Productivity and Influence in Bioinformatics: A Bibliometric Analysis using PubMed Central

Min Song
Department of Library and Information Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, Korea
E-mail: min.song@yonsei.ac.kr

SuYeon Kim
Department of Library and Information Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, Korea

Guo Zhang
School of Library and Information Science, Indiana University Bloomington, IN, USA

Ying Ding
School of Library and Information Science, Indiana University Bloomington, IN, USA

Tamy Chambers
School of Library and Information Science, Indiana University Bloomington, IN, USA

Abstract

Bioinformatics is a fast growing field based on the optimal the use of “big data” gathered in genomic, proteomics, and functional genomics research. In this paper, we conduct a comprehensive and in-depth bibliometric analysis of the field of Bioinformatics by extracting citation data from PubMed Central full-text. Citation data for the period, 2000 to 2011, comprising 20,869 papers with 546,245 citations, was used to evaluate the productivity and influence of this emerging field. Four measures were used to identify productivity; most productive authors, most productive countries, most productive organization, and most popular subject terms. Research impact was analyzed based on the measures of most cited papers, most cited authors, emerging stars, and leading organizations. Results show the overall trends between the periods, 2000 to 2003, and, 2004 to 2007, were dissimilar, while trends between the periods, 2004 to 2007, and, 2008 to 2011, were similar. In addition, the

field of bioinformatics has undergone a significant shift to co-evolve with other biomedical disciplines.

Introduction

The rapid development of powerful computing technology has fueled a global boom in the biomedical industry that has led to the explosive growth of biological information generated by the scientific community. Bioinformatics, a coupling of molecular biology and computing technology, plays an essential role in understanding human diseases by using genomic information to identify new molecular targets for drug discovery. Many universities, government institutions, and pharmaceutical firms have established bioinformatics groups by bring together computational biologists and bioinformatics computer scientists. These groups have made great progresses illustrating and clarifying massive amounts of information and thus directing bioinformatics into an increasingly multidisciplinary field. A deep and appropriate investigation of this field, including quantitative analysis to identify the disciplines that constitute it, is now of paramount importance.

Peer-reviewed scientific literature is regarded as an excellent means of understanding disciplinary evolution, as it reflects worldwide research activities, encompasses all sectors of employment, and provides the opportunity for bibliometric analysis. As a well-established method to map the structure and development of a given field (McCain, 1990; Ding, 2010; Boyack, Klavans, & Börner, 2005; Ding, Chowdhury, & Foo, 2001), Bansard (2007) defines three reasons for the popularity of bibliometric analyses: 1) the availability of full databases of scientific literature with worldwide electronic access; 2) the availability of efficient tools to perform automatic textual analysis; and 3) the major interest of institutions seeking analyses of recent research trends to position national effort outcomes in relation to others.

Therefore, this paper uses citations and publications collected from PubMed Central

full-text database to conduct a bibliometric analysis, and to illustrate the development pattern of Bioinformatics over the past ten years. Our analysis focuses on research productivity and influence, which we measure using the most productive and cited papers, authors, organizations, and countries. We also identify and examine emerging researchers and “new stars” in the field and augment our citation analysis by adopting the topic modeling technique.

The rest of this paper is organized in the following order: Section 2 gives a brief history of bioinformatics; Section 3 reviews related works on bibliometric analysis and its application in bioinformatics; Section 4 presents the research methods used in this study; Section 5 discusses the content analysis by topic modeling, as well as, the productivity and impact of bioinformatics; and Section 6 summarizes the results and provides implications for future research.

Background

An important landmark of the emerging bioinformatics field was the formal initiation of the Human Genome Project (HGP) in 1990, which sought to sequence and map all human genes — more than 30,000. By 1991, a total of 1,879 human genes had been mapped. In 1993, Genethon, a human genome research center in France, produced a physical map of the human genome, and three years later it published the final version of the Human Genetic Map to complete the first phase of the HGP. In 1997, PSI-BLAST (Position-Specific Iterated Basic Local Alignment Search Tool. See Altschul et al., 1997) was invented for investigating sequence similarity. Using this tool, a query protein or nucleotide sequence could be compared to nucleotide or protein sequences in a target database, to identify regions of local alignment and report those alignments with scores above a given score threshold (1 and BLAST chapter). In 2000, a Fly genome was completely sequenced (Adams et al., 2000). In March of the same year, the Drosophila melanogaster genome sequencing project was

essentially completed. The project planned to map large-insert clones for sequencing, but by the end adopted a Whole Genome Shotgun (WGS) approach marking the first time such an approach was used for sequencing in a multicellular organism. The human genome (3 Giga base pairs) was published in 2001 and HGP was completed in 2003. The draft genome sequence of the brown Norway laboratory rat, Rattus norvegicus, was completed by the Rat Genome Sequencing Project Consortium in 2004. Reactome, the knowledge base of biological pathways, was developed in 2005. A major milestone was achieved in September 2008, when the UniProt/Swiss-Prot group completed the manual annotation of the acknowledged full set of human proteins (derived from about 20,000 genes).

Ten years ago, the only way to track genes was to scour large, well-documented, family trees of relatively inbred populations (e.g. Ashkenzai Jews from Europez). Requested by corporate clients, such types of genealogical search may surf 11 million nucleotides a day. Today, the field of bioinformatics is burgeoning because of the increased need to create massive databases (e.g. GenBank, EMBL, and DNA Database of Japan) to store and compare the DNA sequence data from HGP and other genome sequencing projects. Bioinformatics has also expanded to a broader field which includes; protein structure analysis, gene and protein functional information, data from patients, pre-clinical and clinical trials, and the metabolic pathways of numerous species.

Because of rapid development over the last ten years, it is now critical to investigate the current status of bioinformatics, including identifying its major players (e.g. the most productive and highly cited authors) and new driving forces. This will both explain its historical evolution and shed light on its future direction. Additionally, as bioinformatics is a burgeoning field, it has triggered innovations across the fields of genomics, computational biology, and bio-imaging. There is thus a need to evaluate its current research performance

and landscape, so as to facilitate potential interdisciplinary collaboration in the future.

Related Work

Bibliometrics: Exploring research productivity and scholarly impact

Bibliometrics is a well-established quantitative approach used to explore research productivity and scholarly impact, which are two interactive and mutually complementary measures for academic performance. It has been widely used for establishing scholarly performance of authors (e.g. Cronin & Overfelt, 1993; Yan & Ding, 2010), citation patterns of journal articles (Moed, 2005), and the impact of journals (e.g. Garfield, 1955; 2000).

As one of two essential measures, research productivity is usually described in terms of the quantity of publications produced by individuals and institutions. Ramsden (1994) reported that both internal personal variables (e.g. research talents) and structural variables (e.g. institution management) could impact the level of research productivity. Yan and Sugimoto's (2011) exploration of the social, cognitive, and geographic relationships between institutions, based on their citation and collaboration networks, led to findings that institutional citation behaviors are associated with social, topical, and geographical factors and less dependent on the country boundary or physical distance. He, Ding, and Ni (2011) studied the contextual information of scientific collaboration networks and identified that researchers with a broad range of collaborations tended to have increased productivity.

The other measure, scholarly impact, is usually defined as the extent to which a researcher's work (e.g. a paper) has been used by other researchers (Bornmann et al., 2008). Scholar impact can thus be measured by the number of citations made to it by other scholars. As Cronin (1981) stated, "citations are frozen footprints in the landscape of scholarly achievement; footprints which bear witness to the passage of ideas" (p. 16). Nicolaisen (2007) reviewed various theories of citation behavior and citation analysis before introducing

the, now widespread, belief that citing can be regarded as an evolutionary account of science and scholarship, and understood in terms of psychology, the normative theory, and the social constructivist theory. In fact, the process of selecting and dressing a work with references is far from random (Cronin, 1981; Small, 2011). There exists a set of norms—Cronin (2004, p. 43) speaks of "the normative ghost in the machine"—and procedural standards to which scientists typically adhere (e.g., Cronin, 1984; Small, 1976). Therefore, citation analysis, as a major component of bibliometrics, has become an important way to estimate the value, credit, and contribution of a certain paper, journal, institution, or individual (Brown & Gardner, 1985). Recently, a few researches have proposed more refined approaches to measuring scholarly impact. Ding and Cronin (2011) differentiated popularity from prestige by taking the importance of the source of citations into account. Ding (2011) applied weighted PageRank to author citation networks in the information retrieval field. He, Ding, and Yan (2012) proposed a sequence-based mining method to reveal the collaboration patterns for multi-authored papers.

Bibliometric analyses in bioinformatics

Several bioinformatics researchers have applied bibliometric analyses to understand the development of this field. Patra and Mishra (2006) analyzed the growth of the scientific literature in bioinformatics collected from NCBI PubMed using standard bibliometric techniques (e.g. Bradford's law of scattering and Lotka's law). Their study identified core primary journals, productivity patterns of authors and their institutions, publication types, used languages, and countries of publication to conclude that bioinformatics is a relatively new area and still does not have any specific scientific community behind it. Also focusing on literatures, Janssens et al. (2007) and Glänzel et al. (2009) analyzed the core bioinformatics literature by incorporating text mining and bibliometric, citation-based techniques. The primary focus of their study was to improve the classification of literature based on a

combination of linguistic and bibliometric tools.

Manoharan et al. (2011) conducted a bibliometric analysis of the corpus of bioinformatic literature covered by Thompson's Web of Science database for the period ,2000 to 2010, aiming to evaluate the publication frequency, country, individual productivity, and collaboration in the field. Their overall conclusion was that bioinformatics may risk becoming a purely scholarly and unevenly distributed discipline, because only a few countries (e.g. India and China) produce the majority of the publications. Using the same database (Thompson's Web of Science), Huang et al. (2012) analyzed the citation patterns in bioinformatics journals (instead of the citation patterns of individual articles) and their corresponding knowledge subfields by normalizing the journal impact factor available in Journal Citation Report (JCR). Their results showed that bioinformatics journal citations were field-dependent, with scattered patterns in article life span and citing propensity. However, both studies were limited by their data source – only Thompson's Web of Science database – which is biased towards certain domains, languages, and regions, and by their focus on merely journal-level citation patterns.

Seeking to derive potential and beneficial collaboration, Bansard et al. (2007) analyzed the bioinformatic and medical informatic literature to identify present links and potential synergies shared between the two research fields. Their bibliometric analysis used the most significant words and groups of words from the documents to find that bioinformatics and medical informatics were still relatively separate fields, despite both having undergone fast changes and the use by both of advanced computer techniques to process massive biological data. The major limit of their study was their complete dependence on "words" or "word co-occurrences," which should be estimated together with other normalization techniques to decrease contextual errors. In summary, bibliometric analysis has been used to map the

research trends of bioinformatics (e.g. Manoharan et al., 2011; Patra and Mishra, 2006), to compare bioinformatics research in different countries (e.g. Guan & Gao, 2008; Manoharan et al., 2011), and to identify key words, scholars' prominence, and research collaboration (e.g. Glänzel et al., 2009a; Patra & Mishra, 2006). However, as a relatively young field, further study is still needed to identify and define bioinformatics, especially its impact and productivity.

Methods

As bioinformatics is a highly interdisciplinary field, journals that contribute to bioinformatics tend to be cross-disciplinary. The bioinformatics journals in this study were, therefore selected from diverse sources. The selection criteria were originally provided by Huang and his colleagues (2011). We used most of the journals in their study and added a few more sources. Our additional sources were compiled from the following:: 1) The International Society of Computational Biology (http://www.iscb.org/iscb-publications-journals), 2) The bioinformatics journal list on Wikipedia (http://en.wikipedia.org/wiki/List_of_bioinformatics_journals), and 3) The Mathematical and Computational Biology section of the Web of Science's Science Journal Citation Reports (SJCR). From these sources, we compiled a comprehensive list of 48 bioinformatics journals found in PubMed Central (Table 1). The choice of PubMed Central instead of Web of Science, which has been used in previous studies, was influenced by the fact that only 34 (72%), of the 48 journals were indexed in the Web of Science. All full-text articles pertinent to bioinformatics in the 48 journals were collected, which totaled 20,869 papers. However, some journals did not have many full-text articles, which has slightly limited this study.

Table 1: Journals selected and the number of papers included.

Journal	No. Paper	Journal	No. Paper

BMC Bioinformatics	3982	Source Code for Biology and Medicine	53
BMC Genomics	3203	Advanced Bioinformatics	42
PLoS Biology	2648	BioData Mining	32
Genome Biology	2321	Journal of Computational Neuroscience	26
PLoS Genetics	1876	Journal of Proteome Research	23
PLoS Computational Biology	1613	Journal of Biomedical Semantics	18
BMC Research Notes	744	Journal of Computer-Aided Molecular Design	18
Bioinformatics	705	Genome Integration	16
Molecular Systems Biology	485	Journal of Molecular Modeling	12
BMC Systems Biology	480	Bulletin of Mathematical Biology	11
Comparative and Functional Genomics	478	Pharmacogenetics and Genomics	9
Bioinformation	398	Statistical Methods in Medical Research	9
Theoretical Biology and Medical Modeling	256	Neuroinformatics	6
Human Molecular Genetics	223	Genomics	5
The EMBO Journal	215	Protein Science	5
Cancer Informatics	168	Physiological Genomics	4
Genome Medicine	134	Trends in Genetics	4
Evolutionary Bioinformatics	121	Journal of Proteomics	3
Biochemistry	115	Proteomics	3
Algorithms for Molecular Biology	110	Trends in Biochemical Sciences	3
EURASIP Journal on Bioinformatics and Systems Biology	86	Journal of Biotechnology	2
Journal of Molecular Biology	81	Trends in Biotechnology	2
Molecular & Cellular Proteomics	64	Briefings in Functional Genomics & Proteomics	1
Mammalian Genome	55	Journal of Theoretical Biology	1

To extract elements of interest, such as title, abstract, and references from the full text we developed a SAX XML parser in Java. Based on an event-driven sequential access model, this was effective at processing the large dataset due to its low memory requirements. To recognize data elements, we used Journal Publishing DTD made by NLM (National Library of Medicine) available at http://dtd.nlm.nih.gov/publishing/w3c-schema.html. The extracted elements were stored in a relational citation database we built for further analysis. Figure 1 shows the database schema for this citation database.

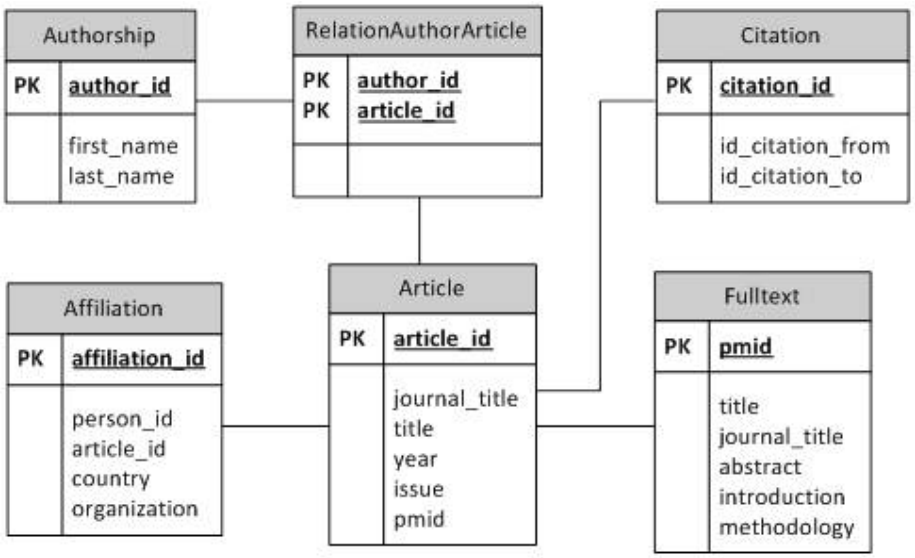

Figure 1. Database schema for a custom-made database.

One of the challenges in building such a citation database was the need to detect duplicate citations, which was made even more difficult by the use of different citation styles in the reference section of the full-text articles. PubMed Central XML data helped detect duplicate records by providing different XML tags to citation elements (e.g. author name, journal title, publication year, etc.) in the reference section. However, these tags could not cover all citations. To improve the accuracy of spotting citations, we employed the edit distance technique, SoftTFIDF, to compare two entities in terms of string similarity (Cohen et al., 2003). Cohen and his colleagues reported that SoftTFIDF outperformed other compared edit distance techniques with 0.91 average precision using the UTA dataset, and 0.914 average precision using the CoraATDV dataset. A pilot test, conducted with our dataset, achieved a 0.92 average precision. After populating the extracted citation data into the tables shown above, we had the following number of instances: Affiliation – 60,263; Articles – 20,869; Authors – 445,034; Citation – 546,245; RelationAuthorArticle – 2,264,079. The RelationAuthorArticle table paired each author to the paper the author (co)authors on it.

A major source of error in processing PubMed Central citation data was related to disambiguation of author names. The problem was exacerbated when the first name was only initialized in the reference section. To solve this problem, we developed an automatic procedure that linked PubMed Central papers to PubMed papers through the PubMed E-Utilis APIs (http://www.ncbi.nlm.nih.gov/books/NBK25500/) to obtain the full first name and the author affiliation information. Due to the low matching rate between the PubMed Central ID and the PubMed ID, we searched PubMed with paper titles including ambiguous author names. Still, a lot of affiliation information was difficult to extract, therefore, we manually checked the top 200 most productive authors and most highly cited authors. We found seven ambiguous authors from the most productive authors list and nine from the most highly cited authors list. We mention in our future work that more comprehensive methods should be applied to disambiguate author names (Tang et al., 2012). Because of this, out of 20,869 papers with 546,245 citations, only 310,002 (57%) citations came from the PubMed database.

As major progress within this field began in early 2000, when it acquired major funding from European Commission and U.S.A., we chose to portrait details of this important phase and better outline the field's dynamic changes, by dividing the period, 2000 to 2011, into three phases; 2000 to 2003, 2004 to 2007, and 2008 to 2011. This resulted in 132,051 citations for the period, 2000 to 2003, 180,570 for the period, 2004 to 2007, and 64,064 for the period, 2008 to 2010. It should be noted that there were 169,560 citations published before 2000.

We divided the time span into three phases for the following reasons: 1) in order to do meaningful topic modeling, we needed to guarantee a certain numbers of articles per period (dividing it into finer-grained levels would have deteriorated the quality of topic modeling); 2) the number of publications per year varied, which could have led to potential bias in the results analysis; and 3) within bioinformatics some noticeable trends are marked by these

three phrases (ie. during the period, 2000 to 2003, the major of topic was the protein study, while during the period, 2004 to 2007, topics diversified to include sequence and structure analysis of genes, brain, cancer, virus, etc.).

To identify author productivity and impact, we divided authors into two categories; first author and second author. The first author category included authors who were indicated as the first author, while the second author category included the remaining set of co-authors. Author order is usually tightly connected to contribution, as first authors tend to be those who contributed the most to the paper and are often the corresponding author. By example, tenure promotion at major universities in the U.S.A., recognize author order as one of the most important indicators for measuring faculty member contribution. While it varies from discipline to discipline and from country to country, from the authors' own experience, significant contribution still comes from the first author in the bioinformatics domain. According to Sekercioglu (2008), author order is of particular importance in bioinformatics.

We applied a topic modeling technique to analyze research productivity and author/ country impact associated with the identified topics. Topic modeling has often been used to identify topics from large-scale document collections. In the model, a topic represents an underlying semantic theme, approximated as an organization of words, and operationalized as a probability distribution over terms in a vocabulary (Blei et al., 2003). The topic modeling technique used in this paper is Dirichlet-multinomial regression (DMR) proposed by Mimno and McCallum (2008), which is an extension of the Latent Dirichlet Allocation (LDA) proposed by Blei et al. (2003). It allows conditioning on arbitrary document features by including a long-linear prior on document-topic distributions that is a function of the features of the document, such as author, publication venue, references, and dates. By applying the

topic modeling technique to the bioinformatics journals we collected, we could examine which leading countries or authors have strengths in each topic.

In addition to extending the paper by incorporating topic modeling for productive authors and countries, we also conducted co-authorship analysis to understand scientific collaboration patterns and the status of bioinformatics researchers. Studies of co-authorship networks have relied on topological features, including centrality, largest component, diameter, clustering coefficient, average separation, average number of collaborator etc. (Yang et al., forthcoming).

Topic modeling was also used to spot thematic development in bioinformatics over time, as defined by our three time periods. We used the MALLET package (McCallum, 2002) as a basis for our system and extended the DMR topic modeling algorithm implemented in MALLET to suit our needs. MALLET was applied on each period of interest to find the top topic groups. We used 1000 iterations with stop word removal.

Results and discussions

Content Analysis by Topic Modeling

By and large, there are two major subfields in bioinformatics: 1) computational bioinformatics and 2) application bioinformatics (Baldi & Brunak, 2001). Computational bioinformatics uses computational work, including algorithm, software development, database construction and curation, to develop applications that are aimed at addressing certain problems in biology. Applications of bioinformatics can be categorized into three groups: sequence, function, and structure analysis. Sequence analysis covers various types of sequence information on genes and proteins. Function analysis analyzes the function expressed within the sequences, and predicts the functional interaction between various proteins or genes. Structure analysis predicts the structure, and possible roles for the structure

of proteins or RNA. We use this general taxonomy of the bioinformatics field to analyze the results of topic modeling (Table 2-4).

During the period, 2000 to 2003, the major topic is protein study (topic 2, topic 4, topic 5, and topic 12) with particular interest in those topics relate to the functional analysis of proteins – a core component of application bioinformatics.

Table 2: DMR-based Topic Modeling Results for the Period of 2000 and 2003.

Topic1: Cell cloning	Topic2: Protein sequence	Topic3: Ontology	Topic4: Protein prediction	Topic5: Protein analysis
cell	sequences	data	model	proteins
cloning	region	information	measures	conserved
genes	alignment	database	predictions	domain
expression	protein	ontology	protein	function
development	algorithm	biological	experiments	family
mapping	method	tools	parameters	analysis
Topic6: Gene study	Topic7: DNA binding	Topic8: Yeast network	Topic9: Gene expression	Topic10: RNA/DNA
genes	dna	yeast	expression	amplification
study	sites	protein	data	rna
identified	binding	networks	gene	dna
gene	transcription	analysis	microarray	rnai
mutations	regulatory	coli	genes	gene
tmc	motifs	mass	analysis	protocol
Topic11: BTBD/Receptor	Topic12: Protein/ arabidopsis	Topic13: Tuberculosis/ Genomics	Topic14: Chromosome/ Mutations	Topic15: Genomics
receptor	proteins	tuberculosis	chromosome	genomics
channel	arabidopsis	functional	mutations	cdca
btbd	plant	comparative	biology	gene
binding	membrane	current	genotyping	expression
mhc	family	awareness	human	genome
olfactory	plants	genomics	snp	sequence

Topics during the period, 2004 to 2007 are more diverse and include sequence and structure analysis of genes, brain, cancer, virus, etc. In addition, two topics directly relate to computation bioinformatics (topic 3 and topic 12), which is different from the first period.

Table 3: DMR-based Topic Modeling Results for the Period of 2004 and 2007.

Topic1: Protein structure	Topic2: Brain	Topic3: Ontology	Topic4: Immune/Virus	Topic5: Gene
protein	brain	gene	immune	gene
structure	neurons	annotation	infection	genomics
family	circadian	ontology	virus	cell
binding	activity	functional	host	sequence
peptide	cortex	terms	viral	genome
method	neural	biological	hiv	est
Topic6: Network pathway	Topic7: Gene expression	Topic8: DNA/Chromosome	Topic9: Cancer research	Topic10: Gene transcription
network	data	dna	mass	transcription
pathway	gene	chromosome	research	genes
interactions	expression	microarray	biology	sites
metabolic	microarray	methylation	spectrometry	binding
yeast	analysis	chromatin	cancer	motifs
protein	profile	hybridization	new	regulatory
Topic11: Gene evolution	Topic12: Database/Software	Topic13: Gene/Genome	Topic14: Cell/Model	Topic15: SNP/Disease
evolution	data	genome	model	genetic
species	database	genes	cell	snps
gene	analysis	genomes	development	disease
phylogenetic	software	bacterial	system	polymorphisms
duplication	tool	sequences	stem	variation
human	information	bacteria	signaling	association

During the period, 2008 to 2011, topics continue to be diverse and similar to the second period. However, new topics like mutation and RNA emerge during this period.

Table 4: DMR-based Topic Modeling Results for the Period of 2008 and 2011.

Topic1: Ontology/Mining	Topic2: Gene sequence	Topic3: Gene/Protein	Topic4: DNA/Chromosome	Topic5: HIV/Virus
information	gene	genes	dna	patients
research	sequence	proteins	methylation	Study
ontology	marker	plant	chromatin	hiv
biomedical	splicing	identified	cells	clinical
terms	genome	analysis	histone	virus
system	analysis	expressed	chromosome	health
Topic6: Mutation	Topic7: Protein binding	Topic8: Cancer	Topic9: Network pathway	Topic10: SNP/Disease
mutations	protein	cells	network	genetic

mice	binding	cancer	pathway	association
disease	molecular	profile	modules	snps
mutation	structure	tumor	interaction	disease
protein	sequence	human	protein	studies
muscle	prediction	breast	biological	polymorphisms
Topic11: Algorithm/Database	Topic12: Neuron/Dynamics	Topic13: Metabolism	Topic14: RNA	Topic15: Cell signaling
data	model	metabolic	rna	cell
method	neuron	metabolism	binding	signaling
algorithm	dynamics	growth	sites	receptor
database	system	coli	transcription	protein
software	time	bacteria	mirnas	kinase
tool	cell	response	regulatory	development

Topics over the entire time period, 2000 to 2011, follow patterns similar to the second and the third periods due to the bulk of datasets coming from those two periods and the sensitivity of topic modeling to the size of datasets. Table 5 shows all 15 topics, which include protein binding, algorithm/method, cell/model, network/interaction, genome sequence, immune/virus, gene expression, genetic/evolution, database/software, gene transcription, DNA/chromosome, ontology/mining, gene/genomics, and cancer/cell.

Table 5: Overall Topic Modeling Results for the Period of 2000 and 2011.

Topic1: Protein binding	Topic2: Algorithm/Method	Topic3: Cell/Model	Topic4: Network/Interaction	Topic5: Genome sequence
protein	data	model	network	genome
binding	method	cell	interactions	genomic
receptor	methods	cells	pathway	sequence
sequence	algorithm	system	gene	dna
structure	model	dynamics	interaction	plant
domain	approach	time	biological	species
Topic6: Immune/Virus	Topic7: Gene expression	Topic8: Generic/Evolution	Topic9: Database/Software	Topic10: Gene transcription
infection	expression	evolution	data	gene
host	gene	selection	database	transcription
strains	genes	evolutionary	analysis	response
immune	microarray	species	information	sites

virus resistance	data analysis	genetic variation	software tool	stress metabolism
Topic 11: DNA/Chromosome	Topic 12: Ontology/Mining	Topic 13: Disease/SNP	Topic 14: Gene/Genomics	Topic 15: Cancer/Cell
dna cells chromatin histone chromosome replication	research biology information biomedical text ontology	genetic disease association snps studies study	genes genome gene genomes species sequence	cancer cell genes tumor expression cells

Productivity

Productive authors

Appendix A shows the top 15 most productive authors in bioinformatics. based on PubMed Central data. Over the entire period, 2000 to 2011, the most productive author is Michael L. Gross, who published 124 papers in the period , 2004 to 2007. In terms of consistent productivity, G.A. Petsko leads, by ranking first or second in all three periods; 2000 to 2003, 2004 to 2007, and 2008 to 2011. R. Robinson is also a highly productive researcher, ranking third in the period, 2004 to 2007, and first in the period, 2008 to 2011.

In the second author category, P. O. Brown ranks fourth in the period, 2000 to 2003 and third in the period, 2004 to 2007. M. Gerstein ranks fourteenth in the period, 2000 to 2003, eleventh in the period, 2004 to 2007, and fifth in the period, 2008 to 2011 respectively, which shows his steady production in the field of bioinformatics over the entire period. Among the top 15 productive authors from 2000 to 2003, no author, except G.A. Petsko, M. Gerstein, and P.O. Brown, is included in the productive author lists for the other two periods. P.E. Bourne emerges in the second period,2004 to 2007, where he ranks seventh in the first author category and fourth in the second author category. He also ranks first in the second author category for the period, 2008 to 2011. In addition, the following authors are productive in first author category for the last two periods, 2004 to 2007 and 2008 to 2011, L. Gross (first

and sixth), R. Robinson (third and first), M. Hoff (fourth and ninth), and R. Jones (seventh and seventh). The list of productive authors for the three consecutive periods reveals that just a few researchers were steadily productive and that highly productive new authors have emerged since 2004.

As we are interested in identifying the topical areas productive authors publish to, we analyzed papers published by top 10 most productive authors in the first author category over the period, 2000 to 2011 using the DMR topic modeling technique to infer the topic distribution of these papers. Figure 2 shows the results of this topic inference on 15 topics. The topic inference was calculated without partitioning the time period to find the general focus of most productive authors up to rank 10 (total 22 authors) on the identified subject areas.

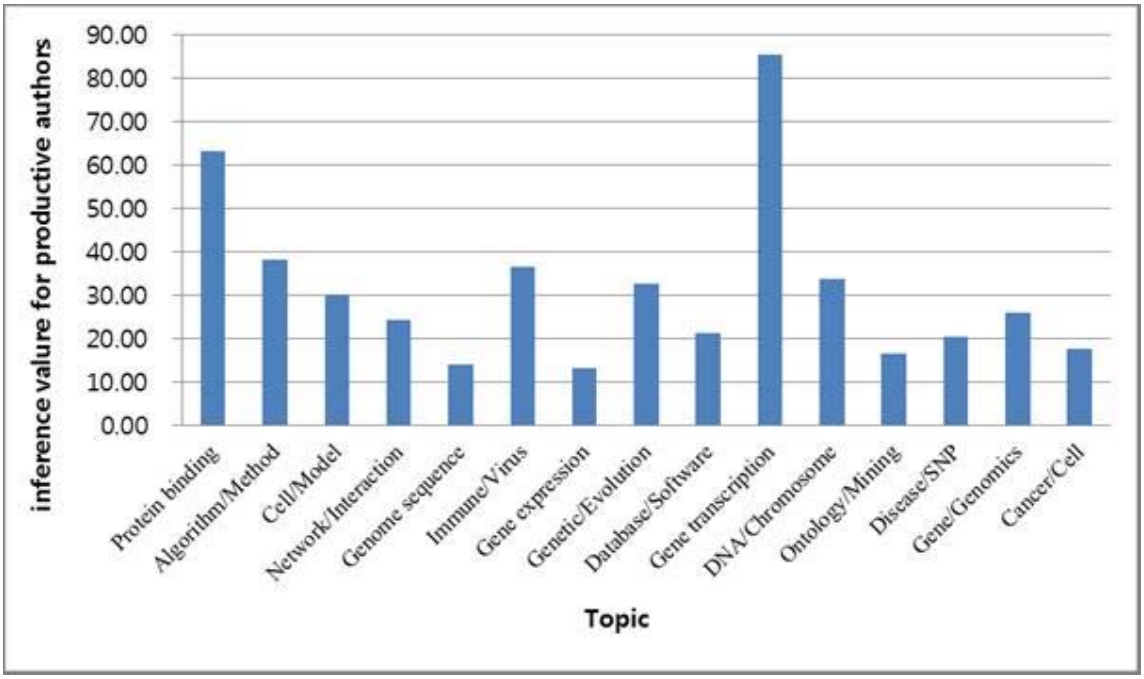

Figure 2. Inference value for topic productive authors.

The average topic inference is 31.49 with a standard deviation of 18.9. Among the 15 topics, two topics exceed the standard deviation. The first topic is related to protein binding and the second is about gene transcription. This implies that the top 10 productive authors focused on those two topical areas more than other topics.

Productive countries

Table 6 shows the top 20 productive countries. Over the entire period, the U.S.A., U.K., and Germany rank first, second, and third respectively. Canada and France rank fourth and fifth alternatively. Denmark is included in the top 20 productive countries in the first two periods, but not in the third period, 2008 to 2011. Belgium is included in the first and the last periods, but not in the period 2004 to 2007. Countries included in the top 20 for the period, 2000 to 2003, but excluded in the last two periods include Norway, Poland, Ireland, and Russia. Since 2004, China, Taiwan, Singapore, and Korea are among the top 20 productive countries. This indicates that, Asian countries have begun to stand out in the field of bioinformatics. Among Asian countries, Japan is the only one included among top 20 countries for all three periods. Other countries shown in Table 6 are included in top 20 but with various rankings in the three periods.

Table 6: Top 20 productive countries.

R	2000-2003		2004-2007		2008-2011	
	Country	no.	Country	no.	Country	no.
1	USA	1090	USA	9314	USA	15683
2	UK	305	UK	1690	UK	3593
3	Germany	114	Germany	1176	Germany	2445
4	Canada	82	France	816	France	1857
5	France	75	Canada	707	Canada	1347
6	Australia	38	Japan	588	China	1153
7	Spain	36	Italy	427	Japan	1145
8	Italy	30	China	385	Italy	895
9	Japan	30	The Netherlands	370	Spain	866
10	Switzerland	27	Australia	334	Australia	833

11	Sweden	25	Spain	318	The Netherlands	832
12	The Netherlands	19	Switzerland	303	Switzerland	622
13	Belgium	18	Sweden	267	Sweden	567
14	Norway	17	Israel	252	India	485
15	Denmark	14	India	247	Israel	459
16	Poland	11	Taiwan	195	Taiwan	427
17	India	10	Singapore	162	Belgium	355
18	Finland	9	Finland	152	Korea	343
19	Ireland		Denmark	146	Singapore	333
20	Israel Russia		Korea	144	Finland	303

To generate topic models with the condition on countries (Figure 3), we selected the top five countries for analysis, U.S.A., U.K., German, Canada, and France, based on consistent ranks within the top 10 during the period of 2000 to 2011. Topic modeling results confirm that the U.S.A. is the leading country. For the period, 2000 to 2003, the top five countries have a strong topical relationship with gene expression and genomics, with all five countries exceeding the inference average. For the period, 2004 to 2007, the top five countries have strong research interests in four topics; gene, gene transcription, gene evolution, and cell/model. In the period, 2008 to 2011, topical interests shift to gene sequence, HIV/virus, metabolism, and algorithm/database. Topic modeling with the condition of a country reveals that the early interests of the top five countries are in gene-centric research, but that recently their interests have expanded to include disease research and computational tools.

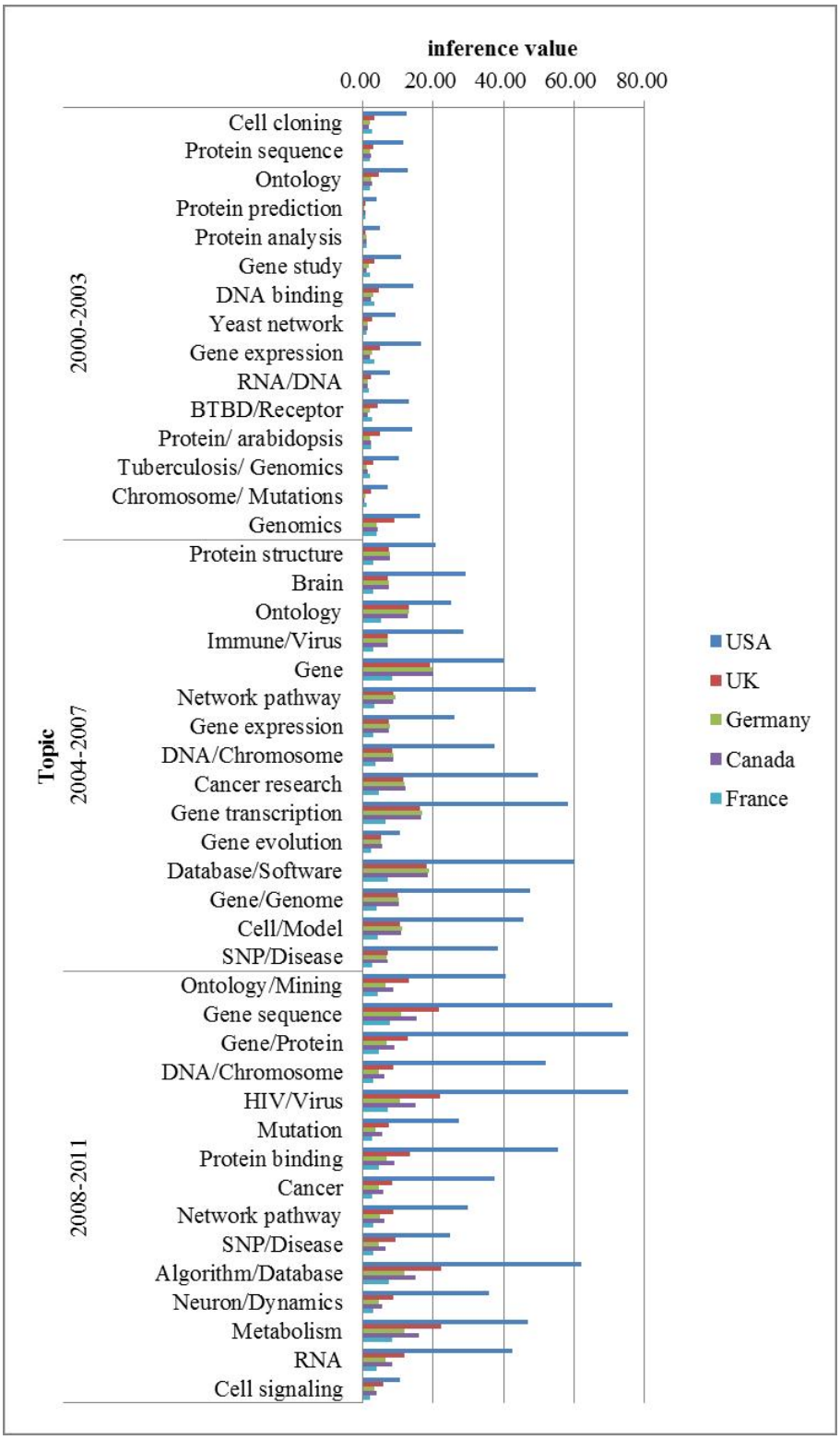

Figure 3. Topic distribution of top countries over the period 2000 and 2011.

Productive organizations

Table 7 shows the leading organizations in the field of bioinformatics. Brandeis University ranks first in the period, 2000 to 2003, and twentieth in the period, 2004 to 2007, but is not included in the period, 2008 to 2011. The University of California Berkeley ranks second for the period, 2000 to 2003, seventh for the period, 2004 to 2007, and fourteenth for the period, 2008 to 2011. Stanford University ranks third in the period, 2000 to 2003, and first since 2004. Harvard University ranks sixth in the period, 2000 to 2003, second in the period, 2004 to 2008, and third in the period, 2009 to 2011. The University of Washington ranks fifth in the first two periods, and second in the period, 2008 to 2011. Two institutions have steadily increasing rankings over the three time periods; the University of Cambridge (eleventh, eighth, and fifth), and the University College London (seventeenth, eleventh, and tenth). The University of Oxford is not included in the period, 2000 to 2003, but ranks tenth the period, 2004 to 2008, and sixth in the period, 2009 to 2011.

Table 7: Top 20 most productive organizations.

R	2000-2003		2004-2007		2008-2011	
	Organization	no.	Organization	no.	organization	no.
1	Brandeis University	47	Stanford University	315	Stanford University	514
2	University of California Berkeley	44	Harvard University	283	University of Washington	506
3	Stanford University	43	University of California at San Diego	206	Harvard University	481
4	National Center for Biotechnology Information	29	University of California San Francisco	188	University of California-San Diego	347
5	University of Washington	27	University of Washington	223	University of Cambridge	307
6	Harvard University		Yale University	165	University of Oxford	281
7	University of Toronto	23	University of California-Berkeley	159	University of California San Francisco	277
8	University of California San Francisco		University of Cambridge	154	University of Toronto	259
9	University of Texas at Austin		University of California Los Angeles	138	Duke University	243
10	Yale University	21	University of Oxford	138	University College London	226

11	University of Cambridge	19	University College London	117	University of California	221
12	European Bioinformatics Institute	14	University of Michigan	116	Yale University	219
13	Duke University		University of Minnesota	114	University of Michigan	205
14	University of Edinburgh	13	Duke University	109	University of California Berkley	204
15	Columbia University	12	Princeton University	107	University of California Los Angeles	200
16	Wellcome Trust Sanger Institute		University of California	106	University of Chicago	190
17	University College London		University of Toronto	105	Princeton University	187
18	New York University		Columbia University	103	University of California-Davis	185
19	Institute for Genomic Research	11	University of Pennsylvania	92	CNRGV	177
20	The Rockefeller University University of Michigan		Brandeis University	91	University of North Carolina	174

Popular subject terms

Subjects assigned to journals in our data collection are listed in Table 8. These subjects are automatically assigned to journals based on the subject heading(s) provided by the Stanford Lane Medical Library. We created an html parser class that connected to the Stanford Lane Medical Library, to query the search engine with the journal title, and parse the extracted subject heading(s) for the corresponding journal. Table 8 show that Molecular Biology and Medical Informatics are the top two subjects, followed by Genetics, Biology, and Biochemistry.

Table 8: Subject terms of journals.

Subject Term	Count
Molecular Biology	11
Medical Informatics	10
Genetics	9
Biology	7
Biochemistry	5
Biomedical Engineering	3
Biotechnology	2
Medicine	2
Neurology	2
A publication of protein society*	1

Computers	1
EURASIP journal on bioinformatics and systems biology*	1
Functional Genomics	1
Genetics, Medical	1
Journal of theoretical biology	1
Molecular cellular proteomics*	1
Oncology	1
protein science	1
Proteomics	1
Technology	1
The EMBO journal*	1
The pharmacogenomics journal*	1
Theoretical biology medical modeling*	1
Trends in genetics*	1

* journal title

Influence

Influential papers

Table 9 shows the top three most cited papers in the field of bioinformatics. We present the rest of the top 20 highly cited papers in Appendix B. Among the papers published in the period 2000 to 2003, the most cited paper is “Gene ontology: tool for the unification of biology,” which was published in Nature Genetics and written by the Gene Ontology Consortium consisting of 20 bioinformatics researchers. Eight authors, among the 20 are included in the top 20 highly cited authors for the same period (D. Botstein, G. Rubin, G. Sherlock, M. Ashburner, J. Cherry, C. Ball, J. Matese, H. Butler). The second most cited paper for this period is “Initial sequencing and analysis of the human genome” published in Nature. The authors of this paper consist of 249 researchers from 48 organizations. The third most cited paper for this period is “Significance analysis of microarrays applied to the ionizing radiation response” written by V. Tusher, R. Tibshirani, and G. Chu, all of whom are affiliated with Stanford University. R. Tibshirani also ranks twelfth in the highly cited author list for the same period.

In the period, 2004 to 2007, the most cited paper is “Bioconductor: open software development for computational biology and bioinformatics” written by 25 authors from 19 organizations. The first author of this paper is R. Gentleman of the Dana-Farber Cancer Institute. Among the 25 authors for this article, four are also included in the highly cited author list for the same period. The second most cited paper for this period is “R: A language and environment for statistical computing” and the third is “Transcriptional regulatory code of a eukaryotic genome” written by 20 authors from four organizations.

During the period, 2008 to 2011, the most cited paper is “The Pfam protein families database” written by 13 authors from three organization. The first author of this paper is A. Bateman, and among the other 13 authors, R. Durbin ranks ninth on the top 20 hightly cited authors list for the period, 2004 to 2007, and first for the period, 2008 to 2011. The second most cited paper for this period is “KEGG for linking genomes to life and the environment” written by 11 authors from three Japanese organizations. The third most cited paper for this period is “Mapping short DNA sequencing reads and calling variants using mapping quality scores” written by H. Li, J Ruan (ninth among highly cited authors for the same period) , and R. Durbin (first among highly cited authors for the same period).

Table 9: Top 3 cited papers.

	2000-2003			2004-2007			2008-2011		
R	paper	journal	no. cited	paper	journal	no. cited	paper	journal	no. cited
1	Gene ontology: tool for the unification of biology. The Gene Ontology Consortium	Nat Genet	948	Bioconductor: open software development for computational biology and bioinformatics	Genome Biol	395	The Pfam protein families database	Nucleic Acids Res	112
2	Initial sequencing and analysis of the human	Nature	465	R: A language and environment for statistical	R: A language and environment	304	KEGG for linking genomes to life and the	Nucleic Acids Res	104

	genome			computing	for statistical computing		environment		
3	Significance analysis of microarrays applied to the ionizing radiation response	Proc Natl Acad Sci USA	349	Transcriptional regulatory code of a eukaryotic genome	Nature	234	Mapping short DNA sequencing reads and calling variants using mapping quality scores	Genome Res.	91

Figure 4 shows topics most pertinent to influential papers. By building three topic models with DMR for three datasets (2000 to 2003, 2004 to 2007, and 2008 to 2011) we are able to select the top 100 most cited papers not part of the datasets used for the topic model, infer topic distribution of each highly cited paper, and sum up an inferred topic value of the highly cited paper. During the period, 2000 to 2003, the average inference value is 6.67 with a standard deviation of 9.099. These statistics imply that influential papers have a significant thematic relationship with topics such as gene expression (26.21) and genomics (30.17). During the period, 2004 to 2007, we observe the topical extension of influential papers with an average influence value of 8.6 and a standard deviation of 7.75. Influential papers, during this period focus on protein structure (19.08), brain (18.18), gene evolution (24.44), and cell/model (21.91). This trend continues in the third period, with an average influence value of 9.29 and a standard deviation of 9.33. Most influential papers during this period focus on the topics of gene sequence (22.79), DNA/chromosome (23.93), SNP/disease (20.48), neuron/dynamics (19.87), and cell signaling (24.01). The diversification of topics starting from the second period is also observed in the thematic focus of the most productive authors, as analyzed in the earlier section.

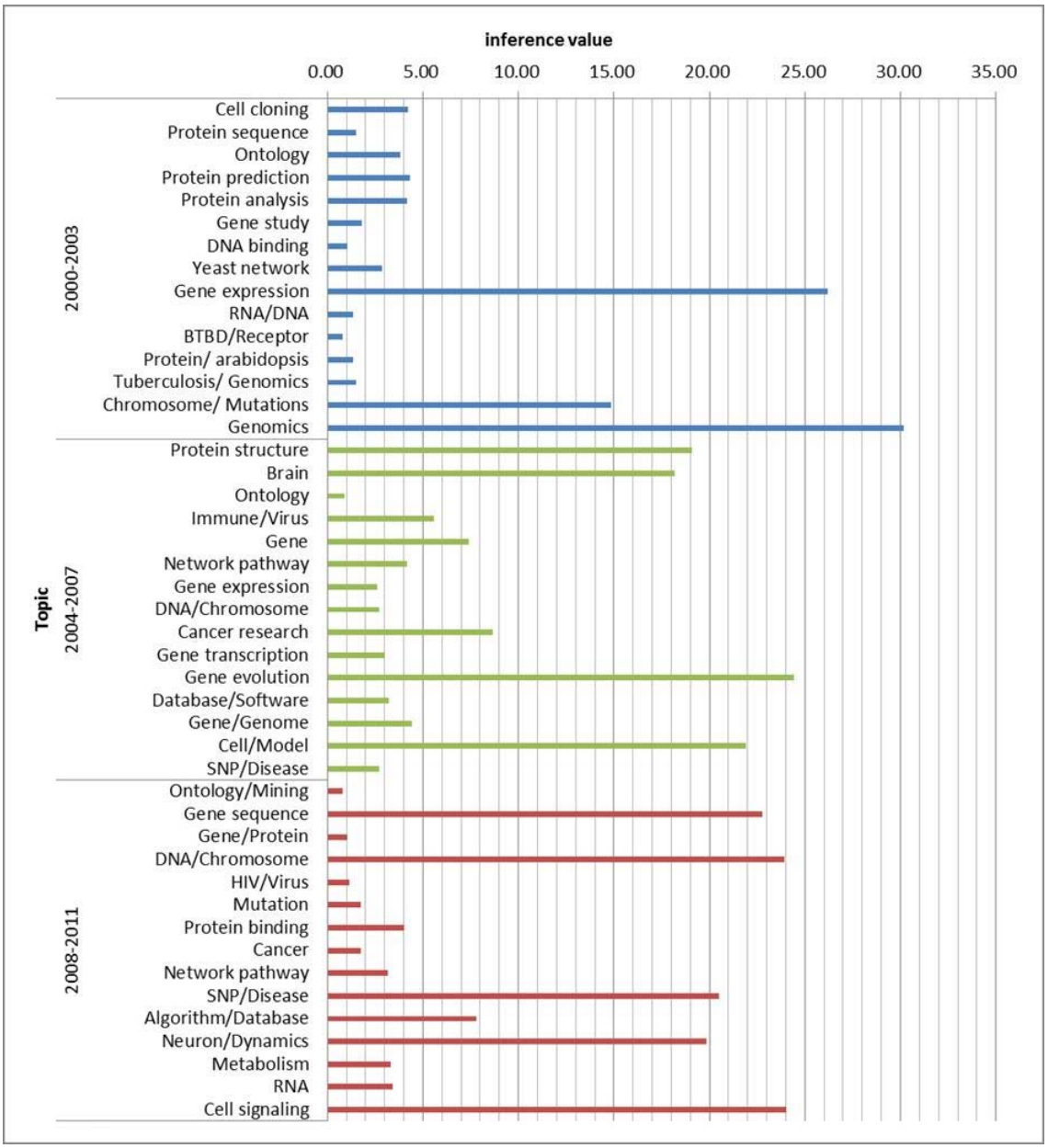

Figure 4. Topic distribution of papers of highly cited authors.

Influential authors

Appendix C shows the top 20 highly cited authors based on 546,245 citations from PubMed Central. In all three periods, M. Gerstein, a professor in computational biology and bioinformatics at Yale University, is both the most highly cited and productive author in the first author category. D. Botstein, a professor in molecular biology at Princeton University, is the most cited author in the second author category for the period, 2000 to 2003. He and his

group have excelled in the research of cellular growth rate in controlled circumstances. E. Lander is the most cited author, as second author, for the period, 2004 to 2007 and ranks third in the second author category and eighteenth in the first author category for the period, 2000 to 2003. He is a professor of Biology at the Massachusetts Institute of Technology, an expert in Genomics, and a collaborator with D. Botstein.

J. Storey is the second most cited author, as first author, for the period, 2000 to 2003. He is a professor of Molecular Biology at Princeton University and, with his research group, is actively involved in genomics. T. Speed is ranked third in the first author category, as well as, fifteenth in the second author category for the period, 2000 to 2003. He is an Australian statistician, who is well known for his contributions to the analysis of variance and bioinformatics. P. Bork is the second most cited author, as second author, in two consecutive periods (2004 to 2007 and 2008 to 2011). He is the head of the division of Bioinformatics at EMBL Heidelberg.

There are a couple of authors who have become more influential in the last two periods; R. Durbin, Joint Head of Human Genetics at the Wellcome Trust Sanger Institute and leader of the Genome Informatics group, ranks ninth in the second author category for the period, 2004 to 2007, and first for the period, 2008 to 2011, and D. Smith, professor of Molecular & Integrative Physiology at University of Michigan and co-director of the A. A. Taubman Consortium for Stem Cell Therapies, ranks third in the first author category for the period, 2004 to 2007, and eighth for the period, 2008 to 2011.

Emerging Stars

Appendix D shows the emerging top 20 researchers for the periods, 2004 to 2007 and 2008 to 2011, using a set notation of A - B for selecting emerging authors. In other words, the top 20 authors in Appendix D are selected when they do not appear in the previous period. For

example, L. Shi does not appear in the period, 2000 to 2003, but for the period, 2004 to 2007, receives 77 citations. He is affiliated with the US Food and Drug Administration and is involved in MicroArray Quality Control (MAQC). Other emerging authors during this time period include; H. Mermjakob, S. Toy, F. Spencer, and G. Smyth. For the period, 2008 to 2011, D. Goldstein, S. Guo, and W. Baumgartner look to be emerging researchers, but it is too early to evaluate their influence.

Influential journals or conferences

Table 10 illustrates leading the journals or conferences in bioinformatics. Analysis of a journal's citation count reveals that throughout the three time periods, journals such as the Proceedings of the National Academy of Sciences (Proc. Natl. Acad. Sci. U.S.A), Nucleic Acids Research (Nucleic Acids Res), Nature, Bioinformatics, and Science rank as the top five leading journals in bioinformatics.

BMC Bioinformatics, ranks sixth for the period, 2004 to 2007, and fifth for the period, 2008 to 2011. Among the top 20 journals, 11 journals are included in the top 20 for the entire period. 2000 to 2011. The EMBO Journal, Current Biology (Curr Biol), Trends in Genetics, and Journal of Bacteriology (J Bacteriol), which are in the top 20 for the period, 2000 to 2003, are not included among the top 20 journals past 2004. Journals such as the Journal of Molecular Biology, Genetic, Genes & Development (Genes Dev), Molecular and Cellular Biology (Mol Cell Biol), Molecular Biology and Evolution (Mol Biol Evol) are included in the top 20 for the period, 2004 to 2007, but their overall rankings decrease over time.

New journals emerging in the period, 2004 to 2007, are BMC Bioinformatics, PLoS Biology, BMC Genomics, and Nature Reviews Genetics (Nat Rev Genet). The rankings of these journals also increase during the period, 2008 to 2011. New journals such as PLoS One, PLoS Genetics, PLoS Computational Biology, Nature Biotechnology (Nat Biotechnol), and

Nature Methods (Nat Methods) are also included in the top 20 leading journals for the period, 2008 to 2011.

Table 10: Leading journals or conferences.

R	2000-2003		2004-2007		2008-2011	
	Journal	No.	Journal	No.	Journal	No.
1	Proc. Natl. Acad. Sci. U.S.A	12796	Nucleic Acids Res	14784	Nucleic Acids Res	3701
2	Nature	11718	Bioinformatics	12766	Nature	2684
3	Nucleic Acids Res	10438	Proc. Natl. Acad. Sci. U.S.A	12718	Proc. Natl. Acad. Sci. U.S.A	2425
4	Science	10174	Nature	10891	Bioinformatics	2220
5	Bioinformatics	8433	Science	8647	BMC Bioinformatics	1866
6	Genome Res	6955	BMC Bioinformatics	7260	Science	1861
7	Nat Genet	5816	Genome Res	5903	Nat Genet	1486
8	J Biol Chem	5266	Cell	5412	Genome Res	1442
9	Cell	3835	Nat Genet	5192	BMC Genomics	1377
10	Journal of Molecular Biology	3689	J Biol Chem	4671	PLoS One	1275
11	Genome Biology	3145	Genome Biology	4299	Cell	1194
12	Genetics	2477	PLoS Biology	3219	PLoS Genetics	1131
13	The EMBO Journal	2169	Genetics	2804	PLoS Computational Biology	934
14	Genes Dev	2112	Journal of Molecular Biology	2468	Genome Biology	818
15	Mol Cell Biol	1972	Mol Biol Evol	2426	J Biol Chem	772
16	Curr Biol	1935	BMC Genomics	2417	PLoS Biology	669
17	Trends in Genetics	1837	Genes Dev	2196	Nat Biotechnol	579
18	Mol Biol Evol	1730	Nat Rev Genet	2182	Nat Rev Genet	557
19	Mol Cell	1691	Mol Cell	2029	Nat Methods	553
20	J Bacteriol	1675	Mol Cell Biol	1992	Mol Cell	494

Co-authorship Analysis

In this section, we attempt to understand the knowledge structure of the field of bioinformatics using co-authorship analysis of the 2,088,356 co-author pairs. Since this network is too big to either analyze or visualize, we focus our analysis on authors collaborating with more than 30 colleagues, which consists of 13,952 pairs. We identify 15 communities of the co-author networks using the modularity algorithm widely used in Social

Network Analysis. We use modality to examine how strongly the groups of the co-author networks are structured. Networks with high modularity tend to show a dense connection between the nodes within groups, whereas networks with slow modularity show sparse connections between nodes in different groups (Newman, 2006). To calculate modularity, we use the open source visualization program called JUNG (http://jung.sourceforge.net/). Table 11 shows the characteristics of these communities. The biggest community has a general research interest of genomics and includes 2,151 authors, which is about 23.5% of authors on the co-author network. The second biggest community has 1,269 authors, with top ranked authors A.G. Uitterlinden, H. Wichmann, and T.D. Spector (0.3095, 0.3084, 0.3084 respectively), and a general research interest in genetics. The third biggest community has 1,118 authors, with top ranked authors S.L. Salzberg, J.A. Eisen, and P. Flicek (0.2873, 0.2946, 0.2723 respectively) and a general research interest of computational biology. The top ranked authors in terms of closeness centrality are Y. Li (0.38) in community 2, D.J.Huter (0.33) in community 15, N. Chatterjee (0.32) in community 1, and S.J.Chanock (0.32) in community 15.

Table 11: Community statistics by modularity.

Community	Topic	Size	%	Top Ranked Author	Degree	Closeness Centrality
1	Genomics	653	7.13	N.G.Martin	117	0.31
				N.Chatterjee	104	0.32
				K.V.Shianna	91	0.27
2	Genomics	2151	23.5	Y.Li	351	0.38
				L.Shi	114	0.31
3	Protein and RNA Sequences	498	5.44	R.D.Finn	70	0.26
				A.Bateman	65	0.25
				E.W.Deutsch	55	0.23
4	Software	216	2.36	J.Anderson	138	0.25
				C.Nguyen	131	0.25
				C.Gonzalez	128	0.25
5	Biomedical Text Mining	271	2.96	W.J. Wilbur	6	0.25
				A. Valencia	5	0.25

				C. Blachke	5	0.23
6	Gene regulation/ Sequence	907	9.91	J.Aerts	125	0.27
				M.A.Quail	96	0.26
				P.J.deJong	82	0.29
7	Functional Genomics	538	5.88	M.Nakao	209	0.29
				P.Carninci	201	0.30
				Y.Hayashizaki	170	0.30
8	Molecular Biology	281	3.07	A.Poustka	321	0.30
				R.Holt	215	0.29
				A.Prasad	205	0.28
9	Computational Biology	1118	12.21	S.L.Salzberg	149	0.29
				J.A.Eisen	139	0.29
				P.Flicek	105	0.27
10	Algorithm	128	1.4	M.Vidal	61	0.25
				A.Oliveira	60	0.20
11	Genetics	1269	13.86	A.G.Uitterlinden	305	0.31
				H.Wichmann	256	0.31
				T.D.Spector	244	0.31
12	System Biology	50	0.55	A.A.Sharov	53	0.27
				Y.Piao	49	0.27
				D.L.Longo	5	0.27
13	Computational Biology	114	1.25	A.Helgason	77	0.26
				G.Hallmans	76	0.28
				U.Styrkarsdottir	68	0.30
14	Evolutionary Genomics	542	5.92	V.Barbe	114	0.25
				J.Johnson	94	0.30
				P.Wincker	89	0.27
15	Functional Genomics	417	4.56	S.J.Chanock	319	0.32
				D.J.Hunter	299	0.33

By visualizing the co-authorship network, we are able to map the topology of the bioinformatics field (Figure 5). Figure 5 illustrates that the major driving force of bioinformatics research is genomics related. The neighbor fields to genomics are gene regulation and sequence, protein and RNA sequence, system biology, and genetics. Figure 5 also denotes that the computational side of the field, such as software and algorithms, is located a distance from the main driving force.

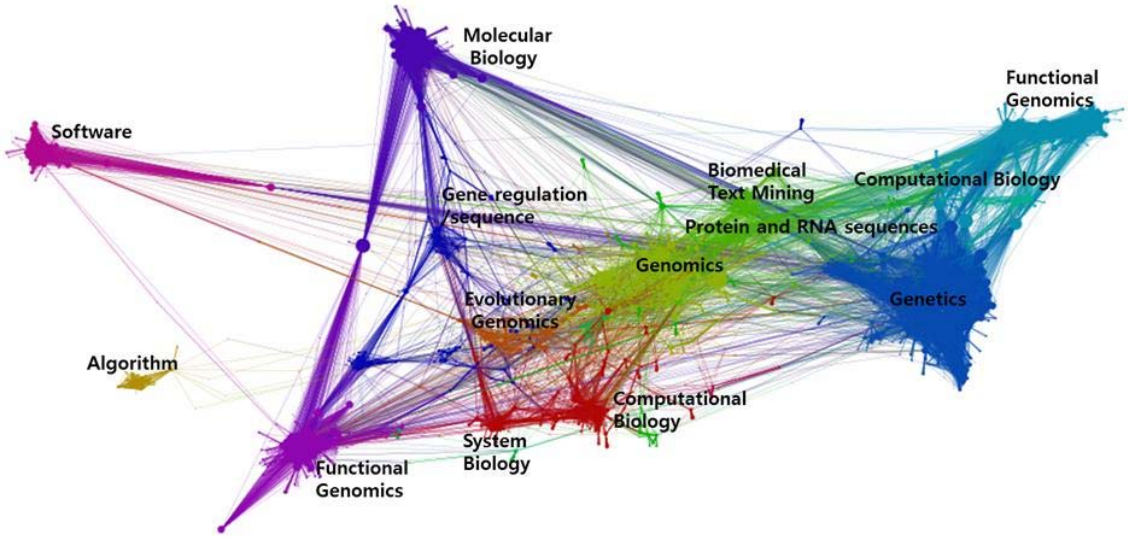

Figure 5. Visualization of author collaboration map in bioinformatics.

Conclusion

In this paper, we conducted a bibliometric analysis of the bioinformatics field using PubMed Central data. Citations were extracted from full-text articles for the period, 2000 to 2011, which were included in PubMed Central collections. Both productivity and impact of the bioinformatic community were analyzed, notably for three periods within the last decade; 2000 to 2003, 2004 to 2007, and 2008 to 2011. For productivity, four measures were used: most productive authors, most productive countries, most productive organization, and most popular subject terms. The most productive first authors were Michael L. Gross, G.A. Petsko, and R. Robinson. In the second author category, E.V. Koonin, Y. Hayashizaki, and P.E. Bourne were the emerging, productive authors. The most productive countries were the U.S.A., the U.K., and Germany. The most productive organizations were Stanford University, Harvard University, the University of California at San Diego, and the University of Washington. The most popular subject terms were Molecular Biology, Medical Informatics, Genetics, and Biology.

Research impact was analyzed based on citation counting. To measure influence, we looked at the following aspects: most cited papers, most cited authors, emerging stars, and leading organizations. For most cited papers, in the entire period,2000 to 2011, we identified the following four: “Gene ontology: tool for the unification of biology,” “Initial sequencing and analysis of the human genome,” “Bioconductor: open software development for computational biology and bioinformatics,” and “R: A language and environment for statistical computing.” M. Gerstein, D. Botstein, and E. Lander were ranked as the top three authors. Upon observing that a few productive authors (G.A. Petsko and J. Wixon) in the period, 2000 to 2003 were not included in the influential authors; we further examined those authors and realized their papers were not research oriented papers, but one page long essays or review papers. This implies that productivity should be considered with the impact measure to evaluate an author’s research performance. For the emerging influential authors, L. Shi, H. Hermjakob, and S. Roy were identified for the period, 2004 to 2007 and D. Goldstein, S. Guo, and W. Baumgartnerfor the period, 2008 to 2011. The highly cited journals and conferences were Proc. Natl. Acad. Sci. U.S.A, Nucleic Acids Research (Nucleic Acids Res), Nature, Bioinformatics, and Science.

The results of productivity and influence analysis indicate that the field of bioinformatics has undergone a significant shift to co-evolve with other biomedical disciplines and that the topical focus has shifted over time. We observed that the growth of computational approaches has facilitated the proliferation of biological databases and methods within various biomedical disciplines, which has become an early driving force for the development of bioinformatics. We found that the use of computational methods became prevalent across biomedical disciplines in the period 2000 to 2003, while the use and application of biological databases have been rapidly increasing since 2004. In addition, we observed that the field of bioinformatics contributed to the wide adoption of molecular sequence databases in

biomedicine, and that microarray analysis and biological network modeling became two major new topics emerging in the bioinformatics community.

Overall, trends between the periods, 2000 to 2003 and 2004 to 2007, were dissimilar, while trends between the period, 2004 to 2007 and 2008 to 2011, were similar. This coincides with the exponential publication growth since 2004. In the collection of test data, the rate of publication increased 114% with the period 2003 to 2004 (499 publications in 2003 and 1068 in 2004). Since then, the rate of increase has been more than 50%. Except for three authors, the top 15 authors included for the period, 2000 to 2003 did not appear in the period, 2004 to 2011. Among the top 20 countries for the period, 2000 to 2003, four countries were not included in the top 20 for the period, 2004 to2011. Among leading organizations for the period, 2000 to2003, eight organizations were not included for the period, 2004 to 2011.

Our future research will include comparing results reported in this paper with citation analysis of Web of Science data to investigate how the field of bioinformatics is represented by PubMed Central. We will also use social network analysis to detect research groups or communities in this field. In addition, a follow-up study will be conducted to identify the knowledge diffusion and transfer patterns in this field using content-based citation analysis. For author name disambiguation, we intend to explore more comprehensive methods to disambiguate author names such as a probabilistic method proposed by Tang et al. (2012).

References

Adams, M. D., Celniker,S. E., Holt, R. A., Evans, C.A., Gocayne, J.D., Amanatides, P.G.,. . . Venter, J.C. (2000). The genome sequence of Drosophila melanogaster. *Science*, 287(5461), 2185-95.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17), 3389-3402.

Baldi, P & Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach, 2nd edition*. MIT Press.

Bansard, J.Y., Rebholz-Schuhmann, D., Cameron, G., Clark, D., van Mulligen, E., Beltrame, F., . . . Coatrieux, J.L. (2007). Medical informatics and bioinformatics: a bibliometric study. *IEEE transactions on information technology in biomedicine*,11(3), 237-243.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1033.

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H-D. (2008). Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics*, 8, 93-102.

Boyack, K. W., Klavans, A. R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351-374.

Brown, L.D., & Gardner, J.C. (1985). Using citation analysis to assess the impact of journals and articles on contemporary accounting research (CAR). *Journal of Accounting Research*, 23(1), 84-109.

Cohen, W.W., Ravikumar, P.D., & Fienberg, S.E. (2003). A comparison of string distance metrics for name-matching tasks. *IIWeb*, 2003, 73-78.

Cronin, B. (1981). The need for a theory of citing. *Journal of documentation*, 37(1), 16-24

Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. Taylor Graham: London.

Cronin, B., & Overfelt, K. (1993). Citation-based auditing of academic performance. *Journal of the American Society for Information Science,* 45(2), 61–72.

Cronin, B. (2004). Normative shaping of scientific practice: the magic of Merton. *Scientometrics*, 60(1), 41-46.

Ding, Y., Chowdhury, G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management,* 37(6), 817-842.

Ding, Y. (2010). Semantic web: Who is who in the field - a bibliometric analysis. *Journal of Information Science,* 36(3), 335-356.

Ding, Y. (2011). Applying weighted PageRank to author citation networks. *Journal of the American Society for Information Science and Technology*, 62(2), 236-245.

Ding, Y., & Cronin, B. (2011). Popular and/or prestigious? Measures of scholarly esteem, *Information Processing and Management*, 47(1), 80-96.

Garfield, E. (1955). Citation indexes to science: A new dimension in documentation through association of ideas. *Science*, 122(3159), 108–111.

Garfield, E. (2000). Use of journal citation reports and journal performance indicators in measuring short and long term journal impact. *Croatian Medical Journal*, 41(4), 368–374.

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.

Glänzel, W., Thijs, B., Schubert, A., & Debackere, K. (2009a). Subfield specific normalized relative parameters and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics*, 78(1), 165–188.

Guan, J., & Gao, X. (2008). Comparison and evaluation of Chinese research performance in the field of bioinformatics. *Scientometrics*, 75(2), 357–379.

He, B., Ding, Y., & Ni, C. (2011). Mining enriched contextual information of scientific collaboration: A meso perspective. *Journal of the American Society for Information Science and Technology*, 62(5), 831-845.

He, B., Ding, Y., & Yan, E. (2012). Mining patterns of author orders in scientific publications. *Journal of Informetrics*, 6(3), 359-367.

Huang, H., Andrews, J., & Tang, J. (2012). Citation characterization and impact normalization in bioinformatics journals. *Journal of the American Society of Information Science and Technology*, 63(3), 490-497.

Janssens, F., Glanzel, W., & De Moor, B. (2007). Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp.360-369), August, San Jose, California,

Manoharan, A., Kanagavel, B., Muthuchidambaram, A., & Kumaravel, J.P.S. (2011). Bioinformatics research – an informetric view. *2011 International Conference on Information Communication and Management* (pp. 199-204). Singapore: IACSIT Press

McCain, K.W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, 41(6), 433-443.

McCallum, A. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Mimno, D. M. & McCallum, A. (2008). Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression. UAI 2008: 411-418

Moed, H.F. (2005). Citation analysis of scientific journals and journal impact measures. *Current Science*, 89, 1990–1996.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of National Academy of Sciences*, 103(23): 8577–8696.

Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology*, 41(1), 609-641.

Patra, S. K. & Mishra, S. (2006). Bibliometric study of bioinformatics literature. *Scientometrics*, 67 (3), 477–489.

Ramsden, P. (1994). Describing and explaining research productivity. *Higher Education*, 28(2), 207-226.

Sekercioglu, C.H., (2008). Quantifying coauthor contributions. *Science*, 322(5900), 371

Small, H. (1976). Structural dynamics of scientific literature. *International Classification*, 3(2), 67-74.

Small, H. (2011). Interpreting maps of science using citation context sentiments: A preliminary investigation. *Scientometrics*, 87, 373-388.

Tang, J., Fong, A.C.M., Wang, B., & Zhang, J. (2012). A Unified Probabilistic Framework for Name Disambiguation in Digital Library. *IEEE Transaction on Knowledge and Data Engineering (TKDE)*, 24(6), 975-987.

Yan, E., & Ding, Y. (2010). Weighted citation: An indicator of an article's prestige. *Journal of the American Society for Information Science and Technology*, 61(8), 1635–1643.

Yan, E., & Sugimoto, C. R. (2011). Institutional interactions: Exploring the social, cognitive, and geographic relationships between institutions as demonstrated through citation networks. *Journal of the American Society for Information Science and Technology*, 62(8), 1498-1514.

Yang, C., Tang, X., & Song, M. (2013). Understanding the development across scientific research domains using a hybrid analysis, forthcoming *Journal of the American Society for Information Science and Technology.*

APPENDIX A: Top 15 most productive authors.

	2000-2003		2004-2007		2008-2011	
R	First author	Second author	First author	Second author	First author	Second author
1	G.A. Petsko 42	E.V. Koonin 11	L. Gross 124	Y. Hayashizaki 23	R. Robinson 28	P.E. Bourne 17
2	J. Wixon 13	L. Aravind 10	G.A. Petsko 49	P. Carninci 21	G.A. Petsko 23	J. Feng 17
3	V. Anantharaman 5	A. Valencia 8	R. Robinson 39	P.O. Brown 17	C. Sedwick 21	L. Peltonen 17
4	S. Brenner 5	P.O. Brown 7	M. Hoff 19	P.E. Bourne 14	J. Gitschier 16	J. Nielsen 16
5	C. Blaschke 5	R. Apweiler 5	F. Chanut 17	M.B. Eisen 14	R. Meadows 14	M.B. Gerstein 16
6	C.A. Semple 4	C.A. Ouzounis 5	H. Parthasarathy 10	D.R. Flower 13	L. Gross 13	B.Ø. Palsson 16
7	S. Oliver 4	L. Wang 5	P.E. Bourne 9	N. Barkai 12	R. Jones 10	Y. Hayashizaki 16
8	L.M. Iyer 4	J. Hinds 4	R. Jones 9	Y.Li 12	D.G. Nathan 10	C.M. van Duijn 15
9	K.C. Woodwark 4	I.B. Rogozin 4	M. Inman 9	S. Pääbo 10	M. Hoff 9	P. Wincker 15
10	L. Aravind 3	G.M. Rubin 4	J. Gitschier 8	M. Tomita 10	J. Bohlin 7	T.D. Spector 15
11	G. Xie 3	K. Hashimoto 4	P.D. Taylor 8	S.L. Salzberg 9	K. Heller 7	
12	K.S. Makarova 3	Y.I. Wolf 4	H. Nicholls 6	M. Gerstein 9	A. Sharma 6	P. Deloukas 14
13	M. Crossley 3	R.A. Jensen 4	V. Gewin 5	S.G. Oliver 9	R. Gupta 5	H. Wichmann

						14
14	E.V. Koonin 3	O.K. Pickeral 3	R. Gowthaman 5	J.A. Eisen 9	Y. Sun 5	A.J. Butte 14
15	J.C. Rockett 3 J.M. Bujnicki 3 D.A. Liberles 3	L. Rychlewski 3 G. Kelsoe 3 S.L. Forsburg 3 P. Bork 3 H. Reichert 3 S.E. Celniker 3 M. Tyers 3 S.W. Scherer 3 J. Greene 3 M. Gerstein 3 R.A. Gibbs 3 R. Gonzalez-Duarte 3 C.A. Bonner 3	L. Kashyap 5 P.R. Painter 5	B.Ø. Palsson 9 T.K. Attwood 9	X.He 5 S. Ranganathan 5 W. Mair 5	E.E. Schadt 13 G.P. Raghava 13 O. Kohlbacher 13 M. Mann 13 E. Barillot 13 S. Ranganathan 13 E. Ruppin 13 A.G. Uitterlinden 13

APPENDIX B: Highly cited paper.

	2000-2003			2004-2007			2008-2011		
R	paper	journal	no. cited	paper	journal	no. cited	paper	journal	no. cited
1	Gene ontology: tool for the unification of biology. The Gene Ontology Consortium	Nat Genet	948	Bioconductor: open software development for computational biology and bioinformatics	Genome Biol	395	The Pfam protein families database	Nucleic Acids Res	112
2	Initial sequencing and analysis of the human genome	Nature	465	R: A language and environment for statistical computing	R: A language and environment for statistical computing	304	KEGG for linking genomes to life and the environment	Nucleic Acids Res	104
3	Significance analysis of microarrays	Proc Natl Acad Sci USA	349	Transcriptional regulatory code of a	Nature	234	Mapping short DNA sequencing	Genome Res.	91

	applied to the ionizing radiation response			eukaryotic genome			reads and calling variants using mapping quality scores		
4	The Protein Data Bank	Nucleic Acids Res	341	Linear models and empirical bayes methods for assessing differential expression in microarray experiments	Stat Appl Genet Mol Biol	222	miRBase: tools for microRNA genomics	Nucleic Acids Res	80
5	Cytoscape: A software environment for integrated models of biomolecular interaction networks	Genome Res	325	Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles	Proc Natl Acad Sci USA	217	Mapping and quantifying mammalian transcriptomes by RNA-Seq.	Nat Methods	71
6	Initial sequencing and comparative analysis of the mouse genome	Nature	287	MUSCLE: Multiple sequence alignment with high accuracy and high throughput	Nucleic Acids Res	215	Database resources of the National Center for Biotechnology Information	Nucleic Acids Res	65
7	Exploration, normalization, and summaries of high density oligonucleotide array probe level data	Biostatistics	267	Genome sequencing in microfabricated high-density picolitre reactors	Nature	193	Ensembl 2008	Nucleic Acids Res	65
8	BLAT--the BLAST-like alignment tool	Genome Res	235	A haplotype map of the human genome	Nature	186	Accurate whole human genome sequencing	Nature	64

							using reversible terminator chemistry		
9	A comparison of normalization methods for high density oligonucleotide array data based on variance and bias	Bioinformatics	221	WebLogo: A sequence logo generator	Genome Res	179	Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources	Nat Protoc	57
10	A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae	Nature	217	A gene atlas of the mouse and human protein-encoding transcriptomes	Proc Natl Acad Sci U S A	174	Alternative isoform regulation in human tissue transcriptomes	Nature	56
11	The sequence of the human genome	Science	216	The Pfam protein families database	Nucleic Acids Res	172	The UCSC Genome Browser Database: 2008 update	Nucleic Acids Res	56
12	Transcriptional regulatory networks in Saccharomyces cerevisiae	Science	216	Network biology: Understanding the cell's functional organization	Nat Rev Genet	167	The complete genome of an individual by massively parallel DNA sequencing	Nature	54
13	David: Database for annotation, visualization, and integrated discovery	Genome Biol	212	MicroRNAs: Genomics, biogenesis, mechanism, and function	Cell	162	SOAP: short oligonucleotide alignment program	Bioinformatics	53
14	Functional organization of the yeast proteome by	Nature	208	Proteome survey reveals modularity	Nature	157	Ultrafast and memory-efficient alignment of	Genome Biol.	51

	systematic analysis of protein complexes			of the yeast cell machinery			short DNA sequences to the human genome		
15	EMBOSS: the European Molecular Biology Open Software Suite	Trends Genet	207	Global landscape of protein complexes in the yeast Saccharomyces cerevisiae	Nature	153	The transcriptional landscape of the yeast genome defined by RNA sequencing	Science	48
16	Genomic expression programs in the response of yeast cells to environmental changes	Mol Biol Cell	207	The Gene Ontology (GO) database and informatics resource	Nucleic Acids Res	149	Ensembl 2009	Nucleic Acids Res	46
17	KEGG: Kyoto Encyclopedia of Genes and Genomes	Nucleic Acids Res	206	Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project	Nature	146	Widespread changes in protein synthesis induced by microRNAs	Nature	46
18	Summaries of Affymetrix GeneChip probe level data	Nucleic Acids Res	200	NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins	Nucleic Acids Res	145	The BioGRID Interaction Database: 2008 update	Nucleic Acids Res	44
19	Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry	Nature	190	Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared	Nature	144	Velvet: algorithms for de novo short read assembly using de Bruijn graphs	Genome Res	44

				controls					
20	Primer3 on the WWW for general users and for biologist programmers	Bioinformatics Methods and Protocols: Methods in Molecular Biology	187	MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment	Brief Bioinform	141	Highly integrated single-base resolution maps of the epigenome in Arabidopsis.	Cell	42

APPENDIX C: Highly cited first and second authors.

R	2000-2003				2004-2007				2008-2011			
	first author	no. cited	second author	no. cited	first author	no. cited	second author	no. cited	first author	no. cited	second author	no. cited
1	M. Gerstein	168	D. Botstein	2752	M. Gerstein	161	E. Lander	1391	M. Gerstein	54	R. Durbin	421
2	J. Storey	160	P. Bork	1933	J. Yates	114	P. Bork	1317	M. Mann	35	E. Birney	295
3	T. Speed	157	E. Lander	1835	D. Smith	109	R. Apweiler	1250	A. Wagner	34	M. Snyder	211
4	E. Koonin	151	P. Brown	1812	M. Mann	94	E. Birney	1091	B. Palsson	33	E. Eichler	206
5	R. Tibshirani	125	G. Rubin	1735	A. Wagner	87	R. Irizarry	1085	M. Kuhn	32	R. Finn	198
6	B. Palsson	103	G. Sherlock	1651	B. Palsson	87	W. Kent	1036	R. Stevens	29	T. Hubbard	195
7	D. Swofford	100	E. Koonin	1558	L. Serrano	82	M. Gerstein	1032	M. Vingron	27	P. Flicek	189
8	J. Yates	100	S. Lewis	1510	B. Smith	81	D. Haussler	1025	D. Smith	26	E. Mardis	185
9	A. Brazma	94	W. Kent	1489	L. Shi	77	R. Durbin	925	Y. Guo	26	J. Ruan	177
10	M. Vingron	94	E. Birney	1464	J. Storey	76	D. Wheeler	907	M. Ritchie	24	D. Wheeler	176
11	S. Henikoff	86	M. Ashburner	1459	H. Hermjakob	72	G. Smith	904	E. Ruppin	24	A. Batemann	176
12	R. Russell	80	R. Tibshirani	1436	P. Bourne	70	B. Palsson	897	H. Hermjakob	23	S. Jones	171
13	G. Churchill	80	D. Haussler	1426	S. Mukherjee	70	R. Gibbs	880	J. Yates	23	J. Vogel	170
14	A. Wagner	72	J. Cherry	1403	R. Russell	69	M. Vidal	857	R. Breaker	23	R. Gibbs	169
15	M. Mann	69	T. Speed	1386	S. Oliver	67	M. Daly	841	E. Koonin	23	J. Smith	167
16	D. Smith	68	C. Ball	1361	J. Reed	67	T. Consortium	826	S. Jones	22	B. Ballester	163
17	I. Kohane	66	J. Matese	1308	S. Roy	66	S. Griffiths-Jones	819	A. Millar	22	E. Kulesha	162
18	E. Lander	65	S. Eddy	1225	A. Brazma	66	D. Bartel	811	A.	21	P. Bork	159

									Johnson			
19	D. Jones	64	H. Butler	1211	J. Johnson	65	R. Gentleman	788	O. Keskin	21	E. Sonnhammer	159
20	J. Johnson	64	M. Harris	1209	R. Stevens	64	R. Edgar	777	N. Barkai	21	S. Haider	155
					W. Noble	64			J. Johnson	21		
					D. Rhodes	64						

APPENDIX D: Emerging stars.

R	2004-2007 New Stars		2008-2011 New Stars	
	Name	Citation Count	Name	Citation Count
1	L. Shi	77	D. Goldstein	18
2	H. Hermjakob	72	S. Guo	18
3	S. Roy	66	W. Baumgartner	18
4	F. Spencer	55	P. Kharchenko	17
5	G. Smyth	53	R. Nussinov	17
6	X. Guo	51	T. Manolio	17
7	B. Shapiro	50	F. Leitner	16
8	K. Strimmer	49	M. Pop	16
9	D. Robertson	48	S. Cheung	16
10	E. Ruppin	47	T. Gibson	16
11	M. Bauer	47	M. Brylinski	15
12	S. Wilhite	46	A. Dunker	14
13	Y. Guo	46	B. Ge	14
14	M. Cortese	44	C. Croce	14
15	C. Myers	40	H. Saini	14
16	G. Ast	39	J. Shendure	14
17	L. Hunter	39	M. Tasan	14
18	L. Pachter	39	P. Froguel	14
19	D. Kell	38	R. Nilsson	14
20	P. Tompa	38	S. Horvath	14
			Y. Nikolsky	14