names were identified with reference to a database of Indian names. While numbers in PubMed and SCI are not strictly comparable, the large number of papers retrieved indicates that the Indian diaspora does indeed need a closer look and its potential needs to be evaluated in more concrete terms. In this paper we have examined various characteristics of this community of scientists of ethnic origin through visualization techniques.

2. Data Search

We selected Pubmed as the database of choice as it is web based and freely available from NCBI, with good coverage in the medical and life sciences and excellent search options. We restricted our search to the last 5 years as Bioinformatics and Genomics are relatively new fields and several journals devoted exclusively to these areas have entered only in the last few years. Earlier, articles were being published in more general life science journals, and searching in them for bioinformatics or genomics content would have been a tedious exercise. To search for Indian names we started with a representative list of more than 5000 Indian names and surnames. Using this we searched for similar names publishing in journals with the keywords 'Bioinformatics' or 'Genomics' in the journal title and with affiliations in the USA (Table 1).

Table 1a: Journals scanned in Bioinformatics

BIOINFORMATICS
Bioinformatics (Oxford, England)
BMC bioinformatics [electronic resource]
Briefings in bioinformatics
Journal of bioinformatics and computational biology
Applied bioinformatics
Proceedings / IEEE Computer Society Bioinformatics Conference

Table 1b: Journals scanned in Genomics

GENOMICS
Am J. of Pharmacogenomics
Annual rev of Genomics and Human genet
BMC Genomics
Breifings Func Genomics and Proteomic
Cytogen Genomic Research
DNA Research
Functional and Integrative Genomics
Genome
Genome Biology
Genome Inf. Ser Workshop
Genome Research
Genomics
Genomics, Proteomics, Bioinformatics
J of Structural and Functional Genomics
Mammalian genomics
Mol Genet Genomics
Pharmacogenomics
Physiological genomics

3. Methodology

We were able to download a total of 10,000 papers authored or co-authored by a researcher of Indian origin. Some of them had to be removed manually as they were from different disciplines. From this set of papers we were able to identify a total of 700 unique Indian names. We selected names with four or more papers in our restricted journal set in the last 5 years as our starting set of researchers. New names encountered were incorporated into the names database and iteratively used for capturing more Indian names. For comparison, a set of papers in Bioinformatics and Genomics published from India were also downloaded from

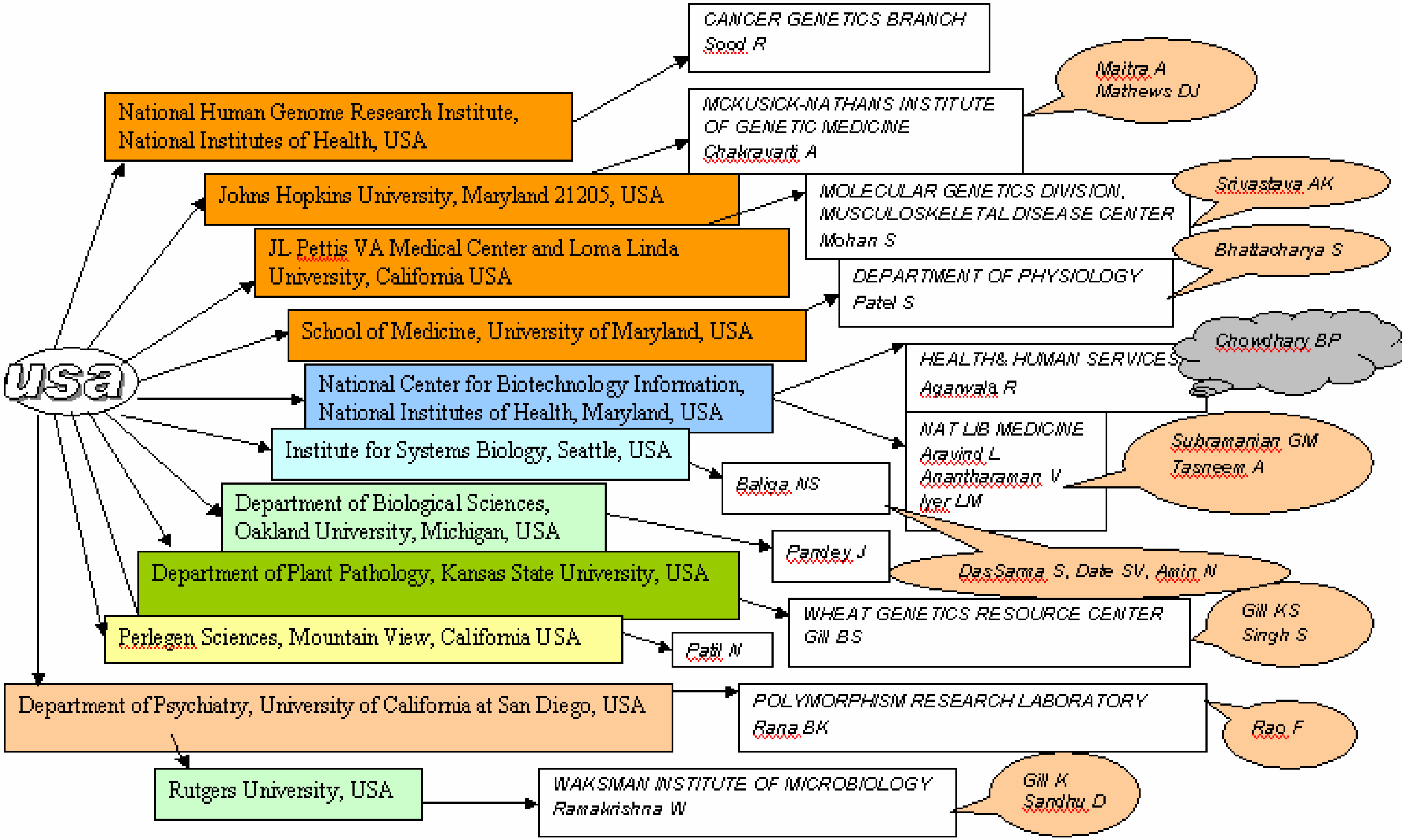
Pubmed. For the starting set we determined the location of the more productive individuals from the affiliating address. Some of the difficulties encountered were:

Use of surname and initial only. This was found to be problematic as there were sometimes 2-4 people with the same initials but different first names, for example, Patel S broke down to 4 individuals, Patel Sandeep, Satish, Satyakam and Shalaka. To take care of this problem, we used Full Author Name in subsequent searches covering other journals in Pubmed by the same authors. This feature was probably introduced in Pubmed after 2002. The data prior to 2002 therefore could remain ambiguous. Wherever possible, the ambiguity was resolved manually by looking at continuity with regard to location/affiliation, collaborators or research topic.

4. Results

4.1. Location and other features

For the starting set we determined where the more active researchers were located. This is shown in Figure 1.


usa
CANCER GENETICS BRANCH
Sood R
National Human Genome Research Institute, National Institutes of Health, USA
MCKUSICK-NATHANS INSTITUTE OF GENETIC MEDICINE
Chakravarti A
Maitra A
Mathews DJ
Johns Hopkins University, Maryland 21205, USA
MOLECULAR GENETICS DIVISION, MUSCULOSKELETAL DISEASE CENTER
Mohan S
Srivastava AK
JL Pettis VA Medical Center and Loma Linda University, California USA
Bhattacharya S
DEPARTMENT OF PHYSIOLOGY
Patel S
School of Medicine, University of Maryland, USA
Chowdhary BP
National Center for Biotechnology Information, National Institutes of Health, Maryland, USA
HEALTH& HUMAN SERVICES
Agarwala R
NAT LIB MEDICINE
Aravind L
Anantharaman V
Iyer LM
Subramanian GM
Tasneem A
Institute for Systems Biology, Seattle, USA
Baliga NS
Department of Biological Sciences, Oakland University, Michigan, USA
Pandey J
DasSarma S, Date SV, Amin N
Department of Plant Pathology, Kansas State University, USA
WHEAT GENETICS RESOURCE CENTER
Gill BS
Gill KS
Singh S
Perlegen Sciences, Mountain View, California USA
Patil N
Department of Psychiatry, University of California at San Diego, USA
POLYMORPHISM RESEARCH LABORATORY
Rana BK
Rao F
Rutgers University, USA
WAKSMAN INSTITUTE OF MICROBIOLOGY
Ramakrishna W
Gill K
Sandhu D

Researchers with 4 or more papers in Genomics journals 2001-Mar 2006 are shown along with their collaborators (colour call out)

Fig. 1 shows a fairly wide geographical distribution. The affiliating departments indicate that the research area covered includes both agricultural or plant genomics as well as animal genomics. There is one example of research collaboration between with an individual currently located outside the US. (BP Cowdhary)

4.2. Author Productivity

Distribution of author productivity was found to follow a Bradford shape but without the falling tail (Figure 2).

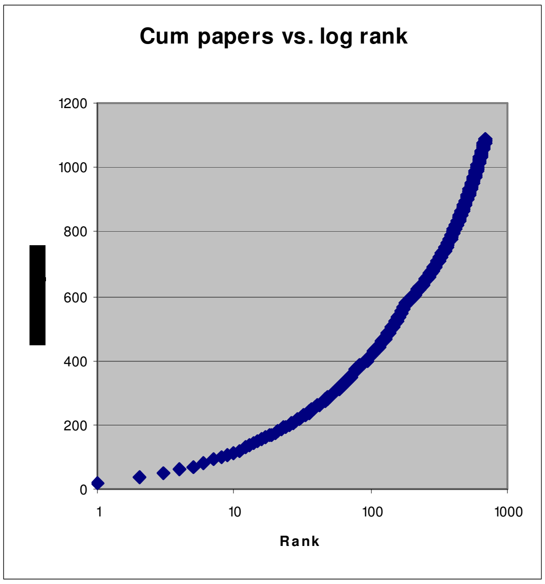


Table 2 shows the productivity of the authors in the starting set. From the date of the earliest paper we have determined the average annual productivity of each author.

Table 2: Average productivity of authors during their active years

NAMES	Papers	Starting Year	current year	years active	Average productivity
Aravind L	172	1997	2006	9	19.11
Mohan S	57	2002	2006	4	14.25
Pandey A	54	2002	2006	4	13.50
Moitra Anirban	142	1995	2006	11	12.91
Ghosh D	49	2002	2006	4	12.25

Chakravarti A	48	2002	2006	4	12.00
Iyer LM	28	2002	2006	4	7.00
Agarwala R	6	2005	2006	1	6.00
Jain AN	20	2002	2006	4	5.00
Anantharaman V	27	2000	2006	6	4.50
Ramanathan M	17	2002	2006	4	4.25
Kumar Anuj	12	2002	2006	4	3.00
Baliga NS	11	2002	2006	4	2.75
Mazumder R	13	2001	2006	5	2.60

We note the very high levels of productivity in the top ranked researchers in this area as compared to other areas in science. A graph of the starting year and productivity is shown in Figure 3.

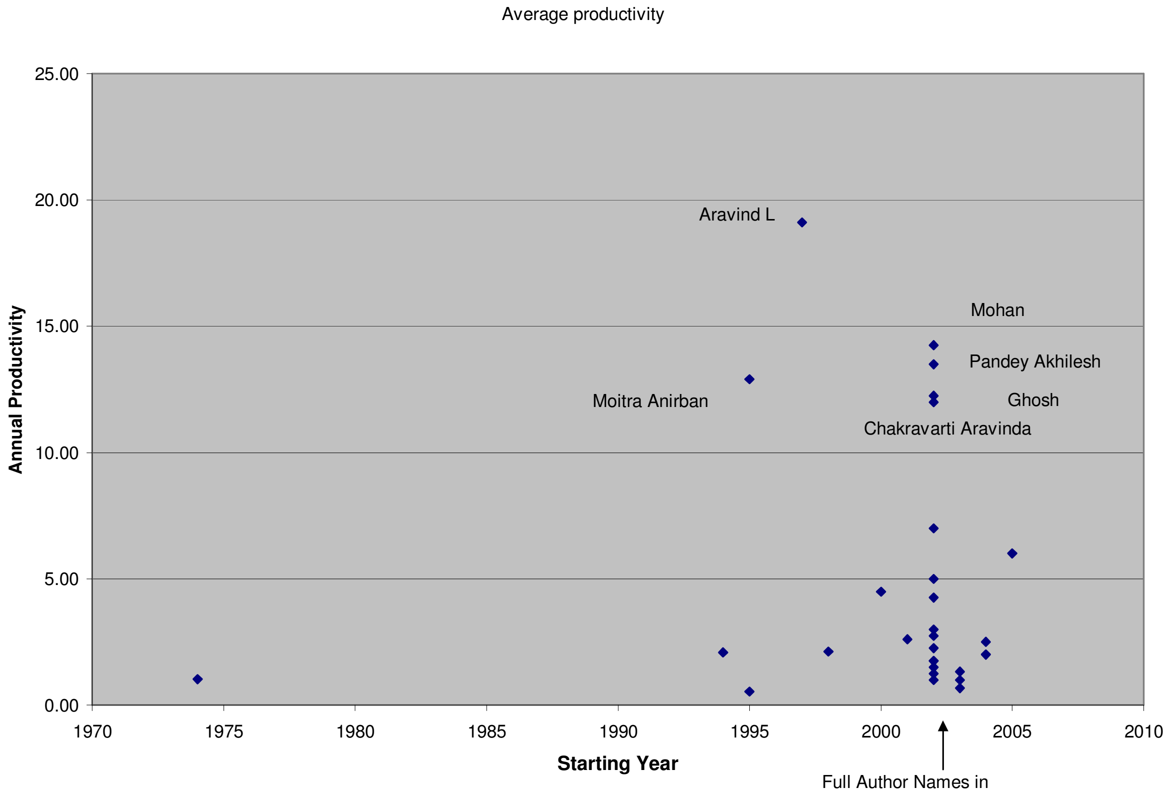


Fig. 3: Graph showing average productivity and the starting year of publication of researchers.(Note that the clustering of points at 2002 is an artifact due to Full Author Names

being introduced in Pubmed at that time. The year 2002 is like an upper bound of the starting year.)

Figure 3 shows a rapid rise in productivity for individuals starting in the post 1995. This is attributable to the new environment created by web based data in the biological sciences and the surge in bioinformatics. There are two discernible groups, one highly productive with average productivity greater than 10 papers annually (Average=14, Std. Dev=2.6) and the other group with less than 10 papers annually (Average=2.5, Std. Dev=1.7).

4.3. Journals used

Papers authored or co-authored by researchers of Indian origin were distributed in 1700 journals. Many of them had only a single paper in one of the 5 years 2001-05. The most frequently used journals are shown in Table 3.

Table 3: Journals used by Bioinformatics and Genomics researchers of Indian origin in the US

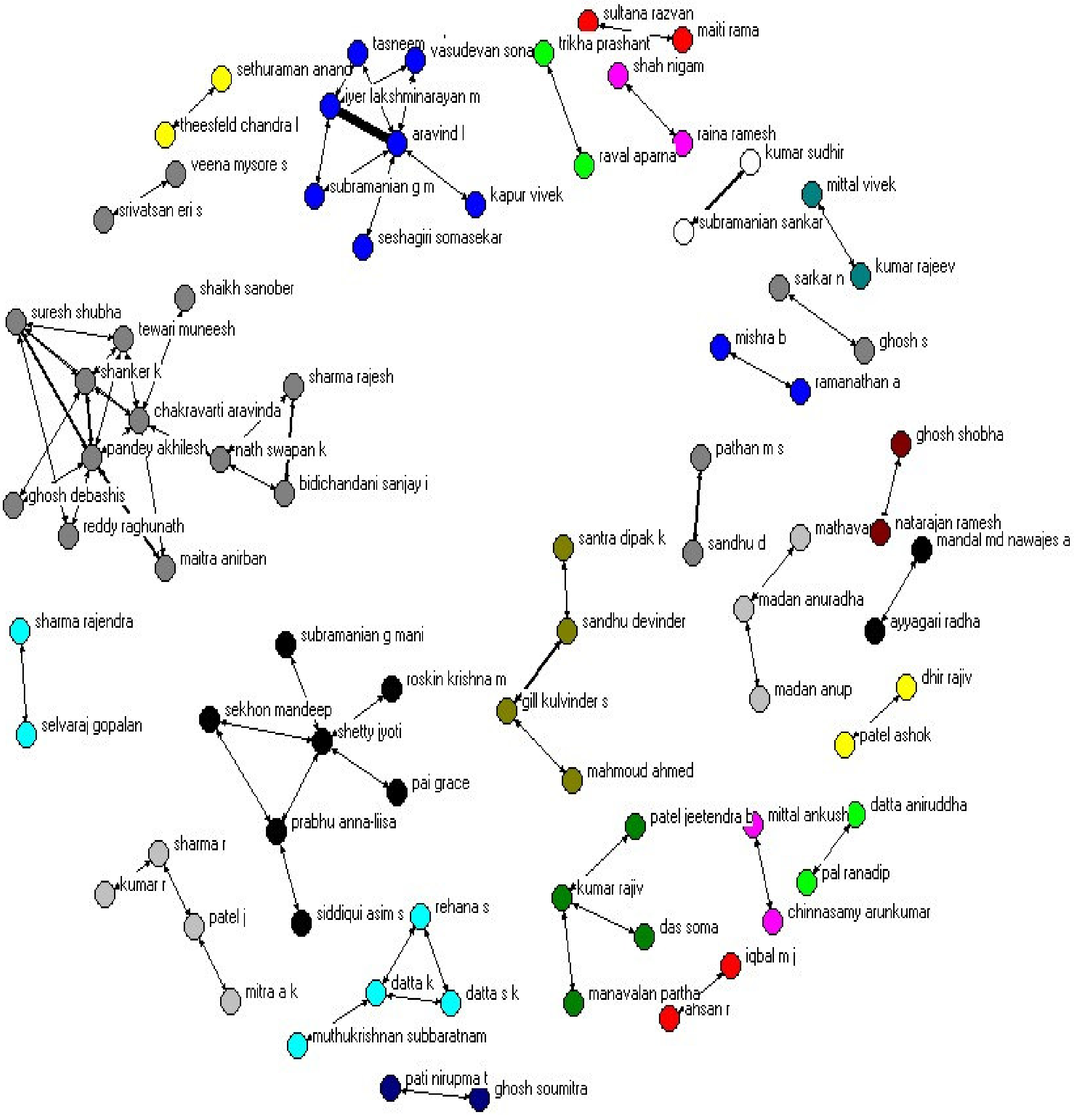
JOURNAL	2001	2002	2003	2004	2005	Total
The Journal of biological chemistry.	42	61	74	31	72	280
Proceedings of the National Academy of Sciences of the United States of America.	17	52	50	53	60	232
Bioinformatics (Oxford, England)	3	20	31	65	112	231
Genome research.	16	27	82	64	40	229
Nucleic acids research.	9	33	47	71	38	198
NOT AVAILABLE	13	11	9	9	98	140
Genomics.	10	25	32	12	51	130
Biochemical and biophysical research communications.	12	24	21	30	28	115
Cancer research.	5	15	21	29	36	106
Genome biology	12	24	18	16	26	96
Nature.	3	25	20	20	23	91
Biochemistry.	14	19	11	18	16	78
Genetics.	9	2	11	8	47	77
Journal of immunology (Baltimore, Md. : 1950)	6	19	13	18	21	77
Physiological genomics.	4	12	9	15	32	72
BMC bioinformatics [electronic resource].	0	10	9	22	30	71
Science.	5	16	10	12	26	69
Clinical cancer research : an official journal of the American Association for Cancer Research.	3	7	7	25	25	67
BMC genomics [electronic resource].	2	10	3	24	24	63
Blood.	3	18	18	5	18	62
Journal of the Indian Medical Association.	0	13	17	10	21	61
Genome / National Research Council Canada = Génome / Conseil national de recherches Canada.	1	11	15	5	23	55

The table shows that on an average there have been more than 45 papers each year in PNAS and BMC Bioinformatics, and 15-20 papers in Science and Nature.

4.4. Collaboration Networks

Most of the papers in the area of Genomics and Bioinformatics are multiauthored as is to be expected in an area that brings together people from diverse disciplines, such as microbiology, genetics, medicine and computer science. All the researchers of Indian origin had a large number of non Indian collaborators. However in this paper we only look at links between Indians to explore leadership or mentoring roles within the community. Out of 700 BIOGEN researchers of Indian origin in the US, about 90 had co-authorship links amongst themselves.

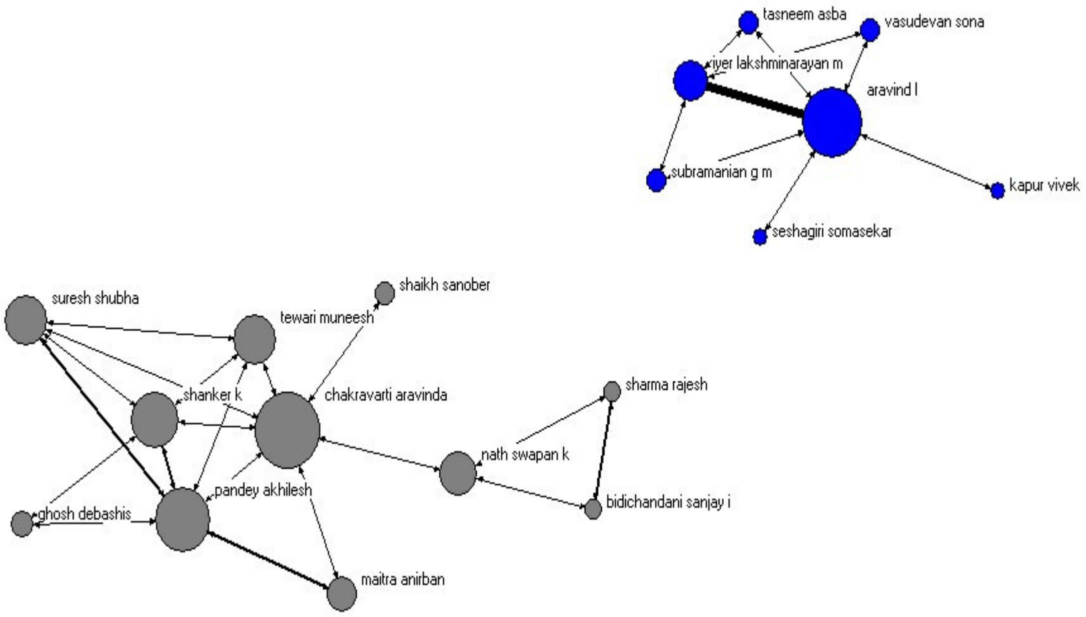
Figure 4. Co-authorship networks among the community of researchers of Indian origin



The co-author networks seen in Fig.4 are drawn using the software UCINET[4]. The thickness of the links is an indicator of the intensity of collaboration. There are a number of disconnected networks, probably indicating local linkages between members.

We examine more closely two of the more extended networks shown in Figure 4, expanded in Figure 4b. The size of the nodes in Fig. 4b indicates the degree of connectivity of the individual in the network.

Figure 4b. Two networks amongst researchers of Indian origin at NCBI and Johns Hopkins University



The first group contains the highly productive individual L Aravind who has collaborated extensively with his colleagues, in particular Iyer, at the National Center for Biological Information and with others, and is tightly linked to all other collaborators. The other group has a more open structure and several highly productive authors at the Johns Hopkins University.

5. Conclusions and Future Work

Our first exploratory study of the Indian community in the US pursuing research in the area of Genomics and Bioinformatics, shows that there are a large number of active individuals, including some who are highly productive. They work in large collaborative groups, and publish their research in highly regarded journals. The annual production of papers from this community is an order of magnitude higher than that from India in the same journal set. The main difficulty in this exercise was to capture Indian names, draw the boundaries of the discipline and distinguish between persons with similar names. In a continuation of this study we will examine the specific research content using multidimensional analysis.

References

1. B. Khadria, Human Resources in Science and Technology in India and the international mobility of highly skilled Indians, OECD STI Working paper Series, No. DSTI/DOC(2004)7
2. P. Stephan and S. Levin Exceptional contributions to US science by foreign born and foreign educated, Population Research and Policy Review, 20(1-2) April 2001
3. National Biotechnology Development Strategy, Department of Biotechnology, Government of India, Draft(20/12/2005).
4. S.P. Borgatti, M.G. Everett and L.C. Freeman, UCINET 6.0 Version 1.00. Analytic Technologies, 1999.