Using A Computational Approach to Study the History of Systems Biology: From Systems to Biology, 1992-2013

by

Yawen Zou

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2016 by the
Graduate Supervisory Committee:

Manfred Laubichler, Co-Chair
Jane Maienschein, Co-Chair
Richard Creath
Karin Ellison
Stuart Newfeld

ARIZONA STATE UNIVERSITY

August 2016

ABSTRACT

Systems biology studies complex biological systems. It is an interdisciplinary field, with biologists working with non-biologists such as computer scientists, engineers, chemists, and mathematicians to address research problems applying systems' perspectives. How these different researchers and their disciplines differently contributed to the advancement of this field over time is a question worth examining. Did systems biology become a systems-oriented science or a biology-oriented science from 1992 to 2013?

This project utilized computational tools to analyze large data sets and interpreted the results from historical and philosophical perspectives. Tools deployed were derived from scientometrics, corpus linguistics, text-based analysis, network analysis, and GIS analysis to analyze more than 9000 articles (metadata and text) on systems biology. The application of these tools to a HPS project represents a novel approach.

The dissertation shows that systems biology has transitioned from a more mathematical, computational, and engineering-oriented discipline focusing on modeling to a more biology-oriented discipline that uses modeling as a means to address real biological problems. Also, the results show that bioengineering and medical research has increased within systems biology. This is reflected in the increase of the centrality of biology-related concepts such as cancer, over time. The dissertation also compares the development of systems biology in China with some other parts of the world, and reveals regional differences, such as a unique trajectory of systems biology in China related to a focus on traditional Chinese medicine.

This dissertation adds to the historiography of modern biology where few studies have focused on systems biology compared with the history of molecular biology and evolutionary biology.

ACKNOWLEDGMENTS

I am very grateful for the help of my advisor Dr. Manfred Laubichler because he has been very supportive all the time and led me intellectually into the world of the computational history and philosophy of science. I am also grateful for another co-chair Dr. Jane Maienschein, who has given me detailed feedback on my drafts and been very patient explaining things. I learned from them the dedication to academia. I also want to thank my other committee members: Dr. Karin Ellison, Dr. Richard Creath, and Dr. Stuart Newfeld, who have been all very helpful in different stages.

I have encountered many excellent teachers at ASU. I want to thank Dr. Michael Simeon and Jacqueline Hettel from the Institute for Humanities Research Nexus Lab, who taught me digital humanities techniques and allowed me to use the resources in their lab, and Dr. Michael Rosenberg, who taught me Python, which has helped my research a lot. I want to say thanks to the staff at the Center for Biology and Society, especially Jessica Ranney, who helped me deal with many challenges with paperwork.

I am lucky to work with Erick Peirson, Julia Damerow, Ken Aiello, Deryc Painter, and Bryan Daniels in my lab. There is a famous saying by Confucius that “If three of us are walking together, at least one of the other two is good enough to be my teacher.” They all gave me great ideas for my research. Adam Staples worked as a student worker in the lab and helped me with my data collection. My best friends Federica Colonna and Lijing Jiang cared about me a lot too. Other graduate students in the center, including Mark Ulett, Erica O’Neil, Valerie Racine, Wes Anderson, Aireona Raschke, Kate MacCord, and Steve Elliot, have helped with my manuscript preparation, presentation skills, giving feedback, and job searching.

I also need to say thank you to my parents Rimin Zou and Lianxiang Wang, who have always been supportive of me. I have met with many good people at ASU and I will remember the wonderful times we spent together. They may not contribute to this dissertation *per se*, but they have enriched my life and provided me emotional support.

Finally, I want to thank my husband Yi Fang. The dissertation is the output of my five years' intellectual work, and there were some difficult moments. Doing a PhD in another language is hard for both me and him, and we gave each other mutual support. I am grateful for it.

TABLE OF CONTENTS

Page

APPENDIX Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1: INTRODUCTION

1.1. Context/Motivation

Systems biology is a new branch of biology. Leroy Hood (2003), the director and founder of the Institute for Systems Biology (ISB) in Seattle, Washington, for example, defined it as a field "studying the interrelationships of all of the elements in a system rather than studying them one at a time" (p. 9). Many scientists believe that systems biology has great potential for health care and could overcome the limitation of reductionism (Kitano, 2002; Hood, Heath, Phelps & Lin, 2004). It has experienced rapid growth because of the invention of high-throughput technologies and computational modeling. However, it is hard to define what exactly systems biology is, as one finds it hard to define molecular biology, because they represent two different ways to approach biology (Powell & Dupré, 2009) rather than well defined disciplines in the traditional sense.

Systems biology is even harder to define precisely because it is even more interdisciplinary than molecular biology (Calvert & Fujimura, 2011). Scientists from different disciplines besides biology, computer science, engineering, physics, and mathematics, to name a few, have variously contributed to its methodology and epistemology. To give the readers an initial understanding of systems biology, I represented this scientific field through a word cloud, which is a graphical representation of word frequency. In a word cloud, the size of a word is proportional to its frequency in a text corpus, which can be used as a proxy for its importance. I used the Paper Machine

to generate a word cloud[1] based on 100 randomly sampled systems biology articles from the Web of Science (WoS) database. Figure 1 shows the word cloud of systems biology, in which words like "data," "model," "network," "metabolic," "information," and "genome" have a big font, because they appear more often. Many of these words are words that would appear in non-biology fields such as computer science or engineering. The goal of this study of the history of systems biology is to see how different disciplines helped shape its development. The motivation for this dissertation as a whole comes from the pilot research project detailed below.

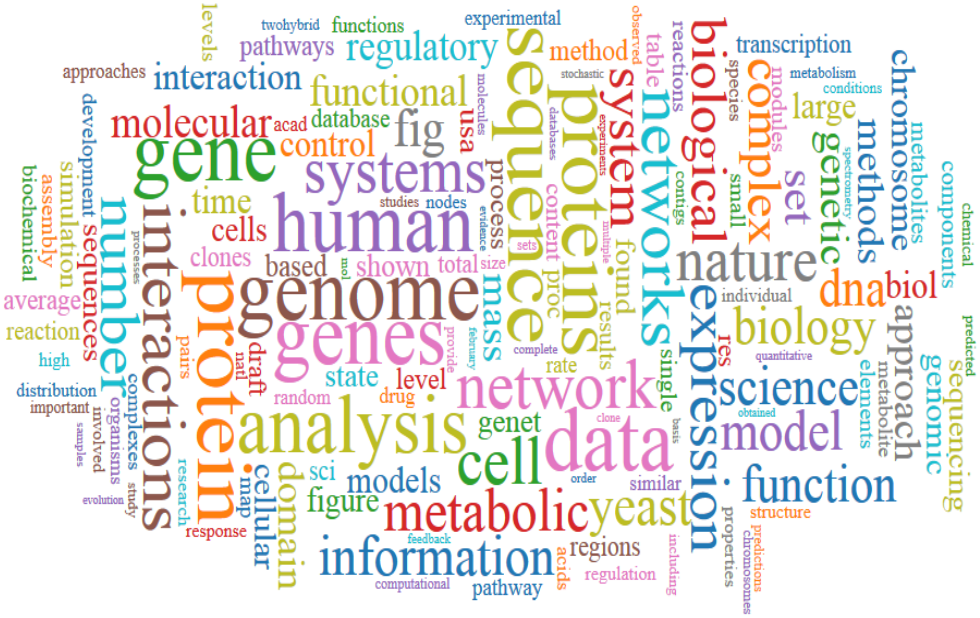


Figure 1. Word cloud of systems biology showing the frequently used words.

1.1.1. Initial research in the Web of Science

1 For more information about Paper Machines, see http://papermachines.org. Paper Machines is a plug in for Zetero, a bibliographic tool. It is designed for humanities researchers to make use of computer science analysis.

To get a better understanding of systems biology, I collected a representative sample of the literature on the topic of "systems biology." A search in the WoS database for articles published between 1900 and 2013 that contain the term "systems biology" in their "topics" (including "titles," "abstracts," and "keywords") returned 9923 articles (The data was collected on February 3, 2014). The language was confined to English. Figure 2 illustrates this rapid development of the field. In the years before 2000, there were fewer than 10 articles listed in the Wo*S* database. From 1997 to 2001, the numbers of articles were 1, 1, 1, 4, 7, but that is not shown in this figure. However, there were 1480 articles published in the year 2012 alone. The number of papers on "systems biology" from 1992 to 2013 is increasing throughout the years. This initial finding is similar to other historians' accounts that the rapid development of systems biology only happened after 2000, but that it is one of the fastest-growing areas of biology (Powell, O'Malley, Müller-Wille & Dupré, 2007).

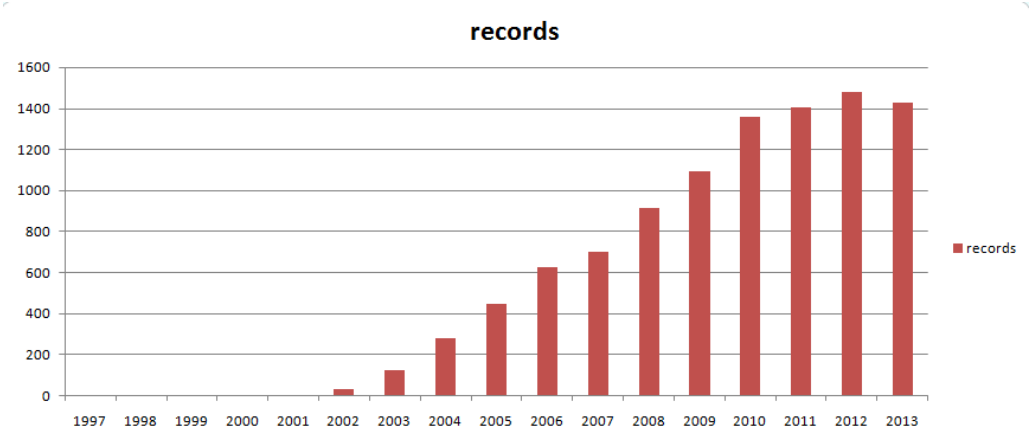


Figure 2. The growth of systems biology literature listed in the WoS. This figure shows the number of articles for each year for the 9923 articles. The *x* axis denotes the year, and the *y* axis denotes the number of articles.

Next, I used the bibliometric tool Citespace[2] to analyze the citation data of the 9923 articles downloaded from the WoS and generated a list (Table 1) of the most highly cited authors to see who have been influential in this field (In my next chapter, I will discuss in detail what Citespace is, and the different types of citation analysis it can carry out). The citation analysis with CiteSpace generates an excel sheet that shows for each author how many articles of my sample have cited his/her work. For example, for an author, if 3 articles out of 9923 articles cite that author's works, the citation count for that author is 3. Table 1 lists the information of their rankings, citation counts calculated by Citespace, author names, affiliations and occupations.

I then researched the most highly cited authors' affiliations and educational background. This analysis led to an interesting observation. From Table 1 one can see that the top five most highly cited authors have a background in engineering, physics, mathematics, or computational science, or a combination of both engineering and biology backgrounds. For example, the first ranking author, Hiroaki Kitano, was an engineer at Sony before becoming a systems biologist. The second ranking author, Minoru Kanehisa holds a PhD in physics, but later became a bioinformatician specializing in building databases such as KEGG, an online database which stores information about genomes, pathways, and biological chemicals (Kanehisa & Goto, 2000). The third ranking author, Trey Ideker got a bachelor's and master's degree in engineering before studying

---

[2] For more information of CiteSpace, please see the following link: http://cluster.cis.drexel.edu/~cchen/citespace/

biotechnology for his PhD. The fourth ranking author, Albert-Laszlo Barabási is a physicist best known for his work in network theory (Barabási & Oltvai, 2004). The fifth ranking author, Michael Hucka, is a mathematician and computational biologist who specializes in designing software infrastructure for systems biology (Hucka et al., 2003).

Table 1 The most highly cited authors according to the data in WoS.

Rank	Cited times	Author Name	Affiliation	Occupation
1	1313	Kitano H	The Systems Biology Institute, Japan	Engineer
2	765	Kanehisa M	Institute for Chemical Research, Kyoto University,	Physicist, bioinformatician
3	730	Ideker T	Chief of Genetics, UCSD	Engineer, biologist
4	651	Barabasi AL	Center for Complex Network Research, Northeastern University	Physicist
5	560	Hucka M	Department of computing and mathematical sciences, Caltech	Mathematician, computer scientist
6	492	Ashburner M	Department of Genetics at University of Cambridge	Biologist, bioinformatician
7	485	Nicholson JK	Department of Surgery & Cancer, Imperial College London	Biologist
8	440	Shannn P	Institute for Systems Biology, Seattle	Mathematician/Computer scientist
9	414	Alon U	Weizmann Institute of Science	Physicist
10	380	Fiehn O	Department of Molecular and Cellular Biology & Genome Center, UC Davis	Biologist

Based on this analysis I observed that among the most highly cited authors, engineers, mathematicians, computational biologists, and physicists have more citations

than biologists while overall traditional biologists seemed to have contributed less to systems biology. This phenomenon is probably not unusual for an interdisciplinary discipline such as systems biology. Nonetheless, I realized that I only selected the top 10 authors instead of looking at a larger sample of authors, so this result just gave me a preliminary impression of the field. It could also be caused by the fact that these scientists invented software or algorithms that tend to get cited more often, but one may wonder why that did not happen to biologists who invented high-throughput biotechnologies.

1.1.2. Further research on my initial findings

The observation that non-biologists took a more leading role led me to read more literature on systems biology. I found that many widely used and influential textbooks of systems biology were written or edited by authors who have a background in engineering or physics, rather than in biology, for example Uri Alon's *An Introduction to Systems Biology: Design Principles of Biological Circuits* (Alon, 2006). The first International Conference of Systems Biology held in Tokyo, Japan in 2000 also reveals the dominance of engineers. The conference was initiated by Kitano, an engineer, and based on proceedings of that conference, Kitano edited and published the first monograph on systems biology, *Foundations of Systems Biology* (Kitano, 2001).

The great influence of non-biology disciplines is also reflected in the institutions of systems biology. Two institutions were particularly important for the promotion of systems biology in the world. The world's first two institutes for systems biology were both set up in 2000. These are the Systems Biology Institute (SBI) in Tokyo, Japan, established by Kitano, and the Institute for Systems Biology (ISB) in Seattle, Washington, with Leroy Hood as the founding director. Having an engineer as a director

or having a biologist as a director influences the atmosphere of the two institutions, which will be explained later. Beside those two institutes in Seattle and Tokyo, many other centers or departments for systems biology have been established since 2000, including the Department of Systems Biology and Bioinformatics at the University of Rostock in Rostock, Germany, the first such institute in Europe, and the Center for Integrative Systems Biology in Manchester, UK. The rapid institutionalization of systems biology is still happening, as an indication of an emerging field (Powell *et al.,*2007). I will focus on the first two institutions, and discuss how they differ in their research agendas.

Kitano, a leading Japanese scientist in systems biology and President of SBI, and his U.S. counterpart Hood have different conceptions for systems biology. Kitano got his PhD degree in Mechanical Engineering in 1991 from Tokyo University. He edited the first ever monograph on systems biology in 2001, and published a 2002 *Science* article titled "Systems Biology: An Overview," which is the most highly cited article in the field of systems biology (Kitano, 2001; Kitano, 2002). Partly because of Kitano's background in engineering, his interpretation of systems biology is "a combination of computational biology and experimental biology," while his own research focuses more on the computational aspects (Kitano, 2002, p. 206). Kitano's study and works of other Japanese scientists in SBI focus mainly on computational systems biology, including developing Systems Biology Markup Language (SBML), a machine readable language based on XML to describe the models of biochemical interactions; the establishment of large-scale database such as KEGG (Kyoto Encyclopedia of Genes and Genomes); and the developing of a digital tool called CellDesigner, which enables scientists to easily create

network models for complex biological networks (Kanehisa & Goto, 2000; Hucka et al., 2003; Funahashi et al., 2008), to name a few.

On the other hand, Leroy Hood got his bachelor's degree in biology from California Institute of Technology (Caltech) in 1960, a MD degree in 1964 from Johns Hopkins University, and then a PhD degree in biochemistry in 1968, again from Caltech. In the 1980s and 1990s, Hood and his colleagues at the Caltech were optimizing the sequencing and synthesizing method of DNAs and proteins (Hood et al., 2004). In 2001, Hood established the Institute of Systems Biology in Seattle. Staring in 2004 Hood proposed that systems biology would facilitate predictive, preventive, and personalized medicine (Hood et al., 2004). This is in line with his life-long commitment to advancing biomedicine. Similar to Kitano, Hood (2003) acknowledges that "computation" is an integral part of systems biology, but he seems to focus more on the other two components: "biology" and "technology," and its medical applications. Besides being keen on developing new biotechnologies, the ISB puts more emphasis on experimental systems biology and solving real biological problems. The comparison between Hood and Kitano is an interesting example of how scientists coming from different backgrounds perceive systems biology differently.

Some systems biologists as well as historians and philosophers of biology have noted the different epistemology and methodology between biologists and scientists from a non-biology background. For example, systems biologists Hans V. Westerhoff and Bernhard O. Palsson thought that there are two lines of research in molecular biology leading to systems biology analysis. The first line of research is about advancements that are more biological: for example, automatic sequencing, high-throughput sequencing

technology, and Human Genome Project; the second line is more systems-rooted, for example, the use of nonequilibrium thermodynamics theory, feedback controls and network studies in biology (Westerhoff & Palsson,2004). I adopted Westerhoff and Palsson's wording and use the word "biology-oriented" to describe research and scholars related to biology, and "systems-oriented" to describe research and scholars related to physics, chemistry, computation, and other non-biology disciplines.

All these observations raised the question: Was systems biology, as a subfield of biology, really influenced more by systems-oriented scientists? If so, has it always been like this from early 1990s to now or can we observe a shift in emphasis? And how could one measure such a shift quantitatively?

1.2. Driving Question

The driving question for my dissertation thus is: **How did systems biology change as a discipline from 1992 to 2013? Did it shift from a more "systems-oriented" to a more "biology-oriented" discipline? What methodology do I need to study such shifts in the history of (recent) science?**

In the following sections, I explain what my driving question means, and then break it into three sets of subsidiary questions.

1.2.1. What is the distinction between "systems-oriented" and "biology-oriented"?

Before I go further, I want to be clear about what my understanding of "systems-oriented" and "biology-oriented" is. The differences between systems-oriented and biology-oriented research are with regards to the following criteria: what tools researchers use, what form of data they deal with, the type of experiment they perform,

the explanations they propose, and their epistemic goals, which are summarized in Table 2.

Table 2: A comparison of systems-oriented and biology-oriented research.

	Systems-oriented research	Biology-oriented research
Tools	Algorithms, mathematical modeling (e.g.: Boolean, Bayesian, and non-differential equations)	High-throughput technologies (e.g.: microarray, fNMR, four-dimensional microscopic imaging)
Data	Often involving data standardization and modeling	Often involving data generation using omics tools
Experiment types	Simulation and prediction (through iterative process); *in silico*	Measuring, perturbation, manipulating and validating; *in vivo* or *in vitro*
Explanations	Mostly mathematical	Mostly functional and mechanistic
Epistemic Goals	Developing generalized software, algorithms, and databases; understanding the abstract properties of systems or networks.	Understanding and solving specific real-world functional problems; application in bioengineering and medical fields

Systems-oriented researchers include mathematicians, physicists, computational biologists, engineers, etc. What they have in common is that they tend to think in an abstract and mathematical way about principles that can be applied to general systems. Systems-oriented thinking can be traced back to systems theories such as cybernetics in the middle of the twentieth century. Mathematician and philosopher Norbert Wiener (1948) defined cybernetics as “the scientific study of control and communication in the animal and the machine”. Systems-oriented researchers often study abstract properties

that are common across a number of different systems, such as complexity, robustness and emergence. They rely heavily on mathematical modeling, a central theme of systems biology. For example, they model gene regulatory networks as Boolean or Bayesian networks based on gene expression data and their epistemic goal is to improve the efficiency of the modeling (Jong, 2002). They also aim to develop algorithms, database, and software for other scientists to use.

"Biology-oriented" studies involve either one or more of the following levels of real biological information: the molecular, genomic, cellular, evolutionary, developmental, and phenotypic levels. Biology-oriented scientists include evolutionary biologists, developmental biologists, zoologists, etc. They study real biological systems such as gene regulatory networks, metabolic networks, and signal transduction networks of concrete model organisms instead of abstract networks. They offer explanations mostly in terms of mechanisms that explain specific biological phenomena. Biology-oriented scientists may work on generating massive biological data through high-throughput technologies, such as microarray analysis and fNMR, or mapping out all the genes and their interactions underlying a disease. Their goals are to understand complex phenotype and sometimes manipulate the functions of real biological systems to have applications in biomedical research, health care and drug development, or even synthetic biology (Kirschner, 2005).

If I say a scientist is "systems-oriented," it does not mean that everything they work on, every publication they have, and every topic they study have no biological component. It just means that they take a stance more of an engineer/ physicist/ mathematician/ computer scientist rather than a biologist, utilizing more of the

knowledge in their own disciplines to solve a problem. The same goes with an article: in a specific systems biology article, if I say it is "systems-oriented," it does not imply that it only talks about mathematical models and algorithms; rather, it may also talk about some biological concepts, but the methodology relies more on systems thinking than biological thinking.

When biology-oriented and systems-oriented scientists do not cooperate with each other and use their own methodologies to study biological systems, it is easy to determine whether a study is systems-oriented or biology-oriented. For example, some biologists only experiment on the upgrade of first-generation technology to second-generation technology, like what Leroy Hood did with his colleagues on second-generation sequencing technology, which can be easily deemed as biology-oriented. For another example, when one reads an article full of algorithms and modeling processes of an abstract biological system without any real biological data, one can easily identify it as systems-oriented. Yet, sometimes there is not a strict demarcation between the two. Some articles fall into a group that can't be easily identified because they may be deemed as both systems-oriented and biology-oriented.

1.2.2. Why did I choose the years 1992 to 2013 as the study frame?

I chose to study the history of systems biology from 1992 to 2013 instead of studying from the 1950s, because I consider that the former period represents "new" systems biology as opposed to "old" systems biology. I will first explain what old and new systems biology mean, and why I am interested in new systems biology instead of old systems biology.

Some scholars argue that the earliest roots of systems biology can be traced back to the middle of the twentieth century, when mostly engineers tried to model biological systems (Krohs & Callebaut, 2007; Levesque, & Benfey, 2004). As mentioned earlier, in 1948, mathematician Norbert Wiener proposed theories about cybernetics, using mostly organisms and device as examples (Kitano, 2001). Others included the modeling of physiological processes, such as research on action potentials, which gave rise to the Hodgkin-Huxley model of neuron activity in 1952, followed by Denis Noble's heart model in 1960 (Hodgkin & Huxley, 1952; Noble 1960). The Hodgkin-Huxley model uses nonlinear differential equations to describe how action potentials in squid work, and the model can explain the experimental data very well. Notably, the model was named after Alan Lloyd Hodgkin and Andrew Huxley, who won the Nobel Prize in 1963. Similarly, Noble developed mathematical model for the pulse and heartbeat. Both models are considered precursors of systems biology (Boogerd, Bruggeman, Hofmeyr, & Westerhoff, 2007; Krohs & Callebaut, 2007). In the 1960s, biologist Ludwig von Bertalanffy and Anatol Rapoport edited a book about General Systems Theory, in which they attempted to develop general laws of biological systems and systems in other social sciences (Bertalanffy & Rapoport, 1963). The success of those scientists encouraged some engineers to get interested in biological systems at that time (Kitano, 2001).

However, old systems biology failed to establish systems biology as a discipline (Powell *et al.,*2007; O'Malley & Dupré, 2005). Critics argue that some of those early engineers often just proposed models or equations that were often too vague to explain real biological problems, because they had insufficient knowledge or lacked an interest in real biological systems. As a result, the trend soon disappeared. It is hard to determine

whether modern physicists and engineers who currently apply what they learned in the physical and engineering system to biology are inspired by those physicists and engineers from the mid-twentieth century.

Some historians have argued that new systems biology, which emerged in the last two decades, is different from the old systems biology, or simply "systems theory applied in biology" in mid-twentieth century (Wimsatt, 2007). Since the 1990s, new technologies have generated big data through the analysis of biological systems that were not available in the middle of the twentieth century, e.g., genomic and proteomic data. A culmination is the Human Genome Project, which elucidated that the human genome is more complex than we initially thought (Powell et al., 2007).

So, why did I choose the year 1992 as a starting point of this new systems biology? Although according to my data, before 2000 there were less than 10 articles having the term "systems biology" in their topics, I argue that technologies that enabled the generation of big data such as the DNA or protein sequencing technologies and modeling techniques were being developed before 2000. Some scholars also argue that systems biology started in the mid-1990s, after the publication of microbial genomes, e.g., those of the *Haemophilus influza*, *E. coli* and yeast (Knuf & Nielsen, 2012).

It was Hood who started to use the term "systems biology" in 1998 for the first time in a journal article. In that article, he suggests that new opportunities would arise from this new field that draws from -omics, or disciplines of biology ending in "omics", such as genomics, proteomics or metabolomics and related high-throughput technologies (Hood, 1998). Historian Alexander Powell and his colleagues also note that Hood predicted in a 1992 book that "the future of biology will depend upon the analysis of

complex systems and networks" (Kevles & Hood, 1992; Powell et al., 2007). Powell and his colleagues think that Hood's claim "clearly captures the spirit of systems biology". Therefore, in the following dissertation, I consider the year 1992 as the beginning year of the new systems biology. In Chapter 2, I will explain in detail how I can overcome the difficulty of sparse publication between 1992 and 2000 using a bibliometric approach. I chose to end in 2013, because that's the time when I finished retrieving my data from the WoS and started to the data analysis for this dissertation.

1.2.3. How to achieve quantitative and objective results through computational analysis of large datasets?

I used computational tools to analyze my data on systems biology. Traditional historical analysis often does not emphasize quantitative data, but instead relies on the expertise of a historian's accumulated through many years of training and practice. The lab that I work in, the Computational History and Philosophy of Science Lab at Arizona State University takes a different approach to writing the history of science[3]. The principle investigator, Manfred Laubichler and his colleagues describe that "Computational history of science introduces big data–based approaches and computational analytical methods…enabling the pursuit of novel types of questions, dramatically expanding the scale of analysis, and offering novel forms of publication that greatly enhance access and transparency" (Laubichler, Maienschein, & Renn, 2013, p. 120).

[3] For more information about the lab, see http://devo-evo.lab.asu.edu/?q=computational-hps

Online scholarly databases contain the metadata of thousands of articles on systems biology, and computer scientists have created many digital tools that can be used to analyze the data, which produces quantitative results. Because the results are based on big data instead of personal interpretation of a few selected publications, they tend to be more objective.

In light of this orientation the research in this dissertation served two purposes: (1) to explore the historical trends in systems biology with an emphasis on the biology-orientation and systems-orientation distinction, and (2) to experiment with a variety of computational approaches that can be utilized by historians and to evaluate their potential.

1.2.4. The subsidiary questions

I broke the driving question into a set of subsidiary questions, and the answers to these question will be explained step by step in the next three chapters of this dissertation. I proposed the subsidiary questions mainly according to two considerations: First, the subsidiary questions should involve different aspects of my driving question. Second, the way to answer these subsidiary questions should involve computational tools to ensure that the results are quantitative and objective. To this end I examined a variety of tools from the digital humanities, scientometrics, network analysis, and text-mining, before applying those to answer the research questions of this dissertation. These are:

First, which and who are the most influential articles and authors in systems biology from 1992 to 2013? Did systems biology articles have a trend of citing increasingly more biology-oriented articles or a trend of citing more systems-oriented articles from 1992 to 2013? What were the institutional backgrounds of the authors

publishing systems biology articles? Did they come from a biology-oriented institution[4] or a systems-oriented one?

Second, what concepts can best characterize systems biology? What are the relationships between those concepts? Is there a change of systems biologists' use of biology-oriented concepts and systems-oriented concepts over time in their articles?

Third, how did the development of systems biology vary between different countries, especially China and the West?

Several of these questions could be addressed by analyzing citation data of systems biology articles (Garfield, Sher, & Torpie, 1964). To explore the relationship between concepts, I used a network approach that is becoming more and more popular in knowledge representation (Barabási, 2011) as knowledge of a scientific field can be represented as networks depicting the relationship between concepts (van Atteveldt, 2008). And as it is impractical to explore the development of systems biology in all countries simultaneously, I analyzed China as a case study, and compared the development of systems biology in China with the development in other major countries. I will further explain the subsidiary questions in detail in the next three chapters, and each of the chapter addresses one set of subsidiary question.

1.3. Methods

Philosopher of biology Werner Callebaut (2012) calls systems biology, along with computational biology, bioinformatics, and synthetic biology, big data biology (BDB), in

[4] By institution I mean an organization where the scientist work in, including the the department level information.

which a deluge of data is produced under the influence of information technology. Interestingly, similar to the way a big data approach is carried out in systems biology, I used the digital tools developed in the past decades to study the bibliographic information of 9923 articles on systems biology as mentioned in Section 1.1.1. Unlike a well-established traditional discipline like evolutionary biology, in which one often needs to go back to classic books like Darwin's *The Origin of Species*, systems biology is such a new discipline with most of its publication as journal articles deposited in online databases, so that looking at journal articles alone can give a good representation of its developments.

Historians and philosophers of science often rely on qualitative descriptions and narratives. This approach can be augmented by employing digital tools to generate quantitative data. My project is an example of "digital history and philosophy of science (digital HPS)." Digital HPS is part of digital humanities, which happens when humanities, like social sciences and natural sciences, take a "computational turn" (Berry, 2011). Digital humanities can be traced back to the 1940s to Father Robert Busa's use of computation in linguistic analysis for his work *Index Thomisticus* (Schreibman, Siemens, & Unsworth, 2008). Because of the advancements in information technologies and especially the World Wide Web, digital humanities have broadened their scope and are widespread in, for example, archaeology, art, linguistics, and music. The data of digital humanities involve not only textual data, but also visual data such as paintings and audio data such as music. Because published knowledge is growing at an unprecedented speed, the analysis of the data needs assistance of computers (Schilling, 2013). According to a study in 2010, half of the tools in digital humanities are used for text analysis, which

suggests that text is a main object of study (Shreibman et al., 2010). In my study, metadata is also a specialized form of text.

Digital HPS is still a very young field. According to a study reviewing all projects listed in the website of the Digital HPS Consortium in 2013, more than half of the digital HPS projects examined concerned mainly digitalizing paper versions of data through scanning and Optical Character Recognition followed by close reading and manual annotation, instead of automatic and computational analysis (Damerow, 2014). These digitalizing efforts include for example, the Darwin Correspondence Project[5], which has digitalized all of Charles Darwin's works and letters (Van Wyhe, 2006), The Alfred Russel Wallace Correspondence Project[6], etc.

My project uses computation to analyze already digitalized metadata of publications because "computational methods allow for automated data extraction, data and text mining, network and other types of visualization, statistical analysis, and causal modeling (such as agent-based models)" (Laubichler et al., 2013). The interpretation of the results from the historical and philosophical perspectives is equally important as computation.

The hypothesis of my dissertation is that systems biology was first systems-oriented, but later became biology-oriented. My hypothesis is addressed by big-data analysis, similar to the way scientists test their hypotheses through experiments. My research is also data-driven and based on computational approaches that can get

[5] See https://www.darwinproject.ac.uk for more information about the Darwin Correspondence Project.

[6] See http://wallaceletters.info/content/homepage for more information about the Alfred Russel Wallace project.

meaningful information from big data. Some scholars have argued that hypothesis-driven research and data-driven research should not reject each other (O'Malley, Elliott, & Burian, 2010). My research confirmed this position as I continuously asked new questions and discussed unexpected outcomes resulting from the analysis of data.

1.3.1. Selection criteria of the online database WoS

My research uses the metadata and data of 9923 articles to represent the whole field of systems biology as described earlier. As Laubichler et al. have said: "As with similar transformations in the life sciences (another fundamentally historical field), the starting point of computational approaches is big data" (Laubichler et al., 2013, p. 121).

In recent years, more and more articles on systems biology have been published and are indexed in a number of different places. Of course I cannot study all the literature on systems biology one by one in all databases. The first thing I needed to do was to select a reputable database. I chose Thomson Reuters' WoS after comparing it with some other databases especially PubMed and Google Scholar.

Compared to other well maintained databases such as PubMed, the WoS database contains more articles on "systems biology." For example, when searching for articles that have "systems biology" in their "titles," PubMed has 1995 articles, whereas the WoS has 3445 articles as of December 6th of 2013. In addition, for the hypothesis that systems biology is becoming more biology-oriented, I needed to choose a database that is not biased toward either systems-oriented articles or biology-oriented articles. The Pubmed database includes mostly articles related to medicine and biological sciences, so it is biased toward biology-oriented research and thus unsuitable for my research.

Google Scholar has more articles on systems biology than the WoS, but the quality of the articles of the two databases is different. The WoS only contains articles that are published by reputable journals, whereas Google Scholar uses its algorithm to maximize the quantity of articles contained and the results are sometimes inconsistent (Falagas, Pitsouni, Malietzis, & Pappas, 2008).

Furthermore, compared with Google Scholar, WoS can export structured metadata files for further analysis. WoS is also famous for its citation indexing, so WoS can export titles, abstracts, publishing years, authors, references and other metadata of 500 articles all at once. Many bibliographic tools can work directly on the metadata exported by WoS, such as the ones that I used for this project. Therefore, based on the above comparisons, I chose WoS.

1.3.2. Computational workflow for my research

This dissertation presents a computational workflow that combines three types of analysis: the first is the analysis of citation data; the second is network analysis and computational linguistic analysis; and the third is geographical analysis and comparative analysis (See Figure 3). Each of the approaches addresses one of the subsidiary questions and discussed in the chapters below.

The analysis of citation data can give us information about the highly cited references, research topics and institutions, which are important in shaping the field. The network and linguistics analysis sheds lights on concepts, their use, and their importance in different times. Geographic analysis and comparative analysis illuminate how social factors such as city, country, and institution, can contribute to difference in quantity, quality, and varieties in research. The three types of analysis complement each other and

reveal a comprehensive picture of a scientific field. The computational flow combines some widely used tools popular in the digital humanities, scientometrics, network analysis, and data-mining communities, but also the Python codes that I wrote to retrieve and analyze data.

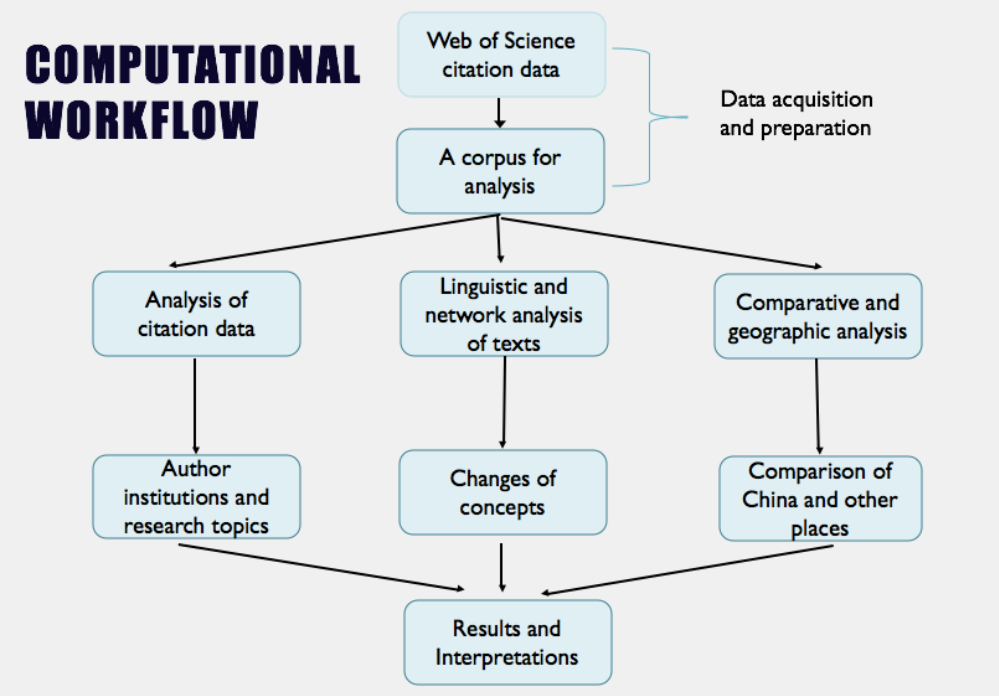


Figure 3. The computational workflow of the dissertation.

The digital tools used in this project include CiteSpace, Tethne, Cytoscape, Mallet, Wordsmith, and Google Fusion Tables, which will be explained in detail in my other chapters. Python was used extensively in my research, which is a programming language that works well with natural language, and is used in industry and research, not only in computer science, but also in the humanities such as linguistics, economics, and history (Bird, Klein, & Loper, 2009). In my research, I used Python to work with texts, which is a form of natural language. Python has modules, such as Xlrd, Xlwt, Scipy,

Numpy, that enables data extraction, data analysis, statistical analysis, data output, and graph design.

1.4. Layout of the Dissertation

This dissertation consists of five chapters. Notably, Chapter One and Chapter Five are written in the form of dissertation chapters, whereas Chapter Two, Three, and Four are written in the form of individual articles aimed for publication. Chapter Two, Three, and Four are analyses based on the same initial dataset, i.e., the metadata of the 9923 articles on systems biology, but each chapter describes different methods to analyze the metadata. Because they were written as individual articles, they all introduce what systems biology is and the dataset in order to be complete, so there might be some overlapping content.

Chapter Two identifies the most highly cited 330 references and 330 authors from 1992 to 2013 by using digital tools to analyze the metadata I downloaded in my pilot study to represent systems biology. I classified those 330 references into biology-oriented research and systems-oriented research. A close reading of those 330 references suggests that during the past few years, articles in -omics research, database research, and medical research increased tremendously. The institutional backgrounds of the most highly cited 330 authors suggest that before 1996, systems-oriented scientists overshadowed biology-oriented scientists. However, after 1996, most of the scientists who published on systems biology are biologists. This chapter explores several turning points in the history of systems biology, and divides systems biology from 1992 to 2013 into the "early roots," "establishing," and "subfield emerging" stages. Right now, more and more subfields are still emerging within systems biology. Topic modeling of the abstracts of articles

published after 2000 highlights the increasing trending of medical research, corroborating my close reading of the 330 most highly cited references.

Chapter Three offers information gathered from a corpus of about 5 millions words built based on abstracts of 9876 articles published between 2003 and 2013 retrieved from the metadata[7]. I used a multi-method analysis to map the co-word/concept networks to show how the networks changed over time. The co-word/concept networks consist of hundreds of words/concepts linked by co-occurrence relationship. I analyzed the network properties of biology-oriented and systems-oriented concepts over time. The network properties of biology-oriented and systems-oriented concepts show different trends over the years. More than half of biology-oriented concepts have increased centrality in co-word/concepts networks and the reverse is true for systems-oriented concepts. Network analysis also allows me to zoom in on one part of the co-word/concepts network to look at a word to see its change over the years. For example, the words of "therapy" and "cancer" were used together with more types of words over time.

Chapter Four describes a case study that explores regional differences. I used computational approach to study the development of systems biology in China. I wanted to know if the global trend of systems biology can also be seen in China. The case study is also a comparative study, comparing research in China with a few major countries, including the US, Germany, England, and Japan. The reason I chose to look at China is

[7] The number is 9876 instead of 9923 is because we look at the years between 2003 and 2013, instead of between 1997 to 2013.

that China ranked NO. 2 (The NO.1 is the US) in publishing the highest number of scholarly articles on systems biology by 2014, and it has had a strong increasing trend. There have been articles on the development of systems biology in the US, Japan, but less so for China. The approaches used include GIS (Geographical Information System), network analysis, and bibliographic analysis. Chapter Four shows that although the quality of Chinese scholars' work is slightly poorer than their counterparts in the US in terms of impact factor of the journals that Chinese scholars publish in, the topics of Chinese scholars were mostly similar to those of the US, Germany, and Japan, but with an exception of also focusing on traditional Chinese medicine. In addition to that, my research reveals the unequal distribution of research power in China.

In the final chapter I first discuss the summary of my research findings and the reflections based on the findings. Systems biology represents a new turn in biology. My analysis of its applications in medicine and bioengineering, its interdisciplinary nature, and its relationship with systems science all offer new insights about this new discipline. Next I summarize the big data and computational approachs that were used in this research and how they enabled me to answer questions that were hard to answer using traditional methods of historiography, and offer my understanding of the digital HPS. Finally, I explain the future directions and the limitations of this kind of research.

CHAPTER 2: FROM SYSTEMS TO BIOLOGY: A BIBLIOGRAPHIC ANALYSIS OF THE RESEARCH ARTICLES ON SYSTEMS BIOLOGY FROM 1992 TO 2013

Summary: Systems biology is a discipline that studies biological systems from a holistic and interdisciplinary perspective. It brings together biologists, mathematicians, computer scientists, physicists, and engineers. We applied several computational tools to analyze the bibliographic information of published articles in systems biology to answer the question: Did the authors and research topics of systems biology become more biology-oriented or more systems-oriented from 1992 to 2013? We analyzed the metadata of 9923 articles on systems biology from the *Web of Science* database. First, we generated co-citation networks for different time slices to visualize the development of systems biology. The co-citation networks reveal three different stages of systems biology and we divided the time between 1992 and 2013 into three stages, titled as "early roots," "establishing," and "subfield emerging" stages. Next, we identified the most highly cited 330 references and through close reading we divided them into nine categories of research types in systems biology, and found that articles in one category, namely systems biology's application in medical research, increased tremendously. Furthermore, we identified the most highly cited 330 authors over time. We found that before mid-1990s, systems-oriented scientists have made the most referenced contributions, but in more recent years, biology-oriented researchers have made more and more of the most referenced contributions. This finding was corroborated by computational analysis of the

abstracts, which also suggests that the percentages of topics on vaccines, diseases, drugs and cancers increased over time.

Keywords: Systems Biology; Bibliometrics; Application; CiteSpace; Tethne; MALLET; Topic Modeling

The development of high-throughput technologies in the 1990s brought forth a deluge of data to biology. Without mathematical models and computational simulations, however, the data could not be understood at the time. Systems biology is an interdisciplinary field, where biologists, referred here as biology-oriented scientists, and engineers, computer scientists, and mathematicians, referred as systems-oriented scientists, both participate.

Systems-oriented scientists have contributed greatly to the advancement of systems biology. For example, Hiroaki Kitano published the most highly cited article in systems biology, edited the first monograph on systems biology, founded the Systems Biology Institute in Tokyo, and organized the first International Conference of Systems Biology in Tokyo in 2000 (Kitano, 2001). Kitano was trained as an engineer and is the head of Sony Computer Science Laboratories. Another example is Albert-Laszlo Barabási, whose work on network theory has won him many awards for systems biology. He was trained as a physicist, yet he publishes widely in systems biology (e.g. Barabási et al., 2004).

Because scientists from different backgrounds have different epistemologies and methodologies, one may wonder how biology-oriented disciplines and systems-oriented

disciplines have shaped the research topics within systems biology. This study examines the history of systems biology from 1992 to 2013, utilizing a set of computational tools to analyze the metadata of the systems biology literature. We answered the following questions: From 1992 to 2013, how can one visualize the evolution of systems biology? How did the topics of systems biology research change and did this change reflect a shift towards more biology-oriented topics? How did biology-oriented and systems-oriented scientists contribute to systems biology at different times?

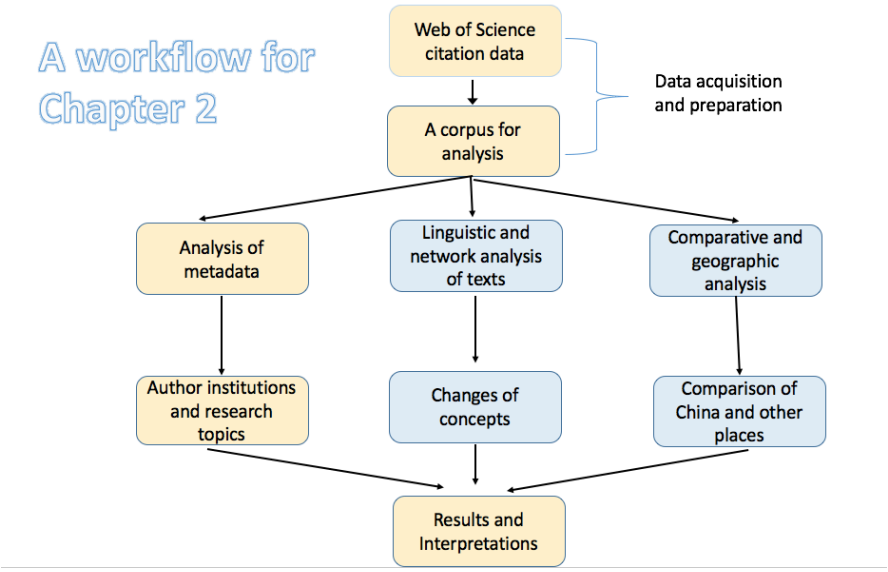
2.1. Methods

This study is one of the first systematic analyses of the history of systems biology that is based on bibliographic information. The growing number of publications in a scientific field like systems biology makes it hard to identify trends and study frontiers simply by analyzing key papers. Bibliometrics can provide analysis tools to address these difficulties. The study of bibliographic data is called bibliometrics and when it concerns scientific information, it is called "scientometrics." They are two closely linked areas (Sengupta, 1992).

Scholars have applied bibliographic analysis to study the history of business, science, art, and engineering (Leonidou, Katsikeas, & Coudounaris, 2010). Bibliographic analysis is a good way to assess the influence and quality of literature by deciding which work gets cited most and which author has the most citations (Moed, 2006). Earlier attempts to analyze citation date form the Web of Science, such as those by STS scholar Susan Cozzens in the 1990s, were limited as many computational tools for the analysis of big data were not yet available (Cozzens, 1997). However, after more than a decade of development, information scientists have produced many tools and approaches for

citation analysis, which can overcome the difficulty faced by Cozzens. These include the ISI citation index, CiteSpace, HistCite, VOSViewer (De Bellis et al., 2009; van Eck & Waltman, 2010; Garfield, 2009).

Because systems biology is a very new discipline, scientists have published their findings mostly in the form of peer-reviewed articles that are accessible online. In our study we used the citation data of 9923 articles in Thomson Reuters's WoS. Based on these data, the study discussed in this chapter ran four kinds of computational analysis and interpreted the results from a historical perspective as shown in Figure 4.



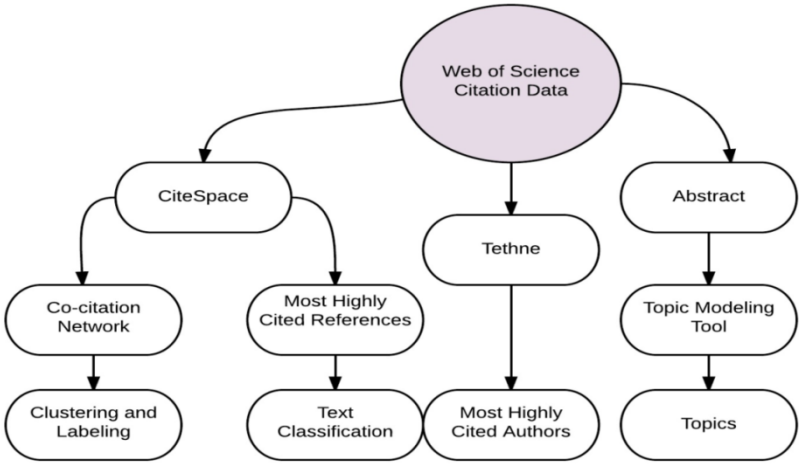


Figure 4. The flowchart for Chapter 2. The figure on the top highlights parts carried out in this chapter, and the parts that are not highlighted are for Chapters Three and Four. The figure on the bottom shows the detailed steps in Chapter two. Explanations of these workflows can be found in the following sections.

2.1.1. Data collection

In the WoS database, we searched for documents containing the term "systems biology" in their topics (including titles, abstracts, and keywords) and published from 1992 to 2013. Then by selecting those published in English, we narrowed the sample down to 9923 articles, which included research articles, reviews, editorial materials, proceeding papers, and meeting abstracts[8]. We downloaded the bibliographic information

[8] In the WoS database, we set the search criteria for year to be from 1992 to 2013. However, we found that the first article that contains the term "systems biology" was published in 1997. Despite this limitation, we can use bibliographic analysis tool to get the relevant references before 1997 even when there are no publications using systems biology as a term at that time.

for these 9923 articles. If each article on average has 10 to 30 references, the total number of references for the 9923 articles should be between 100,000-300,000, which are analyzed using bibliographic tools.

For every article, WoS can bulk export all the bibliographic information, such as authors, title, abstract, references, and publishing year. For more information of the types of bibliographic information that WoS can export, see Appendix A-1. Using computational tools, we followed the steps shown in the flowchart on the bottom in Figure 4. To answer our questions, a systematic analysis extracted four types of information from the bibliographic data. We first visualized the field of systems biology in different stages, and then studied the research categories of the most highly cited 330 references by close (manual) reading. Next we analyzed the affiliations of authors, and finally used machine learning techniques to study the topics of systems biology embedded in the abstracts.

2.1.2. The conceptual model of citation analysis

There are many benefits of citation analysis. First, it can expand the scope of research from a group of articles to the references of these articles (Chen, 2006). Looking at references with the help of bibliometric tools enabled us to study the historical period before 1997, because these tools can use a computational method to analyze the references automatically, and overcome the time limitation that before 1997 no articles use the term "systems biology." However, the foundations of systems biology have been laid in the context of the articles referenced by those early systems biology papers. Second, it can help us identify key papers, authors, etc., and facilitate information retrieval using computational method without being overwhelmed by the large number of

publications or having to spend tremendous time studying the literature (Dunne, Shneiderman, Gove, Klavans & Dorr, 2012). Third, many network visualizations can be generated by computational tools to map the knowledge structure and paint a "big picture" of a scientific field (Börner, Chen, & Boyack, 2003). The second and the third points are explained in other sections, and this section explains the conceptual model of citation analysis as shown in Figure 5, and why it can expand the scope of our research.

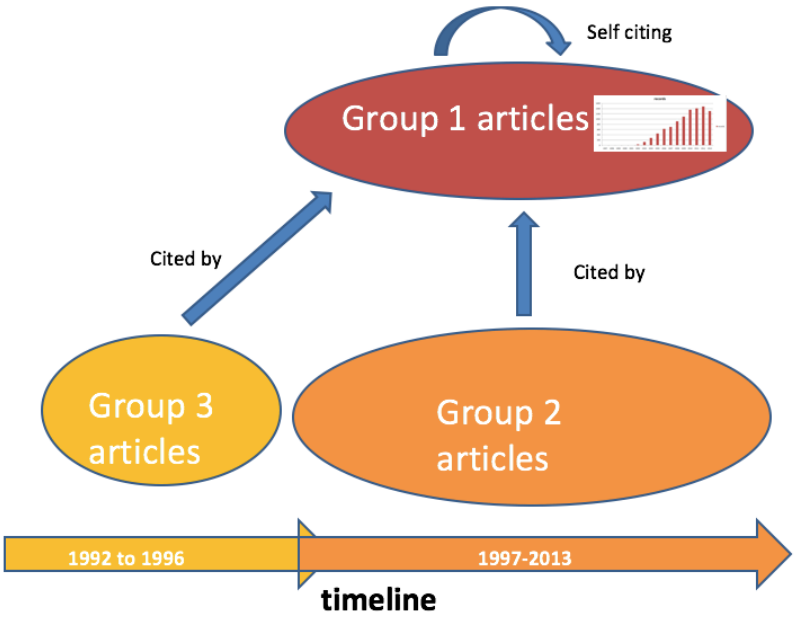


Figure 5. Literature in three categories, Group 1, Group 2 and Group 3. Group 1 literature refers to the 9923 articles, and Group 2 and Group 3 literature refers to references cited by Group 1 articles.

We refer to those 9932 articles as group 1 (G1) literature. These articles were published between 1997 and 2013. However, we believe that some other references that are cited extensively by the G1 articles can also be considered as literature that contributed to systems biology, although they may not be included in the G1 because they do not use the term "systems biology" in their "topics" or they were not deposited in

the WoS database. Yet we were able to find these articles through citation analysis and could therefore expand the scope of our research to years before 1997 using computational tools. We could also find additional information such as who are the most highly cited authors or references.

These cited references can be grouped into two categories: the first category of those published from 1992 to 1997 when the drivers for systems biology, such as various sequencing projects, started to emerge. Although these references may not directly have "systems biology" in their "topics", they have in fact contributed to the G1 literature as they were highly referenced by those early articles. Therefore, we decided to include them in our research scope and call these article group two (G2) literature.

The second category refers to references that were published after 1997, but also do not have the term "system biology" in their "topics". We downloaded and read a sample of these references, and found that they often use the systems biology approach, or are under the guidance of systems thinking. While these did not fit our initial search criteria, they nonetheless contributed to systems biology. Similarly, we can find those references using citation analysis, and call these references group three (G3) literature.

The aggregates of the G1, G2, and G3 literatures can be a good representative sample of the literature on systems biology[9]. From the metadata of G1 articles, we found

[9] For G1 literature, we have the full metadata of them downloaded from the WoS. For G2 and G3 literature, we have the most basic information that can identify them, including the author, the publishing year, the journal name, etc. This most basic information is retrieved from the metadata, for example "Ahuja I, 2010, TRENDS PLANT SCI, V15,

the information of G2 and G3 articles; therefore, this project examines an expanded scope as shown in Figure 6.

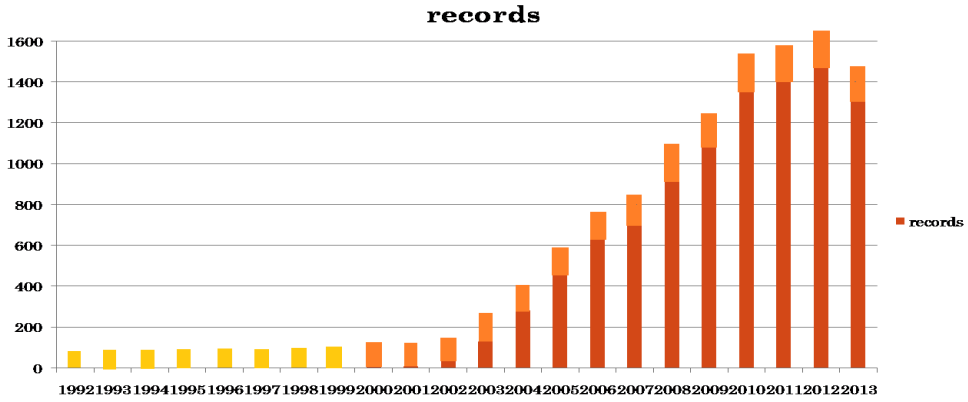


Figure 6. The expanded literature scope. The *x* axis is the time, and *y* axis is the number of publications. The computational tools can add cited references in the research scope. The 9923 articles are the bars in the red color and the scale is matched with real data. The expanded scope is in yellow and orange color, and is not at scale as the numbers are much larger. We just indicated which type of literature was added to each year. The expansion of the literature collection was facilitated by the use of bibliographic tools looking into the metadata of the initial 9923 articles.

2.1.2: Visualization of the evolution of systems biology using co-citation network analysis

---

P664, DOI 10.1016/j.tplants.2010.08.002." Although the information of cited references is minimal, we can do analysis such as figuring out which reference gets cited most, which author gets cited most, etc., and downloading those articles for close reading.

We used CiteSpace to analyze the WoS bibliometric data to visualize the evolution of the field of systems[10]. The references of many articles in a scientific field, can shed light on the intellectual base for that field. Historians often want to look at the intellectual basis of a field to trace where and how knowledge grows (Chen, 2006; Chen, 2009). This bibliometrics tool has been used for determining historical and emerging trends in many area of research, including regenerative medicine, neuroscience, and psychology, etc. (Chen, Hu, Liu, & Tseng, 2012; Wang, Zhang, & Qiu, 2012).

CiteSpace generates co-citation networks by analyzing the bibliographic data of hundreds or thousands of papers automatically (Chen, 2006)[11]. If two references are cited together by another article, then these two references have the relationship of co-citation. The network created in such a way can visualize a scientific field. CiteSpace can also automatically calculate how many times a reference is cited. It can be assumed that the more citations a reference has with other articles, the more important that article is to a field, in our case, systems biology (Dunne et al., 2012). Thus, CiteSpace can be used not only as a visualization tool, but also as a selection tool.

Co-citation networks generated for different time slices can help visualize how a scientific field has evolved. We divided the time span from 1992 to 2013 into 11 time

[10] The software can be accessed at http://cluster.cis.drexel.edu/~cchen/citespace/. There are many kinds of citation analysis that CiteSpace can carry out.

[11] Co-citation analysis is one kind of citation analysis that CiteSpace can carry out. Others include co-author analysis, author co-citation analysis, and hybrid network analysis, which given the space of this dissertation we will not discuss here.

slices: (1992 to 1993), (1994 to 1995), …., (2010 to 2011), (2012 to 2013), and generated co-citation networks for the 11 time slices. CiteSpace also has other functionalities. For example, we can use it to group co-citation networks into clusters and then label the clusters using terms extracted from the titles, keywords, or abstracts.

2.1.3. Analysis of the most cited references for the types of research

The built-in functions of CiteSpace can analyze the citation frequencies of each reference, and produced a list of the most highly cited references from the year 1992 to 2013 (Chen, 2006). Selecting the top 30 references in each period brought a total of 330 references for the 11 time slices. After CiteSpace picked out these 330 references, we manually downloaded them and then analyzed them to determine whether the articles were systems-oriented or biology-oriented.

This requires certain criteria. Because systems biology is an interdisciplinary science, there is no 100 percent biology-oriented or systems-oriented research. However, because any research has to be focused on a certain area, it is possible to classify these articles into a few categories, and for each category, it is easier to say whether each is more systems-oriented or biology-oriented.

Based on the close reading of these 330 articles, we divided them into nine categories of systems biology research to see how the number of articles in each category changed over time. For previous bibliographic analysis, researchers have categorized publications of a field in categories to show the "big picture" of that field (Braisford Braisford, Harper, Patel, &Pitt, 2009; Leonidou, Katsikeas, & Coudounaris, 2010). For our analysis we derived our categories from within the literature (based on close reading) and also from categories identified by other historians and philosophers of biology. We

have to admit that it is a formidable task to categorize publications of a scientific field as diverse as systems biology. However, these categories were based on both reading the articles and examining the work of other historians or biologists.

The historians of science Ulrich Krohs and Werner Callebaut once argued that three roots of systems biology must be discerned to account properly for the structure of the field, namely, pathway modeling, biological cybernetics, and -omics (Ulrich & Callebaut, 2007). We agree that pathway modeling and -omics have fueled the advancements of systems biology during the past two decades. Cybernetics may be very important in the mid-twentieth century and contributed to systems biology's theory as a root; however, in the literature on systems biology published over the last two decades, it is hard to see the influence of cybernetics as comparable to that of pathway modeling and -omics.

We argue that Krohs and Callebaut's categorization is simplistic. Systems biology is a very interdisciplinary field and has many diverse areas; therefore, we used more categories than just the three proposed by Krohs and Callebaut, and most of the articles in systems biology fall in a rather straightforward way into the nine categories we propose.

First, we will introduce what each category means and why each is either systems-oriented or biology-oriented (See Table 3). Biology-oriented research includes the following four categories: -omics-related research, high-throughput technologies, applications in engineering and medicine, and biological mechanisms. Systems-oriented research includes network properties, software development, Metabolic Flux Analysis, database development, and algorithms, equations, and modeling. A more detailed description of these nine categories can be found in Appendix A-2.

Table 3: Nine categories and their descriptions.

Categories	Description	Systems-oriented or Biology-oriented
Metabolic Flux Analysis	Measures the stoichiometric data of metabolites, and relies on modeling using non-differential equations and a few parameters.	Systems-oriented
Development of high-throughput technologies	These technologies include sequencing technologies, protein chips, DNA arrays, and mass spectrometry, etc.	Biology-oriented
Algorithms, equations, and modeling	This category includes development of algorithms, equations, modeling, and simulation techniques that relies heavily on mathematical knowledge.	Systems-oriented
Omics research characterizing a real biological system	Omics research relies on data produced by high-throughput technologies and modeling; the ultimate goal is offering a system-level characterization of a real biological system.	Biology-oriented
Database development	This category involves the launch of databases storing genes, pathways, proteins, etc. It also involves standardization of data and procedures, such as the Systems Biology Markup Language.	Systems-oriented
Software development	Software is developed to process, analyze, and visualize large data.	Systems-oriented
Network properties	These properties include robustness, dynamics, stochasticity, and emergent network properties that can be applied to every system, not just biological systems. The work is mostly mathematical and theoretical.	Systems-oriented
The application of systems biology	Systems biology is especially useful in tackling complex diseases like cancer, and has application in bioengineering and synthetic biology.	Biology-oriented
Biological Mechanisms	This category involves using systems approach to understand a specific biological mechanism.	Biology-oriented

2.1.4. Analysis of authors' affiliations to reveal the institutional context

We used Tethne to determine which authors were the most highly cited over the years (30 authors for each of the 11 time slices). Tethne is a Python package developed by Erick Peirson at the Digital Innovation Group at Arizona State University for

bibliographic and corpus analysis[12]. The tool was written in Python, and it is open source. It works with WoS, JSTOR, and Scopus data to visualize patterns and trends in the scientific literature.

Using Tethne, we analyzed the references for all 9923 articles. Each reference includes the information of the first author's name and the publishing year. For each time slice, we calculated the number of the references an author published in that time slice, and arranged the authors based on that number. Then we picked out the top 30 authors with the highest number of references in each time slice for the analysis in the next step.

We were not only interested in the most highly cited 330 authors, but all the authors of those 9923 articles. Therefore, we retrieved each author's affiliation at the time of publication from either WoS data or Google Scholar. We analyzed these affiliations to determine whether the author was affiliated with a biology-oriented or a systems-oriented institution. The reason we chose their affiliation instead of other information was a trade-off between the content and the accessibility of that information. For example, a person's description on one's own website may be a more accurate assessment of what one is doing, but this information is hard to get for thousands of authors, and harder to compare in an objective way. Affiliations are easier to obtain and can accurately reveal the institutional backgrounds for researchers of systems biology.

We built a word list by retrieving an identifying word from the affiliations of the most highly cited 330 authors. For example, if a department name has the word "anatomy," it is likely to be a biology-oriented institution, and we used "anatomy" as an identifying word. Next, we developed a model written in Python code to study the

[12] For more information about Tethne, see https://github.com/diging/tethne.

affiliations of the 9634 first authors of articles published from 2003 to 2013 in the WoS database[13]. Based on the word list generated by analyzing the 330 authors, we wrote a Python code to design a machine learning model that automatically labels the institutions of thousands authors.

The model is described as follows: the automatic labeling of the institutions was based on the first word, usually the department name that matched the word list. For example, in the affiliation of "Max Planck Inst Mol Plant Physiol, D-14476 Potsdam, Brandenburg, Germany", the first word that matched the word list was "plant," so it was labeled automatically as a "biology-oriented" institution. For institutions names that contained a word that indicated the institution's orientation but is not on our list, we labeled them manually and added the identifying word to our word list. We then ran the process iteratively. However, some institutions were still hard to define because the affiliations retrieved from the WoS citation data did not have a department and only the university, so we labeled these manually as "unidentified", for example, affiliation like "C1 Los Alamos Natl Lab, Los Alamos, NM 87545 USA" would be labeled as unidentified. Another situation where an institution was hard to identify was that it is a foreign institution with a name like "C1 Tech Univ Dresden, Inst Lebensmittel &

[13] The reason for starting from 2003 was that in that year, 118 articles were published whereas in the previous year, only 30 articles were published. In statistics, 30 is usually considered the minimal sample size. Because we started with the year 2003, the articles we looked at was 9634 instead of 9932.

Bioverfahrenstech, D-01069 Dresden, Germany." The word list that we used to identify institution categories contain 193 words and is shown in the Appendix B.

2.1.5. Analysis of topics found in abstracts using topic modeling

Through the previous step, we analyzed the categories of systems biology research through manual reading of 330 articles mentioned in section 2.1.3. However, we wanted to analyze the research categories in a larger sample size, say thousands of articles, and see if the results of the latter corroborated with our close reading. Because we could not have enough time to read all of them, we relied on a machine learning technique: topic modeling. A topic here can be broadly interpreted as a subfield, a category, or a research type. The best thing about topic modeling is that it can analyze millions of words quickly without human reading them (Newman & Block, 2006). Topic modeling has many applications in the humanities, social sciences, and bioinformatics, including the study of Twitter messages to identify trends and studying the corpus of thousands of research papers or newspaper articles to show how ideas in a specific field have changed over time (Hong & Davison, 2010; Hall, Jurafsky, & Manning, 2008; Newman & Block, 2006). We used topic modeling to analyze the 8809 abstracts of articles published from 2003 to 2013 that were retrieved from the WoS Citation data (8809 out of the total of 9634 articles had abstracts)[14].

[14] Before 2003, each year only had a few publications and the number of articles is positively related to the accuracy of the results of topic modeling. Out of the 9634 articles that were published from 2003 to 2013, 8809 articles have abstracts.

Topic modeling uses probabilistic models to generate topics (each topic is represented by a cluster of words) through automatic reading of unstructured natural language (Blei, 2012). One of the most widely used topic model is the Latent Dirichlet Allocation (LDA) model (Blei, 2012).

The main mechanism of topic modeling is as follows: First, the model sets a fixed number of topics and a fixed number of words in each topic. The basic idea is to view a document as a distribution of topics and a topic as a distribution of words. Second, the model randomly assigns the words in a document to a topic and calculates two probabilities: P (topic/document) and P (word/topic). Third, the LDA model utilizes Bayesian inference to adjust the word assignments iteratively until it reaches a relatively stable state. Each iteration involves assigning a word to a topic, and updating the P (word/document) and P (topic/document) to infer and update P (word/topic) (Blei, 2012). This is the simplest explanation of how LDA works. The actual modeling process utilizes more complicated algorithms. Furthermore, since the modeling is based on probability, topic modeling is a close inference of topics, but rather only a representation of the topics through machine learning.

There are many tools that can implement topic modeling, for example the MALLET (Machine Learning for Language E Toolkit), the Stanford Topic Modeling Toolbox (McCallum, 2002; Ramage & Rosen, 2011)[15,16]. We used MALLET, a software

[15] For more information about MALLET, see http://mallet.cs.umass.edu

[16] For more information about the Stanford Topic Modeling Toolbox, see http://nlp.stanford.edu/software.tmt.tmt-0.4/

developed by Andrew McCallum at the University of Massachusetts, and it is based on LDA model and a set of different algorithms. Words like “doi,” “paper,” “ab,” “research,” “results” and “elsevier” were compiled in an extra list of stopwords, which are words that MALLET ignores, in addition to the default one that MALLET includes. The number of topics was set at 20.

We can also analyze how the topics change over time. For a historian, identifying topics in thousands of articles is an interesting task, but it is even more useful to see the topic trends over time (Newman & Block, 2006). MALLET also returned the composition of topics in all the documents over time. In our case, each document is an abstract of article. For example, for a document, MALLET returned the probabilities of each 20 topics. If every topic is equally represented in articles, then the probability of each topic should be 5%. If the probability of a topic is higher, that means that this document has a higher probability of containing that topic. For example, if an article has a topic whose probability is higher than 40%, then it suggests that this article is indeed related to that topic. Similarly, if an article has a topic whose probability is only 1%, then it is unlikely that this article contains that topic. In our study, we consider a probability of 10% to be the threshold for considering that the topic is significant. A Python code was written to calculate the number of articles that contained a topic the probability of which was higher than 10% for a certain year. We then calculated the percentage of that number for all the articles published in that specific year.

2.2. Results

2.2.1. The evolution of the co-citation network of systems biology

We used CiteSpace to answer these questions: How did the field of systems biology evolve from 1992 to 2013 and what stages has the field of systems biology gone through from 1992 to 2013? We first discuss the evolution of co-citation networks, and then discuss three stages of systems biology, followed by the automatic labeling of clusters of co-citation network.

CiteSpace can generate co-citation networks automatically, and the evolution of these networks can shed light on the evolution of the scientific domain of systems biology. By visually analyzing the co-citation networks, one can determine turning points (Chen, 2006). The co-citation networks based on our data over time exhibit three different stages as shown in Figures 7, 8, and 9.

We call the first stage the "early roots" stage, as exemplified by the co-citation network in Figure 7. The characteristics of this network are, first, the nodes are small, meaning that the cited references have fewer citations compared to cited references after 1993. Second, the most highly cited references were published before the 1980s. These most highly cited articles listed in descending order are Gillespie (1977), Kacser (1973), Kauffman (1969), Henrich (1974), and Hodgkin (1952).

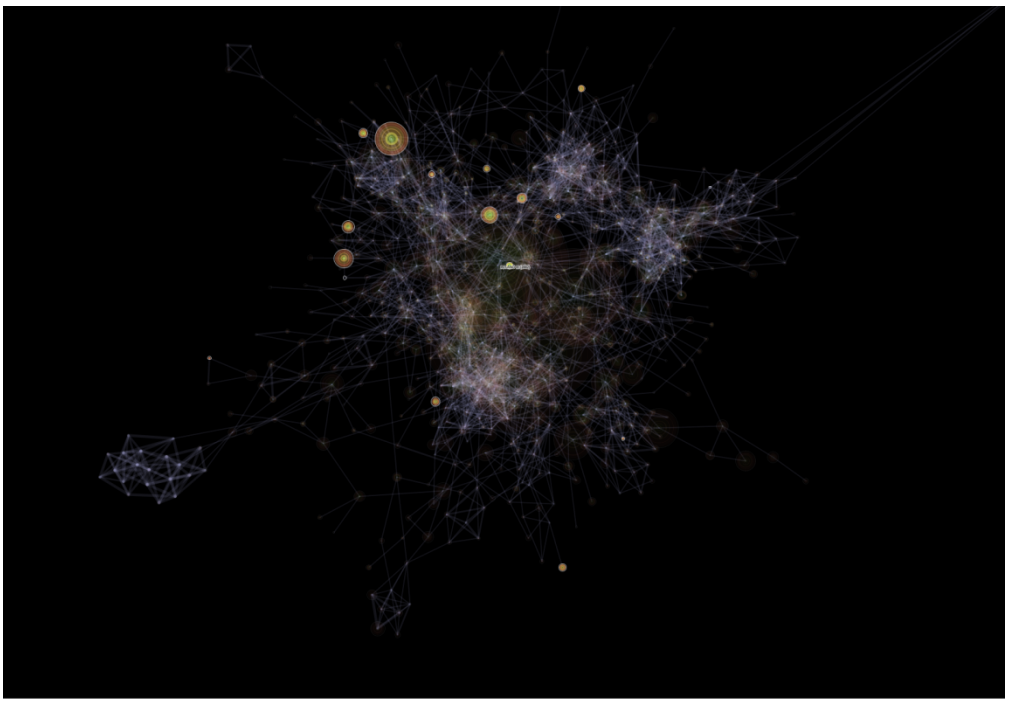


Figure 7. The co-citation network from 1992 to 1993. This network depicts the intellectual base of systems biology up to 1993. Only a few nodes (highlighted ones) appear in the network. In this co-citation network, the size of the node is proportional to the citations that an individual article generates. The bigger a node is, the more citations it has.

In these references, the Hodgkin-Huxley model is believed to be one of the earliest systems biology models. It is a mathematical model that describes the potential actions in squid (Hodgkin & Huxley, 1952). Another prominent article is the Kauffman (1969) paper, where Stuart Kauffman uses a theoretical network of genes to analyze complex network properties. Kauffman is believed by many to be an early pioneer of systems science, and he continues to work on complex systems to this day (Ramage & Shipp, 2009). These highlighted works were published before the 1980s, and no significant references published in 1980s and early 1990s are highlighted. In this stage,

the field of systems biology still mostly referenced certain classic papers published from the 1950s to the 1970s.

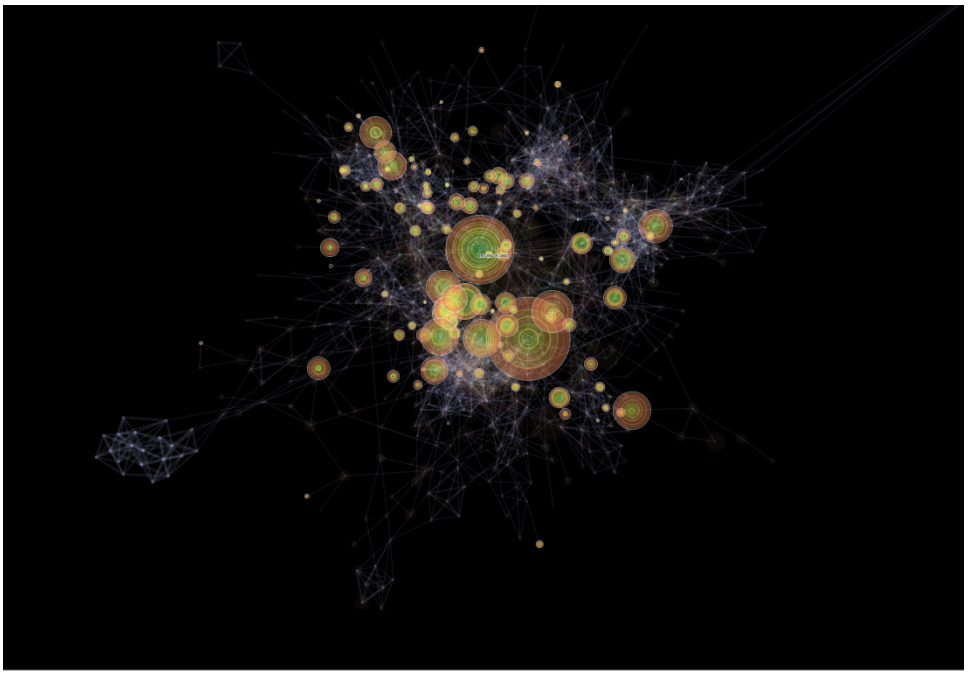


Figure 8. The co-citation network from 1992 to 2001. This figure depicts the intellectual base of systems biology up to 2001. Much more nodes with higher citations appear in the network.

We call the second stage the "establishing" stage, as exemplified by the co-citation network in Figure 8. The key characteristic of this stage is that from 1994 to 2001, more recent nodes emerge in the network. The increase of nodes is gradual and not exponential. By the year 2001, much bigger nodes have appeared compared to those in Figure 7, which means that the new references generated more citations than early classical papers between the 1950s and the 1970s. This suggests that the old systems biology has been replaced by the establishment of new systems biology by 2001.

However, these individual articles are not linked through the edges, suggesting that these articles do not yet form a cluster, which can be interpreted as a specialty, or a subfield (Chen, 2006). However, these individual articles point to the advent of a new discipline.

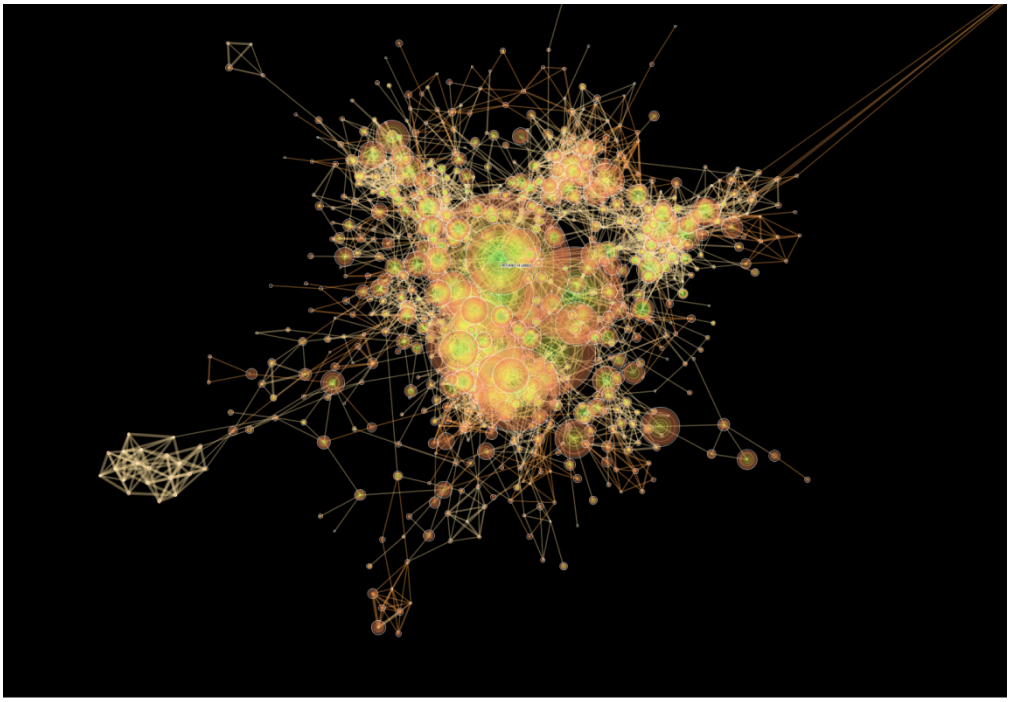


Figure 9. The co-citation network from 1992 to 2013. Clusters of the nodes appear in the network.

We call the third stage "subfields emerging" stage, as exemplified by co-citation network in Figure 9. The key characteristic of this stage is that many clusters have already been formed. On one hand, from 2002 to 2003, the biggest nodes appeared, including Kitano (2002), which generated the highest number of citations within all articles, Fineo (2002), and Hucka (2003). One the other hand and more importantly, by 2013, one sees that many nodes have formed clusters. In a cluster, references are more strongly connected to each other than to references outside that cluster (Chen, Ibekwe-

SanJuan, & Hou, 2010; He, 1999). The emergence of a cluster suggests that a sub-field of systems biology has appeared and the field of systems biology was consolidating around those clusters instead of being composed of loosely connected articles.

CiteSpace enables the automatic clustering of nodes through its built-in clustering algorithm called spectral clustering, a generic clustering technique (Chen, Ibekwe-SanJuan, & Hou, 2010). In addition, by cluster analysis in Citespace, clusters were arranged in ascending order according to the mean publication year (See Table 4). Mean (Year) or the mean publishing year indicates the average publishing year of the references in that cluster. The clusters are arranged in such a way to show which clusters appeared first and which appeared later.

Of all the clusters, the second biggest cluster (ID number 4) contains 202 papers in it, and has systems-oriented labels, such as “parameters” and “stochastic”. These labels suggest that the articles in this cluster are related to the general properties of complex systems. One detail to note is that the cluster was formed around 2003, relatively early stage. From 2007 on, more clusters have labels relate to empirical application of biology, such as “stem,” and “pluripotent” for cluster 10, “vaccines” for cluster 8, and “cancer” for cluster 11, and the years in which these clusters emerged are more recent. This phenomenon suggests that recent trends in systems biology are approaching empirical biological problems.

Table 4: The Clusters arranged according to the mean (year) in descending order.

ClusterID	Size[a]	mean(Year)	Label (LLR)
13	39	2002	pulmonary (28.66, 1.0E-4)[b]; pharmacogenetics (21.49, 1.0E-4); pathogen (21.49, 1.0E-4);
4	202	2003	stochastic (283.42, 1.0E-4); death (179.88, 1.0E-4); parameter (179.88, 1.0E-4);
5	390	2004	network (682.13, 1.0E-4); networks (401.17, 1.0E-4); gene (388.84, 1.0E-4);
7	41	2004	proteomic (147.39, 1.0E-4); proteomics (130.59, 1.0E-4); proteome (112.63, 1.0E-4);
3	126	2005	metabolomics (896.03, 1.0E-4); metabolome (248.31, 1.0E-4); metabonomics (195.96, 1.0E-4);
6	101	2005	metabolic (524.36, 1.0E-4); genome-scale (425.1, 1.0E-4); production (246.8, 1.0E-4);
10	7	2007	stem (73.73, 1.0E-4); pluripotent (50.98, 1.0E-4); unique (40.76, 1.0E-4);
9	6	2007	sequencing (37.46, 1.0E-4); pitfalls (27.46, 1.0E-4); possible (27.46, 1.0E-4);
12	3	2007	microrna (89.16, 1.0E-4); genomes (34.89, 1.0E-4); targeting (34.89, 1.0E-4);
8	19	2008	vaccine (110.81, 1.0E-4); vaccinology (102.26, 1.0E-4); immune (70.18, 1.0E-4);
11	23	2009	laparoscopic (313.93, 1.0E-4); cancer (257.46, 1.0E-4); surgery (216.36, 1.0E-4);
0	2	2010	granularities (11.85, 0.001); answer (11.85, 0.001); description (11.85, 0.001);

Notes: [a] The size of a cluster indicates the number of references in that cluster. This could be translated into the importance of a cluster, as for example, Cluster 4 is much more important than Cluster 1. [b] Each cluster was labeled with information retrieved from the references using the Log Likelihood Ratio (LLR) algorithm, a statistic test algorithm that picks the terms that best represent a specific cluster (Chen, 2009). The first number in the

parenthesis after each label indicates the LLR score (the higher the better), and the second number is the p-value (the lower the better).

2.2.2. Research types of the most highly cited references

CiteSpace automatically generated a list of the most highly cited references, and the top five most highly cited references are shown in Table 5 below. We use the top five most highly cited articles to avoid overlap and get a clearer signal. The table lists the authors, publishing years, titles, journal names and categories of the references. We will first discuss the results of the top-five most highly cited references from the year 1992 to 2013 as an example. Later, we discuss the 330 articles and the changes of number in each category over time.

Table 5: The top five most highly cited references.

Rank	Citations	First Author	Year	Title	Category	Systems-oriented or Biology-oriented
1	737	Kitano H	2002	System Biology: A Brief Overview	Hard to tell	
2	510	Hucka M	2003	The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models	Software development	Systems-oriented
3	465	Ashburner M	2000	Gene Ontology: Tool for the Unification of Biology	Database development	Systems-oriented
4	456	Barabasi AL	2004	Network Biology: Understanding the Cell's Functional Organization	Network Properties	Systems-oriented
5	439	Shannon P	2003	Cytoscape: A Software Environment for Integrated Models of Biomolecular Interation Networks	Software Development	Systems-oriented

An interesting observation here is that the top five most highly cited references are more systems-oriented as judged by reading the paper carefully. For example, from Table 5, it was found that the engineer Kitano, published the most cited article in 2002. In that article, Kitano offers an overview of what he thinks is systems biology is, focusing on the computational perspective (Kitano, 2002)[17]. The second article is a description of the Systems Biology Markup Language, an XML-based language used to describe models so they can be used in a common software platform (Hucka et al., 2003). The third article is about Gene Ontology, a database storing genes and the attributes of genes through a unified representation (Ashburner et al., 2004). The fourth article was written by the physicist Barabási to introduce network theory and try to apply that to biological networks (Barabási & Oltvai, 2004; Barabási, 2014). All of his discussion about biological networks is theoretical, and therefore, systems-oriented. The fifth article introduces a digital tool for the visualization of biological networks, Cytoscape (Shannon et al., 2003).

The above table shows the result of the top five most highly cited references for the entire field of systems biology in the entire time period. With the same method we analyzed the category of all 330 most highly references from 1992 to 2013, as we want to explore patterns of temporal change.

[17] Except Kitano's article is identified as hard to tell, other four articles can be easily put in a category.

CiteSpace also picked out 18 books out of the most highly cited references. Those books were mainly published in the early years. Especially between the year 1992 and 1993, 8 out of the 30 most highly cited references are books, for example, Stuart Kauffman (1993)'s *The Origin of Orders: Self Organization and Selection in Evolution*. This book incorporates the findings in physics, chemistry, and mathematics to study the origin of organism as a complex adaptive system, in which computer simulations are used to model how self-organization occurs. The number of books among most highly cited references decreased after 2000, with only two books have high citations that ranked among top 30. Those two books are Erberhard Voit (2000)'s *Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists* and Bernhard Palsson (2006)'s *Systems biology: properties of reconstructed networks*.

Excluding books and references that are hard to label[18], the results of these 330 references that fall into biology-oriented and systems oriented categories are demonstrated in Figures 10 and 11, respectively. Each line represents a category.

Figure 10 show the number of references in four biology-oriented categories: omics research, development of high-throughput technologies, the application of systems biology, and biological mechanisms. The development of high-throughput technologies is the category that has the highest number of articles in early stages, but this category

[18] Out of the 330 articles, CiteSpace picked out 14 references that could not be placed into these nine categories. They either focus only on topics tangentially related to systems biology, or are hard to put into any category, so they were not included in the nine categories.

decreased afterwards. This suggests that early biologists involved in systems biology by developing new technologies.

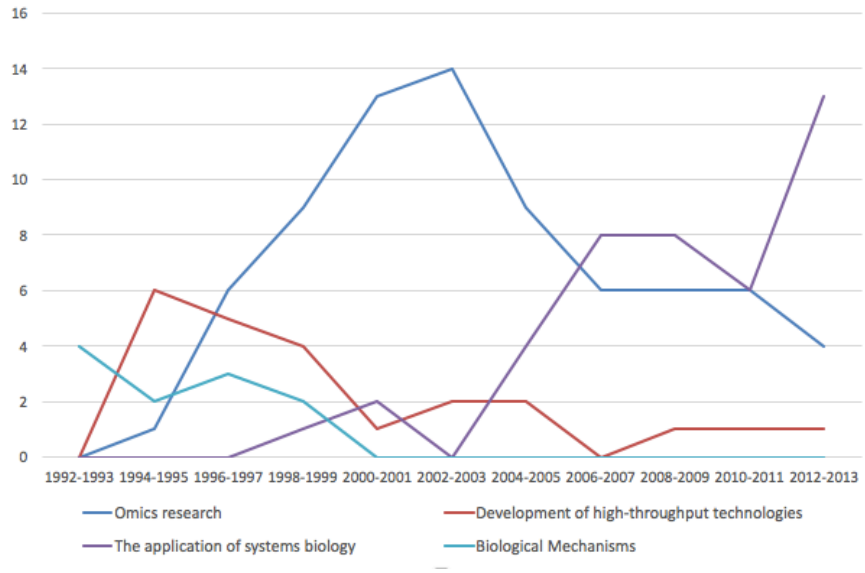


Figure 10. The number of most highly cited articles in biology-oriented research categories. The *x* axis stands for the 11 time slices and the *y* axis stands for the number of articles among 30 most highly cited articles for a category.

Omics research (blue line) began to emerge in the late 1990s and peaked around 2002. Omics research has changed from simply getting the sequence of a genome in the 1990s to actually mapping the interactions of biological molecule, be it proteins, or genes, or metabolites after 2000. An example of an article in this category is the comprehensive study of protein-protein interactions in yeast. Peter Uets and his colleagues discovered 957 possible interactions of more than 1000 proteins (Uetz et al. 2000). After that omics research decreased while the application of systems biology (the purple line) began to increase in early 2000s, and reached a plateau from 2006 to 2010.

Then in the last four years, the rise of systems biology's application in medical fields and engineering became very significant. For example, one such most highly cited reference in 2010 is on systems vaccinology, describing how systems approach has changed the way scientists develop vaccines (Pulendran, Li, & Nakaya, 2010). Despite the fact that vaccines work well in preventing diseases, the mechanism for how they work remained largely unknown before application of a systems approach. Scientists are now starting to use systems approaches to identify the gene regulatory network after the injection of vaccines, and predict the later responses. It can help identify high-risk individuals and prevent potential harmful consequences of vaccine to those individuals (Pulendran, Li, & Nakaya, 2010).

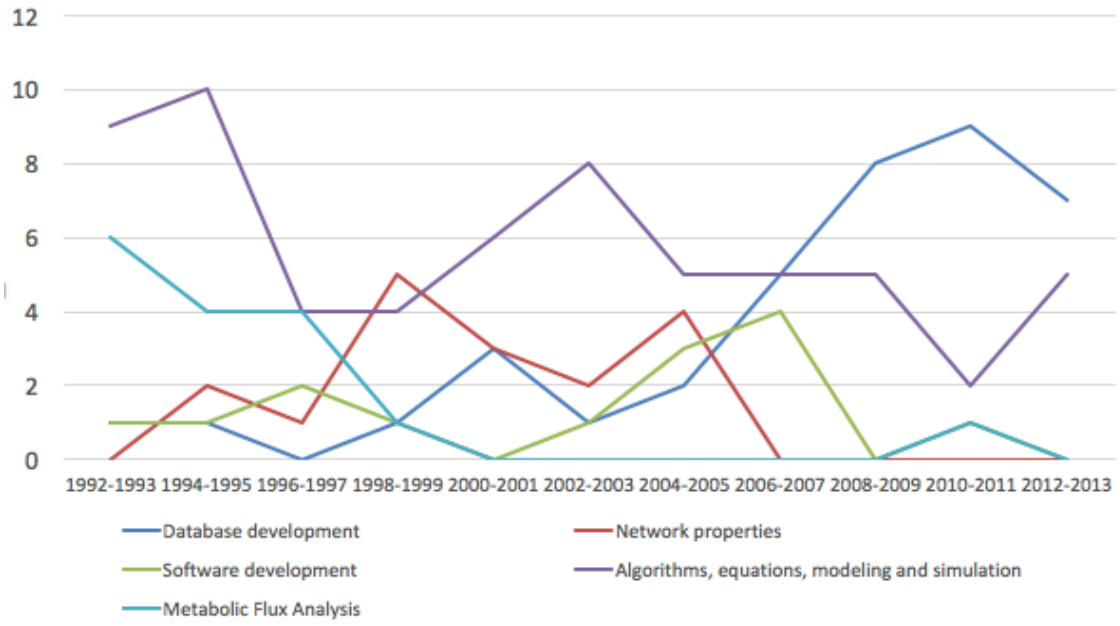


Figure 11. The number of most highly cited articles in systems-oriented research categories. The *x* axis stands for the 11 time slices and the *y* axis stands for the number of articles among 30 most highly cited articles for each time slice.

Figure 11 shows the number of references in five systems-oriented categories. From 1992 to 1995, two categories, namely, metabolic flux analysis (purple line), algorithms equations and models (light blue line), are the top two categories. However, only the category of algorithms, equations and models maintains the same importance, and metabolic flux analysis decrease over time. This suggests that algorithms, equations and models are still very central to systems biology research, which is claimed by other scholars (Machado et al., 2011).

Database management (Dark blue line) emerged around 2000 and continue to be a strong presence in later years, including the Gene Ontology and KEGG database discussed earlier, but also BioGRID, Reactome, BiGG, IntAct, to name a few, which are more recent databases (Stark et al., 2006; Matthews, 2009). A database is not a place where biologists dump their data, because scientists need to figure out how to store the data, how to search for data quickly, how to manage database structure, and how to develop a standard of data format that is compatible to more databases (Leonelli, & Ankeny, 2012). This knowledge can be classified as data science in general, which requires the input of systems-oriented scientists.

The overall trend of biology-oriented research was proceeding strongly as systems-oriented research was becoming weaker. In Figure 12, two lines represent system-oriented and biology-oriented research respectively. From 1992 to 1995, systems biology was more systems-oriented than biology-oriented. After that time, systems biology became more biology-oriented. Figure 12 suggests that two special turning points were the years 1996, when systems-oriented research was superseded by biology-oriented research and 2012, when the gap between systems oriented and biology-oriented research

became bigger than ever. Especially from 2012 to 2013, the articles became predominantly biology-oriented. For the exact number of 330 most highly cited references in different categories over time, see Appendix A-3.

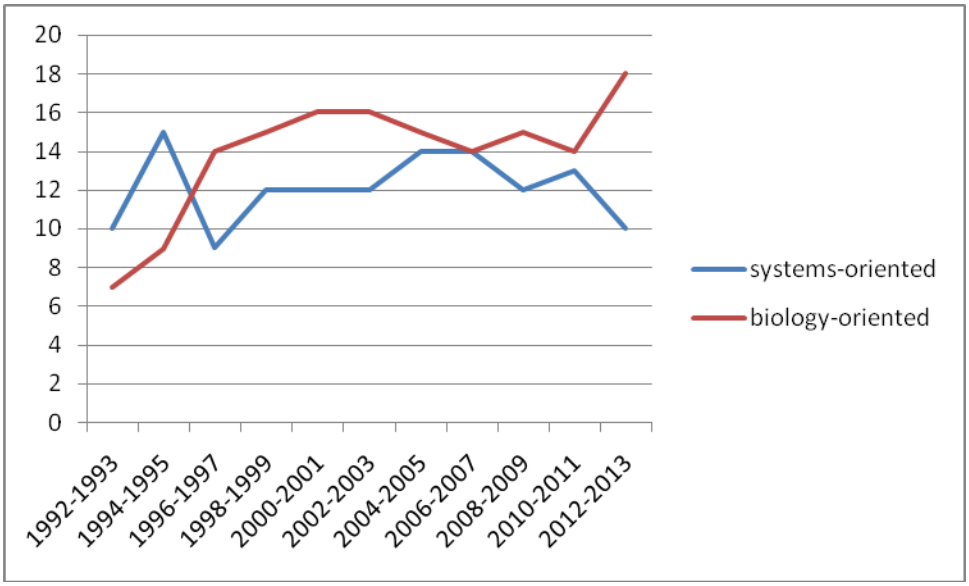


Figure 12. The trend of biology-oriented (red line) and systems-oriented (blue line) articles. The *x* axis stands for the 11 time slices and the *y* axis stands for the number of references among 30 most highly cited articles for each time slice.

2.2.3. The institutional contexts for systems biologists

Table 6 lists the affiliations of the first most highly cited authors in each time slice. The affiliations show the department, the university or institution, and geographical information such as city and country. The table shows that the most highly cited authors in each time slice changed quickly throughout the years, suggesting that systems biology was evolving quickly. The institutions they are affiliated are very diverse and interdisciplinary. We classified the institutions of the top 30 authors from 11 time slices

into four categories: biology-oriented, systems-oriented, interdisciplinary and systems biology institutions.

Table 6: The most highly cited authors.

Time Slice	The most cited author in each time slice	Authors' Institutes	Category
1992-1993	KAUFFMAN S	Department of Biochemistry and Biophysics, School of Medicine, University of Pennsylvania and Sante Fe Institute, Sante Fe, New Mexico, U.S.A.	Interdisciplinary/Systems Biology
1994-1995	BENJAMINI Y	Department of Statistics, School of Mechanical Studies, Tel Aviv University, Tel Aviv, Israel	Systems-oriented
1996-1997	HEINRICH R	Theoretical Biophysics Group, Institute for Biology, Humboldt University Berlin, 10115 Berlin, Germany	Interdisciplinary
1998-1999	HARTWELL LH	Fred Hutchinson Cancer Center, Seattle, Washington 98109, USA.	Biology-oriented
2000-2001	IDEKER T	Inst Syst Biol, Seattle, WA 98105 USA.	Systems biology
2002-2003	KITANO H	Sony Comp Sci Labs Inc, Tokyo 1410022, Japan.	Systems-oriented
2004-2005	BARABASI AL	Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556, USA	Systems-oriented
2006-2007	ALON U	Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 76100, Israel.	Biology-oriented
2008-2009	FEIST AM	Department of Bioengineering, University of California, San Diego, La Jolla, California 92093, USA.	Interdisciplinary
2010-2011	ROUKOS DH	Univ Ioannina, Biosyst & Synthet Genom Network Med Ctr BioSynGen, Ioannina, Greece.	Systems biology
2012-2013	ZHANG AH	Heilongjiang Univ Chinese Med, Natl TCM Key Lab Serum Pharmacochem, Key Lab Chinmed, Dept Pharmaceut Anal, Harbin 150040, China	Biology-oriented

Biology-oriented institutions have words like “cancer”, or “genetics” that are related to the life sciences. We noticed that many authors are from a medical institution, and even pharmaceutical companies, like Glaxosmithkline and Syngenta.

Systems-oriented institutions include those related to “statistics,” “mathematics,” “physics,” “chemistry,” and other non-biology disciplines. We also found that some researchers are from the industry, such as Microsoft, Siemens, and Sony.

“Interdisciplinary institutions” refer to those that related to interdisciplinary field such as “biochemistry,” “biophysics,” “bioengineering,” “biotechnology,” and “bioinformatics.”

Systems biology institutes include those that specifically use the words like “systems biology,” or “biosystem,” or “biosyst.” For some institutions that are hard to tell which category they belong to, we did not label them and include them in the calculation.

The results of the categorization of the affiliations of 330 highly cited authors are shown in Figure 13. In each time slice, the top 30 authors’ institutions were plotted in four color-coded bars to represent the four categories.

The figure shows that the number of authors from systems-oriented institutions (red bar) was first almost the same as biology-oriented institutions, but their number diminished over the years. For example, in the time frame of 1992 and 1993, one of the most cited authors is Daniel T. Gillespie, a physicist working at the Research Department of Naval Weapons Center at the time. His work on stochastic simulation in chemical kinetics contributed to the simulation method adopted by later systems biologists (Gillespie 1992). He contributed to the foundation of systems biology while this discipline was still in its “early roots” stage, and naturally he did mention systems

biology in his work and probably did not know the term. However, he published an article in 2008 on simulation methods in systems biology.

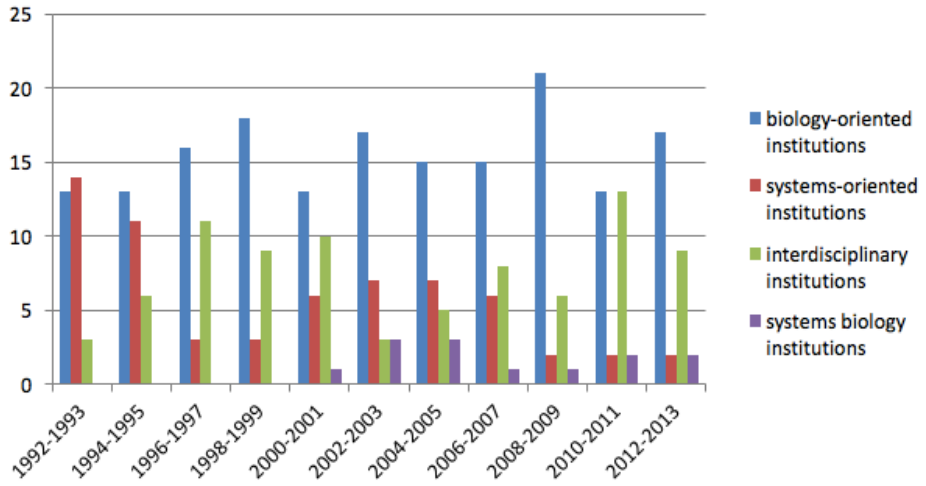


Figure 13. The most highly cited authors' institutions. The *x* axis stands for the year and the *y* axis stands for the number of authors in a category among the top most highly cited 30 authors.

Scientists from interdisciplinary institutions (green bar) have fueled the advancement of systems biology not only in the early days of systems biology, but also in more recent years. These institutions have provided a place for early systems biologists to stay. While in the 1990s, there were no systems biology institutions (purple bar), which only emerged within the slice from 2000 to 2001. In each time slice after 2001, there were a few authors coming from systems biology institutions. The number of the most cited authors from a biology-oriented institution (blue) tends to fluctuate over the years, but they have the highest numbers compared to other categories in every slice except in the time slice from 1992 to 2013.

Figure 14 shows the result of all the first authors who published between 2003 and 2013 retrieved from WoS citation data[19]. After taking out the unidentifiable authors' affiliations, it shows that the percentage of each category has remained quite constant, which means that from 2003 to 2013, the institutional context for systems biologists did not change much.

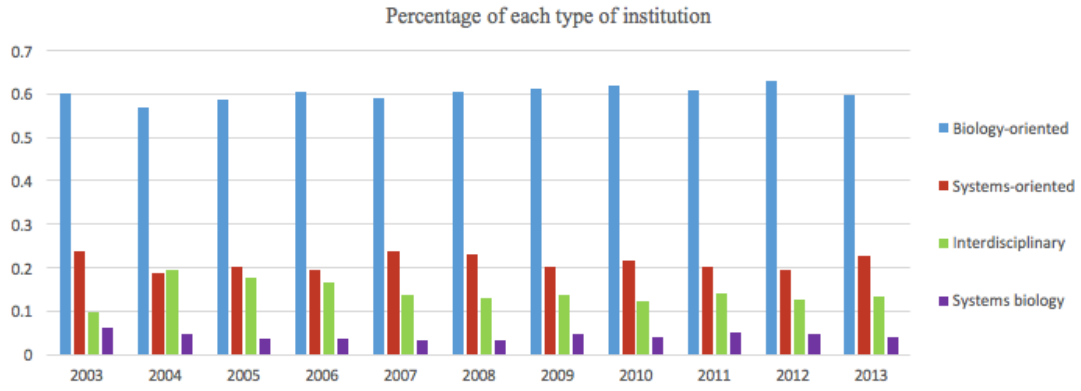


Figure 14. All authors' institutions from 2003 to 2013**.** The number in each category changed little from 2003 to 2013 as shown in four color-coded bars representing the four categories. The *x* axis stands for the year, and the *y* axis stands for the percentage of each type of institution.

Between the years 2003 and 2013, on average 60.85% of authors came from a biology-oriented institution; 21.16% came from a systems-oriented institution; 13.75 % were from an interdisciplinary institution; and 4.23% were from a systems biology institution. These statistics suggest that the institutional context for all authors publishing on systems biology is different than that for the people who published the most highly cited articles on systems biology. Notably, the latter had a larger percentage of authors

[19] Out of all 9876 authors who published between the years 2003 and 2013, a total of 779 (7.89%) were unidentifiable. We excluded those articles when calculating the percentage of four categories.

who were affiliated with interdisciplinary and systems-oriented institutions. This is an interesting observation that the people who were cited most and the people who published in a field have different patterns in terms of their affiliation, and we will discuss that in our conclusion.

2.2.4. Topics found in the abstracts

We were not only interested in the research types of the most highly cited references, but also in general systems biology articles. The result of topic modeling based on thousands of articles is similar to my manual reading of the most highly cited articles. Appendix C-1 shows the machine learning results of topics using MALLET and the labels that we assigned by reading the words in topics. The machine learning model returns the following 20 topics: biology, models, metabolomics, diseases, proteomics, synthetic biology, database and software, cell biology, systems theory, algorithms, immune system, network properties, network, genomics, technologies and tools, drug and cancer, regulation, pathway.

We were more interested in the temporal change of topics. Figures 15 and 16 show the topics that have significant patterns of increasing or decreasing. Figure 15 shows that the percentage of articles that contain Topics 11, 14 and 17 increased over time. Topic 11 is about the research related to immune systems and vaccines, Topic 14 is about disease, and Topic 17 is about drugs and cancer. There are related to the application of systems biology, which is similar to my finding about the articles in this category among the most highly cited articles in section 2.2.2, which showed that they are more biology-oriented.

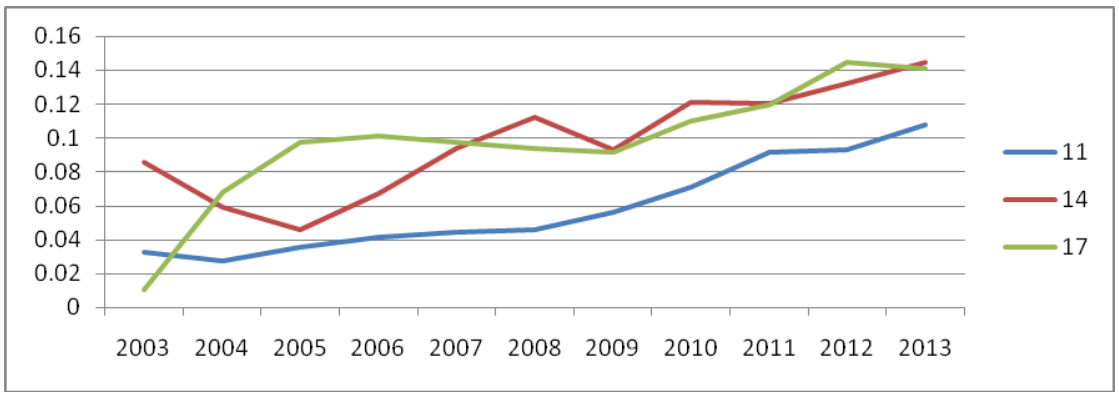


Figure 15. The trends of Topics 11, 14, 17. The *x* axis of the graph represents the year, and the *y* axis represents the percentage of articles published in that year that contain a topic with a probability higher than ten percent.

Figure 16 shows how the percentage of articles that contained topics 9, and 16 decreases. Topic 9 includes some general terms about systems biology, and Topic 16 is about high-throughput technologies, which includes words like "high," "throughput," "technologies," and "techniques." The result is similar to my findings about the percentage of articles in the category of high-throughput technologies among the most highly cited articles, which is decreasing over the years. The percentage for articles that contained each of the 20 topics is shown in Appendix C-2.

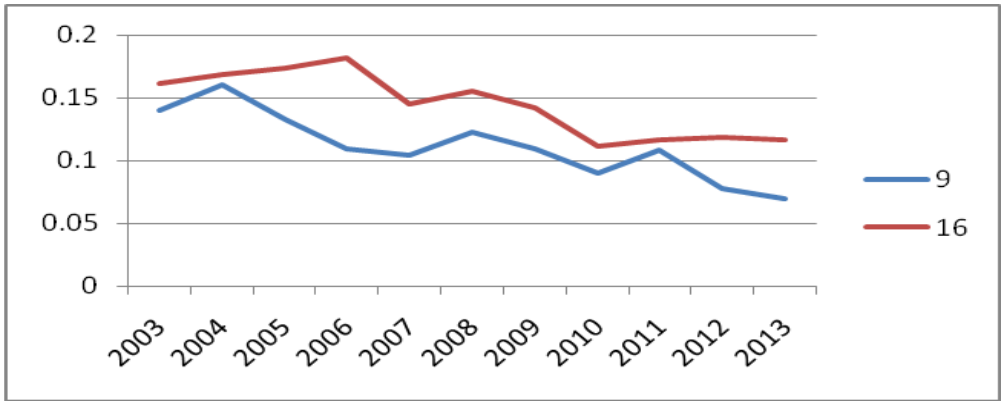


Figure 16. The trends of Topics 9 and 16. The *x* axis of the graph represents the year, and the *y* axis represents the percentage of articles published in that year that contain a topic with a probability higher than ten percent.

2.3. Conclusions and Discussion

Systems biology is a new discipline, and this study offers a bibliographic analysis of its history from 1992 to 2013. We examined the research categories as well as the institutional contexts of researchers in the field of systems biology. Bibliographic analysis allowed us to pick the most highly cited authors and references, because it can be assumed that they have bigger contributions to the field than the less well cited authors or articles. We carefully categorized those authors and articles through close reading, but also applied machine learning technique to analyze a much larger number of articles and authors to overcome the limitation that we could not read them all.

The main conclusion is that systems biology has become more biology-oriented in research categories and the most cited authors. Our findings are echoed in some other scholars' observations about systems biology. Alan Aderem, who works at the ISB, argues that "biology dictates what new technology and computational tools should be developed, and once developed, these tools open new frontiers in biology for exploration. Thus, biology drives technology and computation, and in turn, technology and computation revolutionize biology" (Aderem, 2005). Some of them suggest that biology needs to be more important in systems biology in two aspects. First is the claim that more biologists should be involved in systems biology, and second that more empirical biological problems need to be addressed (Calvert & Fujimura, 2011). Jane Calvert and Joan H. Fujimura carried out interviews with many researchers in systems biology,

including biologists, mathematicians, physicists, and computer scientists. One of the biologists said: "I think biologists need to drive systems biology, because if it's driven by computation or engineers, without a depth of training in biology, they lose that sense, they tend to treat molecules as nodes and edges without a sense of how they're performing their functions" (Calvert & Fujimura, 2011, p. 161). Physicists and computer scientists might also agree with that, because often those who want to model a biological system cannot find a biological expert to help link the model to specific biological problems (Hlavacek, 2011).

Along with the finding of the interviews, a systems biologist at the ISI, Sui Huang, claims that biologists have become active players in systems biology because what they need to understand now is not a single gene or a protein, but networks of genes or proteins, and systems biology approach can help address their needs (Huang, 2007). Huang's article claims that systems biologists should divert their research "back to biology in systems biology." Systems biology has captured more biological phenomena, properties, objects such as various regulatory pathways, data, theories, and methods than its precursor did in the middle of twentieth century. Biologists have seen the utility of applying a systems biology approach to understand the evolution and function of biological networks.

Another interesting finding for biologists is the upward trend toward application in systems biology. The application has been centered around understanding cancer better, transforming drug discovery, and making preventative vaccines. Previous studies might have mentioned the potential of systems biology in application based on a few articles about this field, but this study was able to quantitatively and qualitatively provide

data for this intuition through the percentage of articles in this category, the trends of topic based on topic modeling, and the result of cluster labeling. In choosing their future topics, biologists should be thinking about picking an application-oriented research.

Our result about the institutional context of scholars who published in this field also has implication for policy makers or funding agencies. The result shows that scholars from an interdisciplinary field or a field outside of biology have contributed greatly to systems biology, in terms of first, the contrasting percentage of those scholars in the most highly cited authors compared to percentage of them in all authors, and second, the big influence of systems oriented research types in the early stages when systems biology started to emerge. The institutional contexts of the most cited authors and general authors in this study suggest that the scholars who lead a field are sometimes different from those who publish in that same field. It has been suggested by previous scholars that we should create interdisciplinary environment on purpose, and our results agree with that suggestion (Lattuca, 2001). That result that the percentage of scholars from an interdisciplinary field or a field outside of biology could be translated into a message that the funding agencies should prioritize the funding of interdisciplinary projects and institutions, because they may better lead to the starting of a new discipline, or making a higher impact.

Equally important is the methodology presented in this paper. First, utilizing computational tools allowed us to expand our research years to the period when the term "systems biology" was not yet invented. Although the first article that has the term "systems biology" was not published until 1997, we can find the references that

contributed to systems biology through bibliographic analysis tools, and also identify individuals who are the most highly cited.

Second, this methodology is an example of combination of close-reading and distant reading, as exemplified in other computational history of science projects (Laubichler, Peirson, & Damerow, 2013). The use of CiteSpace and Tethne allowed us to select the most highly cited references and authors to represent the field using computational approaches and build-in functions, and apply close reading on a selected few, so we don't need to read all the articles. Later, the automatic labeling of clusters Citespace based on LLR algorithm, and topic modeling employ statistical models to understand the topics within the literature, which is a form of distant reading. Close-reading and distant reading can benefit from each other. For example, the result of close reading of most highly cited authors allows us to retrieve the identifying word for designing the model that studies all the authors. The results of machine learning technique can be viewed in parallel with the results from manual reading; the results of manual reading or close reading can either support or refute the results of machine reading.

Third, computational history of science blurs the line between natural sciences and humanities. Natural sciences often involve a hypothesis that is testable by experiments, having quantitative, precise, and objective results that are repeatable, whereas humanities are perceived by some in an opposite way, for example relying more on narrative style and speculative method (Gardiner & Musto, 2015). Our research is on history of science; however, the result in this research is quantitative, and most of these results are repeatable using the same WoS data and same parameters.

For historians, the approach used in this chapter can also be used to provide a historical perspective for literature in other fields of research. For historians who don't know how to use programming language, the easier parts would be downloading the metadata from the WoS database, and learning how to use CiteSpace to perform citation analysis, or MALLET to perform topic modeling. For historians with programming skills, they can find the online tutorial for Tethne and the codes used in this research are shared online. This approach is not merely a combination of tools that are developed by other researchers, but the mastery of Python allows us to design our own flowchart to perform tasks tailored to our own needs, and connect the WoS data with tools.

CHAPTER 3: UNDERSTANDING SYSTEMS BIOLOGY'S CONCEPTUAL HISTORY USING CO-WORD NETWORKS

Abstract: Systems biology studies complex biological systems such as gene regulatory networks. This chapter utilizes computational approaches to understand how knowledge of the field of systems biology changed from 2003 to 2013 through co-word/concept analysis, with "co-word" meaning words occurring in close proximity in the same sentence. Our research provides an example of how big data science can be used to study the change of concepts. We retrieved the available abstracts of a total of 9876 systems biology articles published between 2003 and 2013 from Web of Science and build a corpus of over 5 million words. We used co-word networks of 300 keywords identified by corpus linguistics techniques to represent the knowledge of systems biology. We discovered that the majority of biology-oriented words have increased in centrality in the co-word/concept network over time. We visualize the sub-network of "cancer" as an example to show that the words that co-occur with a word can visualize the conceptual change of that word.

Keywords: Systems biology; Co-word analysis; Network Text Analysis; Corpus Linguistics; History of Science; Text-mining.

3.1. Introduction

In systems biology, scientists from different backgrounds have different epistemic goals. For example, engineers might be more interested in generating rules that apply to all systems, whereas biologists care more about specific biological problems (Calvert &

Fujimura, 2011). Scientists from a field other than biology often focus on concepts such as "feedback loops" and "circuit," which we call "systems-oriented" concepts, as opposed to concepts like "evolution" and "development," which we call "biology-oriented" concepts. Concepts are abstract entities that are central to the epistemology of science, and philosophers have traditionally used technical ways to analyze the conceptual structure of science, such as the semantic method of analyzing the predicates and propositions in a sentence structure (Carnap, 1991).

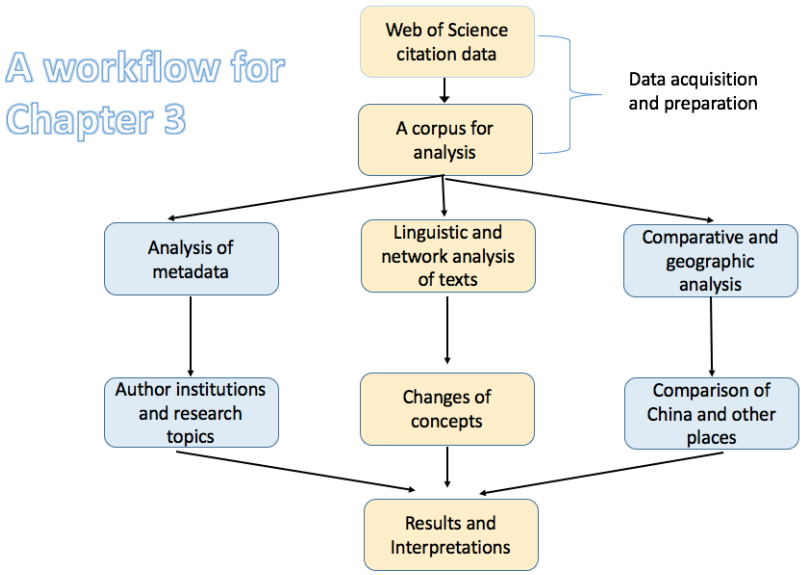
More recently, scientists have argued that knowledge structures can be modeled as networks (Sowa, 1984). According to Popping (2000), "When concepts are depicted as networks, one is afforded more information than the frequency at which specific concepts are linked in each block of text; one is also able to characterize concepts and/or linkages according to their position within the network" (p. 30). Our assumption that observing how a word links to other words that co-occur with it in texts over time can reveal the conceptual change of that word is based on such a network view. Collectively, such relational information of many words in a large network can give a general picture of the entire scientific field, and adding a temporal dimension to the networks can reveal the historical change of that knowledge field.

Here we aim to answer the following questions: How did scientists use biology-oriented words and systems-oriented words differently between 2003 and 2013 in the systems biology literature and how can we visualize and quantify these changes? To answer this question, we analyzed the abstracts of 9876 systems biology research articles retrieved from the Web of Science (WoS) database, and generated networks to visualize the knowledge embedded in those abstracts. One innovation of the research in this

chapter is combining two computational approaches, a corpus linguistic approach and a network approach. This research serves two purposes: We are not only interested in the historical change of concepts in systems biology, but also have the historiographical motivation to explore how computational methods can transform the way we represent change of concepts.

3.2. Methodology

We first describe what we mean by a co-word/concept network in detail and then introduce the steps that we took to generate co-word networks: a) building a corpus, b) generating a codebook, c) text-mining the corpus using the codebook, and d) visualizing, analyzing, and comparing the networks. The four processes are shown in the flowchart in Figure 17, and we introduce the steps in detail in the following sections.



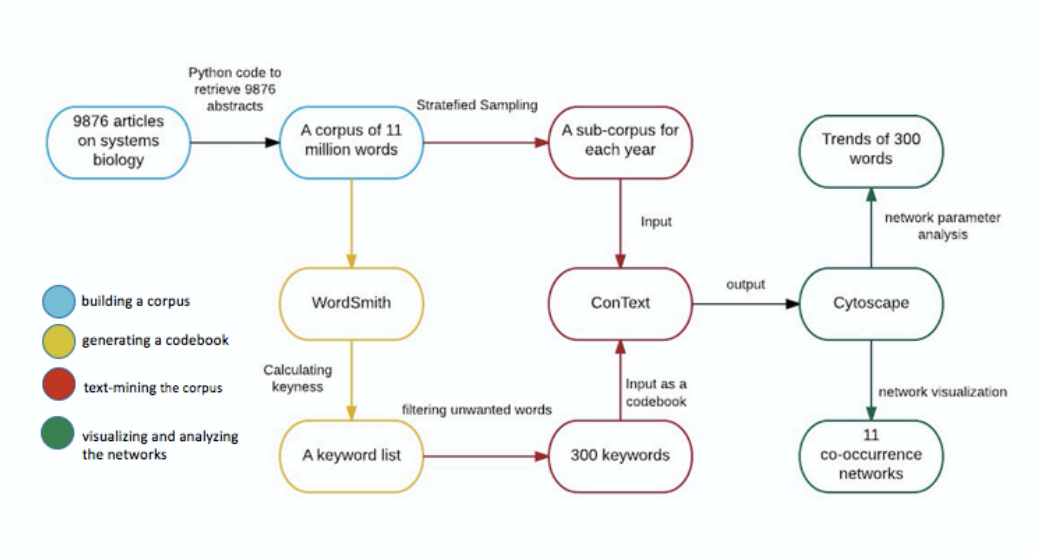


Figure 17. The flowchart for Chapter 3. The figure on the top highlights parts carried out in this chapter. The figure on the bottom shows the detailed steps in Chapter three.

3.2.1. What is a co-word/concept network?

A co-word/concept network is a network composed of words that co-occur in either titles, abstracts, or main texts (He 1999; Callon, Courtial, Turner, & Bauin, 1983). In our case, we examined words that co-occur in abstracts of research articles because research articles are a genre that plays a pivotal role in scientific discourse (Tessuto, 2015). Scientists have not reached a consensus on the naming of such a network. Some might also call it a lexical co-occurrence network (Edmonds, 1997), or a word co-occurrence network (Veling & Van Der Weerd, 1999), or a collocation network (Lahiri, Choudhury, & Caragea, 2014). A co-word network built based on the literature from a scientific field can shed light on the knowledge structure of that field (Ferrer & Solé, 2001). While we focus on words that represent concepts meaningful for systems biology, we could call our network a co-concept network, but for simplicity we will refer to it as co-word network from now on.

Co-word network analysis has been used by researchers from a great variety of disciplines, including health care research (Jang, Lee, & An, 2012), nanotechnology development (Darvish & Tonta, 2016), and strategy management (Ronda-Pupo & Guerras-Martin, 2012), to name a few. Previous research shows that co-word analysis has been used by humanists to represent the field they study as well. For example, researchers use keywords that co-occur in an article to build networks in order to map the knowledge structure of "technology foresight" (Su & Lee, 2010). In another example, researchers looked into how the adjectives that co-occur with "old," "young," "female," and "male" in fairy tales reveal the societal values that shape the gender identity in kids (Weingart & Jorgensen, 2012). They discovered that the young males have the fewest adjectives that are used to describe them, which suggests that young males are less described and assumed a universal position. As far as we know, our research is the first co-word network research on systems biology literature.

Co-word network analysis, along with co-citation analysis, co-author analysis, and co-journal analysis are the most widely used techniques in scientometrics and bibliometrics (Scharnhorst & Garfield, 2010; He 1999). In our previous study (see Chapter 2), we did co-citation analysis. Here we want to point out that one advantage of co-word network analysis over co-citation analysis or co-author analysis is that the nodes of a co-word network are words that won't change and we can easily label the node with the word itself, whereas the nodes of a co-citation network or a co-author network are a paper or a scientist (He, 1999). If we want to know who the scientist is, that requires close-reading, for example looking up their information in Google. Second, co-word analysis enables one to look closely at one part of a network as well as looking at a

network at a whole, thus combining macro-level and micro-level of examination of the knowledge structure of a field (He, 1999). Third, a word itself does not change, and only the meaning changes over time. Therefore, we can compare a word in different times, with its meaning represented as the co-words of the word.

Our analysis can also be categorized under Network Text Analysis (NTA). NTA is "a semi-automated knowledge discovery technique in which entities and their relations are extracted from unstructured texts (e.g., newspaper articles, interview transcripts)" (Martin, Pfeffer, & Carley, 2013, p. 1166). NTA is a type of computer-assisted graph-based knowledge representation that has been deemed central for artificial intelligence (Chein & Mugnier, 2008). Our approach also can be categorized under data-mining techniques, because Ian H. Witten and Eibe Frank (2005) define data-mining as using computer-assisted methods to retrieve information that has a clear structure and predictive value from unstructured data and in our research it is clearly the case. In our research, we used a way to generate networks through readily available software, which can be mastered without extensive technical skills; thus, our approach is especially useful for humanists.

3.2.2. Steps taken to generate co-word/concept networks

We built a corpus consisting of the available abstracts of 9876 systems biology research articles published from 2003 to 2013 from the WoS database, which will be explained later. We then generated a machine identified keywords list using a corpus linguistics tool and manually analyzed which keywords are biology-oriented keywords and which are systems-oriented keywords. Next we generated co-word networks of the above keywords over time, and examined the changes of the centrality of 300 keywords

over time using Python code. As an example to show how a co-word network can visualize the conception of a word on a finer scale, we look at the sub-network centered around the keyword “cancer.”

3.2.2.1: Building a Corpus Representing the Field of Systems Biology

The research in this chapter is a corpus-based analysis. A corpus is a collection of textual data. More and more corpora have been created for certain languages and for specialized scientific domains. For example, the Genia Project has built a corpus containing 2000 annotated Medline abstracts of research articles on molecular biology (Ohta, Tateisi, & Kim, 2002). For another example, Telecommunication Engineering Corpus contains 5.5 million words of the professional and academic written English on Telecommunication Engineering (Rea, 2010). A corpus usually contains millions of words so that one can derive quantitative and statistical results from it, instead of relying on just a few selected texts.

An important question to ask before building a corpus is how we can ensure that a corpus is representative (Hettel, 2013). It is impossible to download all the research articles on systems biology, and as Douglas Biber (1993) points out, one must balance efficiency and cost effectiveness against higher level of representativeness. We consider our sampling frame to be representative because it includes all the research articles on systems biology in a well-curated and large database, which is explained in the next paragraph.

We chose Thomson Reuters’ WoS database because it is a well-curated database with credible journals, and contains structured data for each research article (Falagas et al., 2008). A search in the WoS database for articles published between the years 2003 to

2013 that included the term “systems biology” in their topics (including titles, abstracts, and keywords) returned 9876 articles. We downloaded the full bibliographic records of all 9876 articles in text files, and then used a Python code to extract the abstracts of all the articles to build a corpus. We chose to analyze the abstracts because the access of full-texts is time-consuming in many databases currently and requires more computational power to analyze them, and many corpus-based analyses choose abstracts to analyze (He, 1999). We then divided the corpus into 11 sub-corpora according to the publication year. The corpus size is 11, 821,632 words, and the sizes for the sub-corpora are listed in Table 7.

Table 7: Corpus sizes for each sub-corpus.

Year	Number of Abstracts	Corpus size
2003	93	104538
2004	219	254765
2005	369	430705
2006	533	665418
2007	606	773700
2008	800	1033792
2009	995	1338717
2010	1225	1685112
2011	1286	1774022
2012	1369	1907143
2013	1314	1853720

Because the corpus size for each sub-corpus differs significantly, we randomly picked 90 abstracts for each year and created 11 new sample sub-corpora, with each corresponding to a year. This is called stratified random sampling of sub-corpora, because the population of sampling units are divided on the basis of time (Hettel, 2013).

Compared with simple random sampling, stratified random sampling ensured that each year the co-word networks generated based on them could be compared on the same basis.

3.2.2.2: Lexical Profile Analysis to Generate a Codebook.

Detecting co-word relationships between words of an existing word list is a common way to generate a co-word network (Bullinaria & Levy, 2007). The word list, often referred to as a codebook, can be a dictionary, an ontology, a thesaurus, or a keyword list (Bullinaria & Levy, 2007). The choosing of the word list is critical to the results of co-word analysis and one wants to make sure that the word list contains the most important words that characterize a scientific field (He 1999). Previous studies have used many different approaches to building a word list: manual picking a word list, tf-idf[20], using MeSH[21] terms, etc. The innovation of our research is that we used corpus linguistic approach to ensure that our word list is not randomly picked and can indeed represent the field of systems biology. The tool that we employed is a widely used tool of corpus linguistics: WordSmith (Scot, 1996).

Corpus linguistics is a branch of linguistics that studies linguistic features such as collocates, word list, and keywords "based on examples of real life language use"

[20] Tf-idf stands for term frequency–inverse document frequency, which is a text-mining approach to determine how important or unique a word is to a document compared with to other documents in a collection of documents.

[21] MeSh stands for Medical Subject Headings, a thesaurus that facilitates indexing and search of vocabulary in the life sciences.

(McEnery & Wilson, 2001). Collocates are the words that are habitually juxtaposed. A word list is a list of words arranged according to the frequency of their occurrences in texts. For example, researchers analyzed the frequency of the word "experiment" in early literature to reveal that experiment was used more frequently in English religious literature than in scientific literature before the 1800s and the meaning before 1800 were different from its current meaning (Pumfrey, Rayson, & Mariani, 2012). Currently, experiment denotes an activity in science, whereas in 1700s, it refers to an activity in religion. According to Scot (1997), a keyword is "a word which occurs with unusual frequency in a given text" when compared to another corpus.

WordSmith has a tool called *Keywords*, which compares the word lists of a corpus and a reference corpus and picks out the unusual words of that corpus. A reference corpus represents the general use of a language (Leech, 2002). An unusual word has a high keyness, which stands for the quality a word has of being key in its context (Scot, 1996). The keyness can be determined based on the Chi-square or the log likelihood statistical tests. In this study, we used the log likelihood tests. To calculate the keyness of a word, suppose that this word occurs *a* times in a corpus of *c* words, and occurs *b* times in the reference corpus of *d* words. First the expected normalized frequencies of this word occurs in the first corpus (E1) and that value for the reference corpus (E2) are calculated as:

$$E1 = c * (a + b) / (c + d)$$

$$E2 = d * (a + b) / (c + d)$$

Later, the log-likehood value (LL) is calculated as:

$$LL = 2 * ((a * \ln(a/E1)) + (b * \ln(b/E2)))$$

The higher the LL value is, the more significant the difference is, which means that the word has a high keyness. This keyword list generated by WordSmith is different from the keywords of a research article because the former is identified using the above statistic measure whereas the latter is self-identified by the author(s) of a paper.

Using this tool, we generated a keyword list based on the corpus of systems biology compared to a reference corpus. In this research, we considered several widely used reference corpora: Brown corpus, British National Corpus (BNC), Corpus of Contemporary American English (COCA), and American National Corpus (ANC)). The Brown corpus was compiled by researchers from the Brown University in the 1960s and is the first computer corpus (Francis & Kucera, 1979). The BNC was compiled in the 1990s and contains more than 1000 million words to represent British English (British National Corpus Consortium, 2007). ANC is a corpus compiled since the 1990s containing about 22 million words of written and spoken American English (Ide, & Macleod, 2001). COCA contains more than 500 million words to represent American English and is still growing (Davies, 2008).

We ultimately picked the BNC for several reasons. First, it is freely downloadable in the XML format, whereas COCA is not. Second, it is a more recent corpus compared to the Brown corpus. Third, it is a corpus of more than 90 million words, which has more words than ANC.

The use of a "stop list" can significantly influence the final result of keywords (Bullinaria & Levy, 2007). By implementing the stop list in WordSmith, the digital tool ignored the words in the stop list and did not count their frequencies. Our stop words included 524 common English stop words and 137 words that are specific to common

research articles. For example, we are not interested in certain words because they are not related to systems biology, such as "and" and "is," or words such as "figure" and "argued," despite the fact that their keyness can be very high. For a full list of stop words, see Appendix D. We also filter out a list of words that we determined would not be significant in systems biology.

The last step is normalizing the remaining interesting keywords. We eliminated all the verbs and adverbs, retaining only nouns and adjectives and picking only the singular form of the nouns. For example, we only picked "model" instead of "models," so that we were able to examine more types of words. We picked 300 keywords for our codebook.

3.2.2.3: Text-mining Using ConText.

We built a co-word network based on the codebook using open access software called ConText, one of a few tools currently available for co-word network generation. Others include Linguistic Networks Systems (LNS) (Mehler & Gleim, 2016) and GraphColl (Brezina, McEnery, & Wattam, 2015) for example. We picked ConText because it is a network and text analysis tool designed especially for scholars in the digital humanities and social sciences and is easy to use[22] (Diesner, 2014).

We used ConText to determine how many times two keywords in the codebook co-occur in a set distance for each sample sub-corpus. In this study, the distance was set as left 5 and right 5, which is the most commonly used window size (Bullinaria & Levy, 2007). The two keywords must appear in the same sentence. Using the same codebook,

[22] For more information about ConText, see http://context.lis.illinois.edu

we generated co-word networks for each of the 11 sub-corpora. ConText returned 11 data tables, which were then visualized in the next step.

3.2.2.4: Analyzing and Comparing the Co-Word Networks in Cytoscape.

For the data tables (containing the co-word frequencies of pairs of keywords found in the sample sub-corpora), we imported them in Cytoscape to generate network visualization. Cytoscape was initially developed to visualize mostly biochemical networks, but could be used for network visualization of other fields (Shannon et al., 2003). If two keywords, such as "computational" and "modeling" co-occur, Cytoscape would link the two words with an edge, resulting in "computational" and "modeling" as two vertices with a link. Cytoscape also enables the automatic calculations of network parameters and exported the results in spreadsheets. In this article, we focus on one network parameter—namely, SDC, a centrality measure.

Centrality is a widely used measure that that can help us identify the importance of keywords in a network (Davish & Tonta, 2016). There are three basic types of network centrality: betweenness centrality, closeness centrality, and degree centrality (Borgatti, 1995). Degree centrality measures the number of links to a node. Betweenness centrality measures the percentage of the number of shortest paths that pass through a node (Girvan & Newman, 2002). Closeness centrality is the sum of the length of their shortest paths.

Because we wanted to compare networks with different numbers of nodes, we used standardized degree centrality (SDC), which is explained below, instead of degree centrality to offset the influence of varying network size (Faust, 2006). Previous studies also compare SDC of a node across networks with different numbers of nodes to evaluate the importance of an individual node in health care literature and business literature

(Jiang el al., 2012; Cobbs, 2001). In our case, a node is a word. The comparison of the SDC of the same word over time can shed light on the importance of that word. SDC of a node *v* is calculated as the degree centrality $\deg(v)$ normalized over the number of nodes in a network minus one $(n-1)$ as show in the following equation:

$$SDC(v) = \frac{\deg(v)}{n-1}$$

One assumption of this research is that, if a word has more conceptual connections with other words in the literature, the words that co-occur with it will increase; hence, its centrality in the co-word network will show an upward trend. To assess whether a node has increased SDC or not, we needed to analyze its centralities in the networks between 2003 and 2013. If it is just for one word, we can add a trend line in Excel and observe whether the slope of the line is upward or downward. However, to do this for all 300 nodes manually is time-consuming. Therefore, we used a Python code to implement linear regression to automatically do this. One application of linear regression is trend estimation (Bianchi, Boyle, & Hollingsworth, 1999). In our study, the predictive variable is time and the dependent variable is SDC values. We computed the slope, the R-squared, and the p-value for the regression line for all 300 nodes. If the slope is positive, it means that the trend of the centrality of a node is upward and vice versa. The R-squared (ranging between 0 and 1), also called the correlation coefficient, measures the degree of linear dependence between the *x* variable and *y* variable (in our case, time and SDC values). If the R-squared value is 1, that means that the regression line fits the real data; if the R-squared value is 0, that means that there is no dependence between the *x* and the *y*

variable. We do not claim a causal relationship between time and SDC values; what we are interested in is the trend.

3.3. Results

3.3.1. Keyword list of systems biology research articles.

Table 8 lists the top 20 keywords with the highest keyness. The frequency indicates how many times a word appears in our corpus. The percentage stands for the percentage of articles that contain the word. For example, the word "model" appears in our abstracts 5362 times, and appears in 31% of texts. The software picked some keywords that are keywords for almost all branches of biology like "biology," "gene," "cell," and "protein," but also picked out keywords that are more unique to systems biology like "data," "network," "pathway," and "computational."

Table 8: Keyword list generated by WordSmith.

Rank	Keyword	Frequency	Percentage	Keyness
1	Biology	8335	0.49	61205.57
2	Systems	10445	0.61	49064.73
3	Gene	5372	0.31	34567.33
4	Data	7977	0.47	33467.55
5	Biological	4867	0.28	31459.65
6	Protein	5095	0.30	31102.29
7	Cell	5272	0.31	28220.92
8	Metabolic	3752	0.22	28147.20
9	Network	5246	0.31	26004.05
10	Molecular	3775	0.22	24884.41
11	Model	5362	0.31	21781.84
12	Expression	4041	0.24	18439.70
13	Signaling	2082	0.12	16897.99
14	Cellular	2486	0.15	16835.05
15	Genome	2048	0.12	15319.07
16	Pathway	2152	0.13	15100.70
17	Computational	1908	0.11	13685.21
18	Regulatory	2177	0.13	13360.91
19	Modeling	1633	0.10	13167.17
20	Experimental	2131	0.12	11239.31

The final list of keywords after filtering out unwanted words that are not related to systems biology contains 300 words, including 203 nouns and 97 adjectives. We categorized the words as biology-oriented words and systems-oriented words as well as words that are neutral based on our extensive reading on systems biology literature. According to our categorization, 180 words were biology-oriented words while 47 words were systems-oriented words, and 73 words were neutral. For a full list of the 300 keywords with their categorization, frequency, and keyness, see Appendix E.

3.3.2. Visualizing the co-word networks and computing SDC values

Figure 18 shows the co-word network for the year 2013. This network has 262 nodes and 2973 edges. The size and the label of the node are scaled according to the SDC so that one can determine which nodes are more central in the network by simply looking at it.

The left figure offers a whole view of the network but one could not see the label of the node. Therefore, we zoomed into on the center part of the network as shown in the figure on the right. The two biggest nodes are “systems” and “biology,” not surprisingly. The next tier of biggest nodes includes “data,” “metabolic,” “network,” “expression,” and “cell.” The third tier of nodes include “drug,” “development,” and many others. Given the space limit, we did not show the co-word networks for other years here.

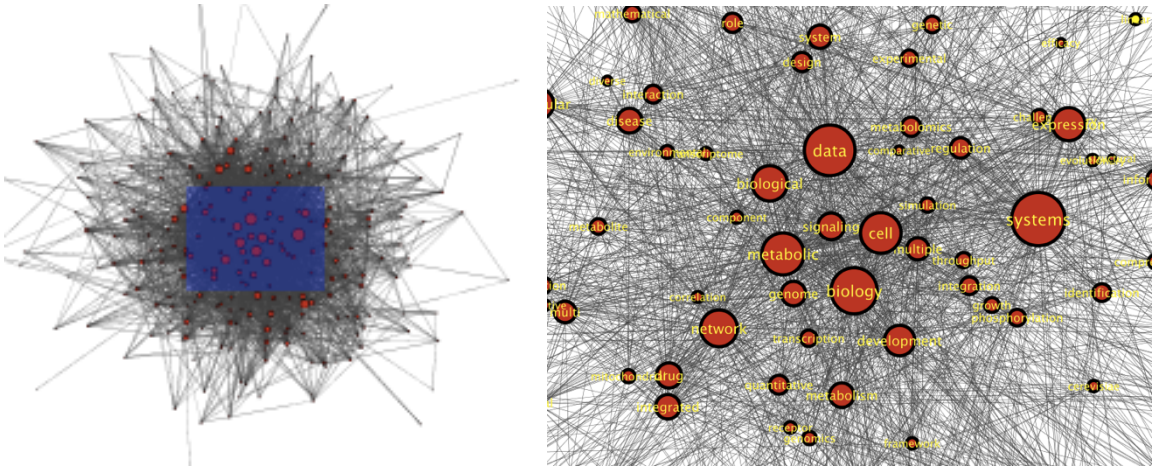


Figure 18. Co-word network in 2013. Nodes are scaled according to their SDC values. The layout is a force directed layout. The links are not scaled so they have similar width.

3.3.3. Change of SDC of 300 keywords over time

To answer the driving question proposed earlier, we wanted to know which keywords are central to the networks at different times, and whether biology-oriented or systems-oriented keywords became more central in the network. Figure 19 shows an example: how the SDCs of the word "therapy" changed from 2003 to 2013. The trend line has a positive slope of 0.0067, and the R-squared value is 0.589, which is quite high, suggesting that the SCD for the word "therapy" is indeed increasing and the word is becoming more central in the co-word network. For a full list of the SDC in all years for all 300 words, see Appendix F.

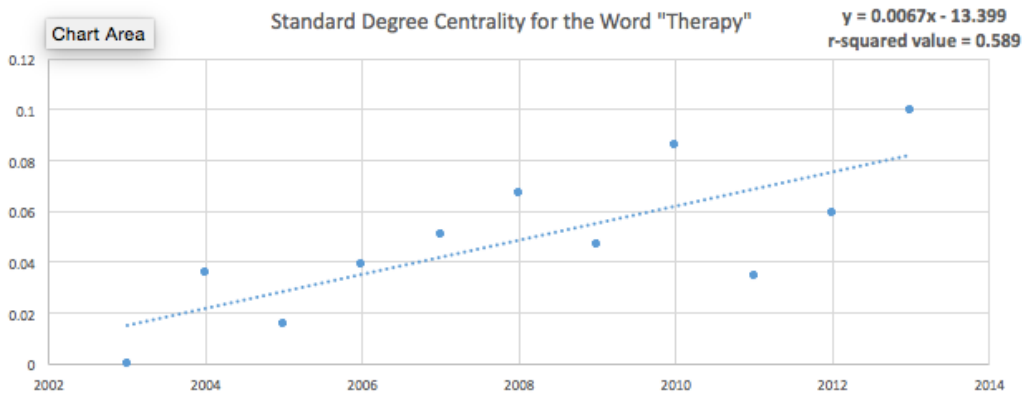


Figure 19. SDC for the word "therapy" over time. The *x* axis shows the years, and the *y* axis shows the SDC values. In 2003, the word has zero centrality, meaning that it does not exist in the network, and then afterwards, its centrality gradually increased over time.

Later we computed the slope and R-squared for all 300 nodes. We plotted the slope and R-squared for all the biology-oriented words, systems-oriented words, and neutral words separately in Figure 20, 21, and 22. For 180 biology-oriented words, 112 words have increased SDC over the years and 67 have decreased SDC over the years, and one has a slope of 0 (Figure 20). From the figure we can discover many interesting words. In the right side of the figure, words like "treatment," "synthetic," "therapy," "infection," "evolutionary," "phenotype," "clinical," "omics," and "epigenetic" have a positive slope and high R-squared value, meaning that their centrality indeed increased over the years.

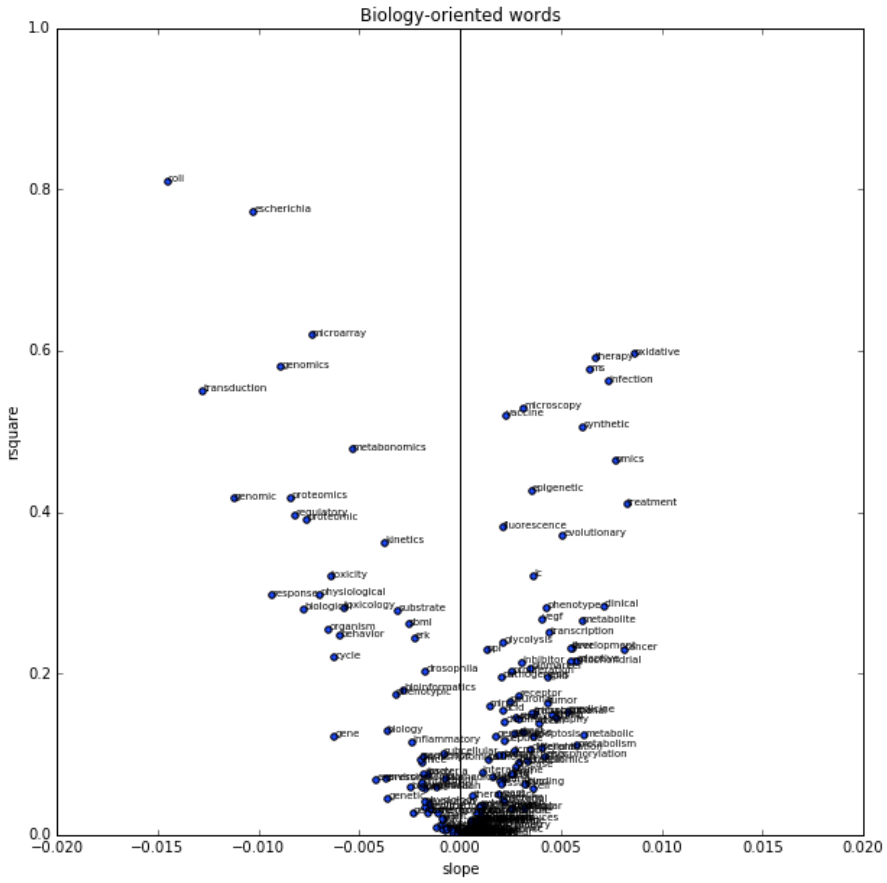


Figure 20. The slope and R-squared of biology-oriented words. The *x* axis stands for the slope of fitted linear line by which a word's SDC over the years is modeled. The *y* axis stands for the R-quared value.

For example, the SCD of the word "synthetic" has increased over the years. This result is interesting because of two reasons. First, synthetic biology depends on systems biology for understanding and reconstructing the regulatory systems (O'Malley & Soyer, 2012). Second, synthetic biology has the potential in changing human lives, for example building re-engineered organisms that can produce green energy such as alcohol, or attacking "superbugs" that are a threat to the environment (Khalil & Collins, 2010).

Many of these interesting words are related to medicine, like “therapy,” “infection,” and “clinical.”

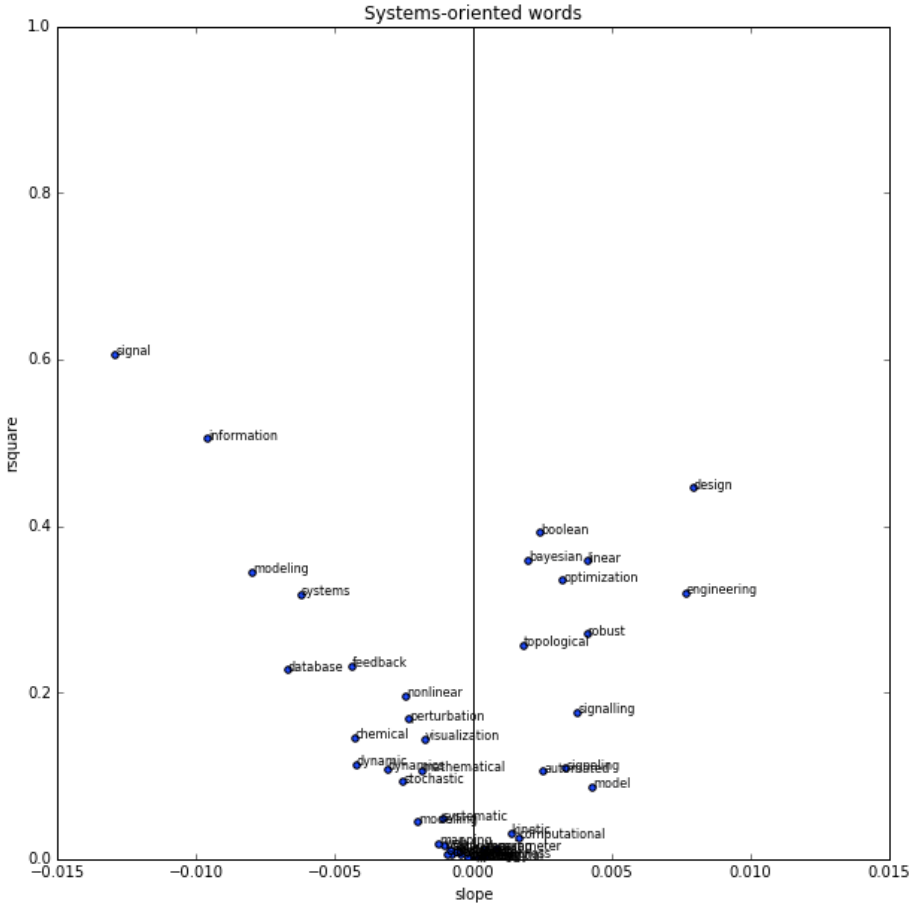


Figure 21. The slope and R-squared of systems-oriented words. The *x* axis stands for the slope of fitted linear line by which a word’s SDC over the years is modeled. The *y* axis stands for the R-squared value.

For 47 systems-oriented words, 20 words have increased SDC and 27 have decreased SDC (Figure 21). Some interesting words include “design” and “engineering.” These two words have the largest slope values, meaning that their SDC values has increased significantly.

For 73 neutral words, 37 of them have increased SDC and 36 have decreased SDC (Figure 22). The neutral words fit binomial distribution very well (p value is 0). If there is not a trend for biology-oriented words and systems-oriented words, we should observe a binomial distribution just like for neutral words.

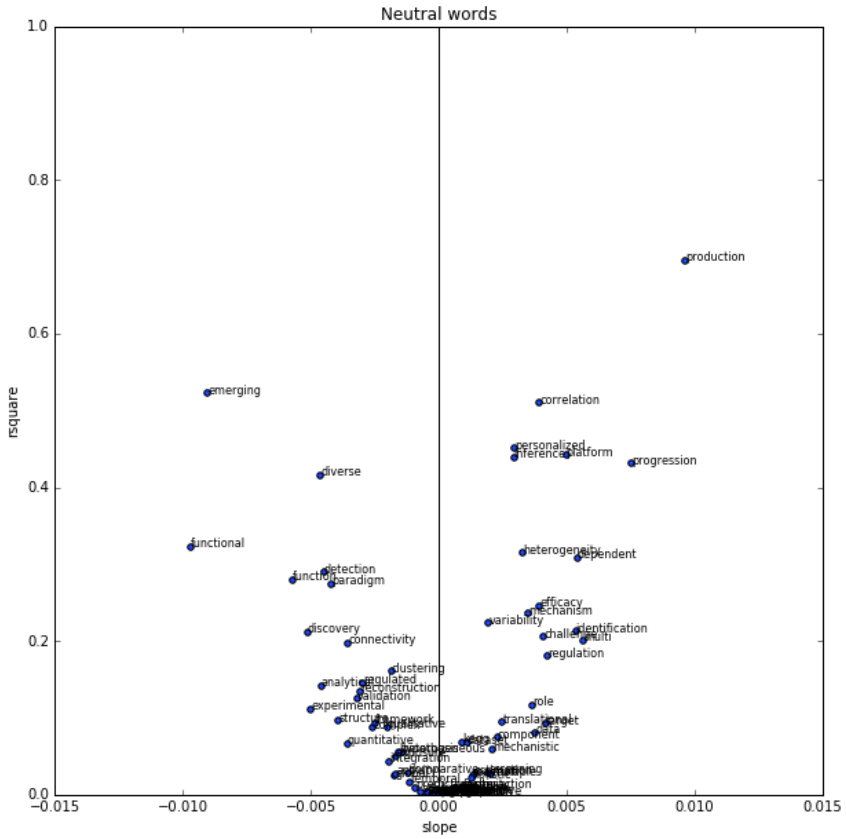


Figure 22. The slope and R-squared of neutral words. The *x* axis stands for the slope of fitted linear line by which a word's SDC over the years is modeled. The *y* axis stands for the R-squared value.

We used a binomial test for the significance of the deviation for biology-oriented words and systems-oriented words. The null hypothesis is that biology-oriented words and the systems-oriented words would be equally distributed in the left side and the right

side. The p-value for the null hypothesis that biology-oriented words does not have an increasing trend is 0.001, and for systems-oriented words that value is 0.381. Therefore, we can say statistically that biology-oriented words have become more dominant in systems biologists' discourse.

3.3.4. The finer-scale look at the sub-network

The result in the previous section offers a macroscopic view of all the 300 words, and identified some interesting words. But if we want to examine the trend of each individual word, we need a microscopic view. One advantage of co-word network analysis is its flexibility, namely, one can zoom in on a word to generate a sub-network to see one word and the connections that word has (He, 1999). Given the space of this dissertation, we could not show the co-word network of all the words that are interesting, but will pick a word to show how sub-networks can be a visualization of change in the use of a word[23].

To see how the use of each individual word has changed over time, we generated a sub-network for an important concept in systems biology--namely, cancer. The word cancer is important because firstly one of promising applications of systems biology is treating cancer (Kreeger & Lauffenburger, 2010). Systems biology aims to reveal the interactions of genes and cancer is a complex disease that has many genes involved

[23] For more co-word networks in different years and sub-networks for more keywords in different years, as well as all the Python codes and other data, please see our open-access online repository developed by the Digital Innovation Group (DigInG) at Arizona State University.

instead of one single gene, and cancer systems biology has emerged as an important branch of systems biology (Kreeger & Lauffenburger, 2010). Secondly, the co-word networks centered on cancer in different years show great change in network structure and the number of co-words. Given the space of the dissertation, we only listed the comparison of the sub-network of cancer in the years 2003 and 2013 instead of in all years as shown in Figure 23.

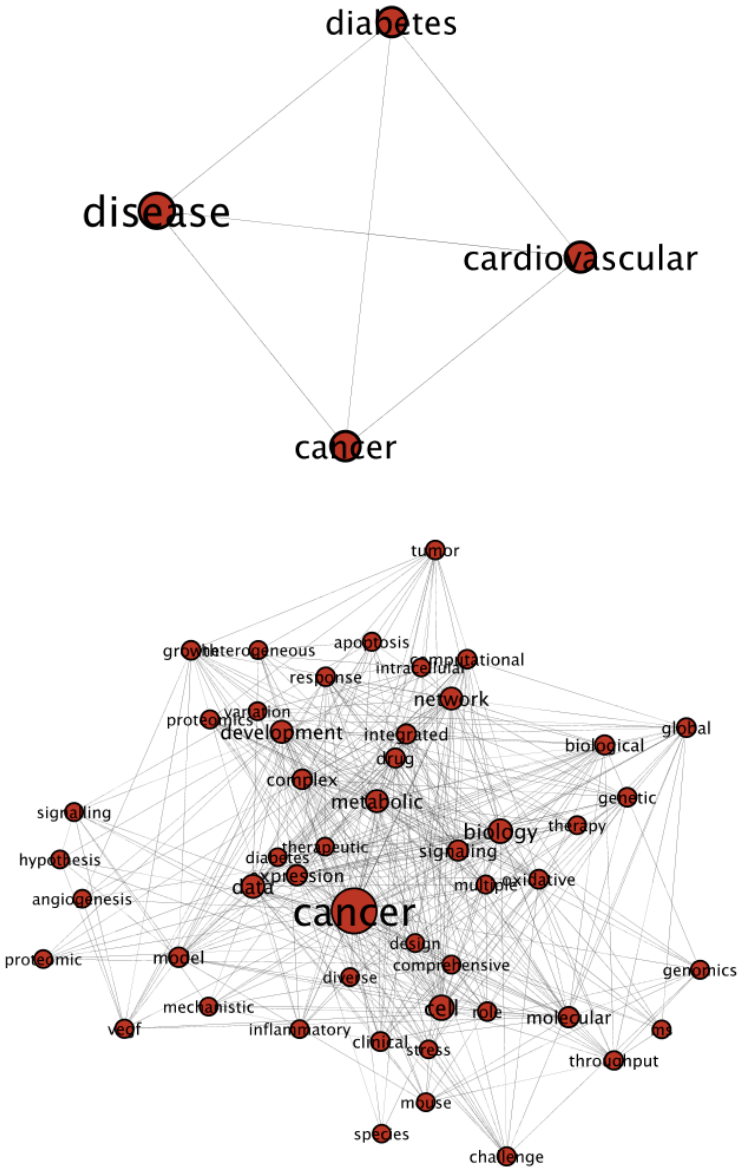


Figure 23. The co-word networks of "cancer" in 2003 and 2013. The top figure shows the network in 2003 and the bottom figure shows the network in 2013. A link indicates a co-occurrence relationship. The size of a node is proportional to its centrality in the sub-network[24].

The sub-network of cancer shows that the co-words of the word "cancer" has changed significantly over time with only a few nodes and linkages in 2003 and many more nodes in 2013. According to Courtial (1998), a word in a network cannot convey the exact meaning of that word, but the changing linkages of that word can be an indicator that the word probably has changed meanings. This might be an indication that systems biologists have deepened its understanding of cancer. In the literature, they did not talk about cancer very much in 2003, but in 2013 they did talk about cancer, and used it around many words.

To know how exactly the meaning of cancer changed over time, it is necessary to go back to original documents to find out (He, 1999). This inference can be further supported by reading the original articles. In our sample of 90 systems biology research articles from 2003, "cancer" only appears in the abstract of one article and co-occurs with 3 out of other 299 keywords as shown in Figure 23 (top). In that article, cancer was only

[24] The centrality in the sub-network for a node, which is also called local centrality, is different from the centrality of a node in the whole network, which is called global centrality. For example, words like "systems" and "biology" have higher global centrality than the word "cancer," but in the sub-network centered on cancer, cancer has the highest centrality and is bigger than the words systems and biology.

treated as a type of disease, and listed along with other diseases such as diabetes and cardiovascular disease, not having a special status in systems biology literature (Fairweather-Tait, 2003). In contrast, among the abstracts of 90 sample articles in 2013 "cancer" co-occurs with 49 out of 299 keywords as shown in Figure 23 (bottom), suggesting that systems biologists have linked cancer to so many different concepts. For example, "cancer" co-occurs with "networks" 19 times, and our result concurs with the claim of other researchers that cancer is a disease of complex networks, instead of a single-cause disease, that involves multiple genetic and environmental causes (Harrold, Ramanathan, & Mager, 2013).

3.4. Conclusions and Discussion

Systems biology is an interdisciplinary research area in which biologists and non-biologists bring in new concepts that they are familiar with, and their understanding and use of concepts change over time. The computational approach employed in this paper demonstrates that the knowledge of systems biology can be visualized and analyzed in the form of network of words, and by analyzing the networks over time we can shed light on the changing use of those words.

Our analysis shows that more than half of the biology-oriented keywords have increased SDC and more than half of systems-oriented keywords have decreased SDC. This implies that within systems biology a biology-oriented focus associated with these concepts has become more dominant. That confirms our initial hypothesis that systems biology has become a more biology-oriented science. To illustrate this trend at a more fine-grained scale we also picked an interesting concept, cancer. Here the change in the conceptual networks related to cancer illustrates (1) how cancer became a more dominant

focus of systems biology and (2) how the number of concepts linked with cancer increased dramatically, which implies an increased systems biology focus within cancer research.

This study also demonstrated the usefulness of combining corpus linguistics and network approaches for analyzing conceptual change in scientific fields. The corpus linguistic analysis enabled us to understand which keywords can actually characterize systems biology. Co-word networks over time enable visualizations of the conceptual history of systems biology on a macroscopic perspective and the conceptual history of an individual keyword on a microscopic perspective.

Quantitative results based on statistics are important not only in the natural sciences but also in the social sciences, for example when psychologists need to use SPSS to process the interview data of subjects (Stevens, 2012). In corpus linguistics as well as in network approaches, statistics is also important in generating the keyword list based on log-likelihood statistics, or determining the trends of hundreds of words without manually examining each word. So far in history and philosophy of science, quantitative reasoning is used less frequently.

Our approach thus offers historians and philosophers of science new perspectives that are based on quantitative and other technical approaches like semantic analysis and corpus linguistics. Historians and philosophers of science can apply our approach to study the historical and epistemological change in other disciplines or represent the knowledge domain that they study.

We hope that our methodology will also be interesting to humanists more broadly. Our approach used a simple coding scheme, and integrated several easy-to-use digital

tools. Humanists, like historians and philosophers, are interested in concepts and their uses. However, it is estimated that the information stored in various databases doubles every 20 months, and humanists will encounter the problem of tackling big data and need to master various data-mining techniques (Witten & Frank, 2005). We hope our methodology can be adopted by more humanists without requiring extensive computational knowledge.

CHAPTER 4: MEASURING THE CONTRIBUTIONS OF CHINESE SCHOLARS TO THE RESEARCH FIELD OF SYSTEMS BIOLOGY FROM 2005 TO 2013

Abstract: Systems biology is a new field of biology that has great implications for agriculture, medicine, and sustainability. In this chapter we explore the contributions of Chinese authors to systems biology through analyzing the metadata of more than 9000 articles on systems biology. Our big-data approach includes scientometric analysis, GIS analysis, co-word network analysis, and comparative analysis. By 2013 China is second in the number of publication on systems biology. Similar to previous studies on Chinese science, we find an unequal distribution of research power in China, favoring big cities and coastal cities. Overall, 75% of the articles in systems biology were published by scholars from universities, 15% by scholars from the Chinese of Academy of Sciences institutions, and 9% from other institutions. Many Chinese scholars' research topics are similar to those in the US, Japan, and Germany, but one salient difference is that traditional Chinese medicine is an important topic among Chinese systems biologists. 25% of Chinese systems biologists cooperate with scholars abroad, suggesting that they could take advantage of the opening-up policy. From the year 2011 to 2013, the average impact factor of the journals that Chinese scholars publish in is generally lower than that of their counterparts in the US, but the trend points to a gradual increase in impact.

Keywords: Systems biology; Chinese Scholars; Scientometrics; Network analysis; GIS analysis; Comparative analysis; Traditional Chinese Medicine.

Along with the economic liberalization of China, the scientific impact of the country is also increasing. According to a report by the Nature Publishing Group, China's expenditure on research and development in 2014 was 207 billion US dollars, second only to the US. In 2014 there are 213,000 scientific papers in Thomson Reuters' SCI database, which represents 15% of the world's total (Nature Publishing Group, 2015). However, it is often criticized that the average quality of Chinese scholars' work is not as compelling as the quantity of their work. According to the SCImago Journal & Country Rank, which uses data from the Scopus database, the overall citations of citable documents by Chinese authors from 1996 to 2014 is 19,110,353, ranking No. 6 in the world; the citations per citable document is 7.44, below the world average.[25] This means that the quality of Chinese scholars' work is generally lower than the world average. However, some scholars also point out the unequal research strength among different disciplines in China, using scientometric methods to show that China is stronger in areas related to physics, engineering, and chemistry than in other disciplines (Zhou & Leydesdorff, 2006).

In this article, we focus on the development of systems biology and explore the contribution of Chinese authors to this area. Systems biology is a burgeoning discipline of biology that involves studying biological systems at a holistic level, combining big data generated from high-through technologies and mathematical modeling. The director of the Institute of Systems Biology at Seattle, Leroy Hood, remarked that biology in the 21st century will be dominated by systems biology (Hood, 2003). Systems biology has

[25] For the rankings of more countries, see http://www.scimagojr.com/countryrank.php

great potential in health care, synthetic biology, and agriculture (Hood et al., 2004; Church 2005; Gutiérrez, 2012).

Systems biology, along with genomics, bioinformatics, computational biology, is what philosopher of biology Werner Callebaut called big data biology (BDB), which benefits greatly from genome sequencing and post-genome analysis (Callebaut, 2012). China is one of the countries that participated in the Human Genome Project (International Human Genome Sequencing Consortium, 2001). Since then, the Ministry of Science and Technology in China has been investing heavily in establishing institutions that are dedicated to post-genomics studies, such as the Beijing Genomics Institute and the Chinese National Human Genome Center, to name a few (Wu, Xiao, Zhang, & Yu, 2011). These infrastructure investments pave the way for advancement of genomics, informatics, and systems biology in China. Also, according to the *National Guidelines on the Planning of Midterm and Long Term Development of Science and Technology (2006 to 2020)*, one of the most high-profile documents that influence policy making in science, systems biology is listed as one of the research fronts in the basic research, which means that systems biology research is put on the priority list of the Chinese State Council (State Council of China, 2006).

Previously there have been studies on various disciplines of biology in China; for example, bibliographic analysis of biochemistry and molecular biology, and surveys about the plant biotechnology in China (He, Zhang, & Teng, 2005; Huang, Rozelle, Pray, & Wang, 2002). However, since systems biology is relatively new, there has not yet been any historical research on systems biology in China as far as we know. We intend to fill this gap for systems biology in China. We asked the following questions:

1. What percentage of systems biologists are from China over time?
2. Where do Chinese authors come from in terms of their geographical locations and institutional affiliations?
3. Do Chinese systems biologists share the same research topics as authors from other countries?
4. Do Chinese systems biologists work in a closed environment or an open environment where international cooperation is abundant?
5. How high is the quality of Chinese systems biologists' work in contrast with the quantity of their work?

4.1. Methods

Our research utilized a variety of computational methods to analyze the metadata of systems biology articles, including scientometric analysis, geographic information system (GIS) analysis, and network analysis. Our metadata is the bibliographic data of 9923 articles published between 1997 and 2013. Our research is also a comparative study to reveal the differences in terms of country, region, institution type, and research topic, and research quality.

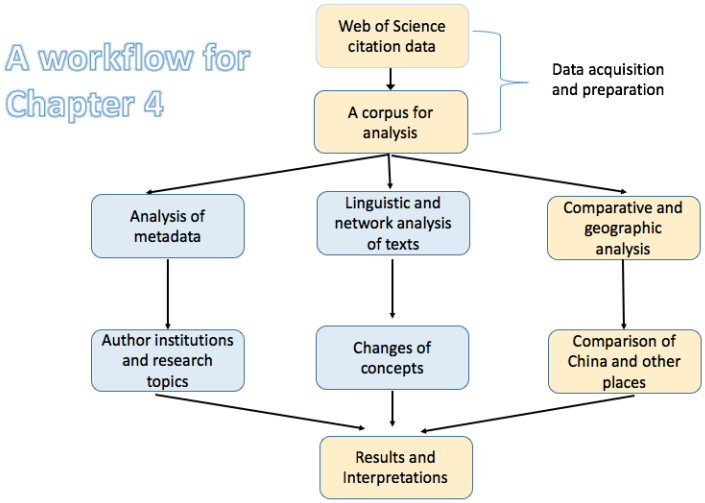


Figure 24. The computational workflow for Chapter 4.

4.1.1. Data collection

In January 2014, we searched for articles that have the term "systems biology" in the "topics," which include "titles," "abstracts," and "keywords," and published from 1900 to 2013 in the Web of Science (WoS) database. Systems biology is highly interdisciplinary, with scientists from a broad range of disciplines publishing on it: molecular biologists, evolutionary biologists, physicists, engineers, and computer scientists, to name a few. Our definition of systems biologists is in a broad sense: authors who publish the articles that fit our search criteria. The search returned 9923 articles. We then downloaded the metadata of those articles. From our data, we discovered that 2005 is the first year when Chinese authors started to publish the articles. Therefore, this study examines the contribution of Chinese scholars to systems biology from 2005 to 2013.

The WoS database is developed by the Institute of Scientific Information (ISI) of Thomson Reuters. ISI is famous for its publication of Journal Citation Report (JCR) and the analysis of journal impact factor (IF), which evaluates the influence of publications through citation counts (Russ-Eft, 2008). Its science citation index (SCI), social sciences

citation index (SSCI), and other indexes are widely used to assess the quality of journals and the articles in them, especially in China (Xue, 2008). We chose the WoS database for the following reasons: first, the WoS database is a successful commercial database, well maintained and updated, and with higher accuracy than the Google Scholar database (Falagas et al., 2008). Second, for each article that is included in the Web of Science database, it exports the title, keywords, author, publication name, publishing year, author address, and other useful metadata using different field tags. The output file itself is a big data file that can be analyzed using computational approach to get meaningful results.

4.1.2. The percentage of articles published by Chinese authors

How did we determine whether a paper was published by a Chinese author? We used the straightforward criterion that the paper must have a Chinese address as the reprint address. That precludes two situations: first, many Chinese scholars go abroad to study and do not use a Chinese address, or a Chinese author participates in the research for a paper but is not its reprint author (also known as corresponding author). In those two scenarios, we do not consider that the credit of those publications should be given to China. In the following sections, when we say Chinese scholars, we refer only to authors who have a Chinese address as a reprint author. We used Python code to get the addresses of reprint authors of the articles, and we broke down each address into institution, city, and country. We then analyzed the country of all the reprint authors and compared the percentage of articles written by reprint authors coming from different countries.

4.1.3. Geographical and institutional analysis of Chinese authors

For Chinese authors we analyzed the provinces that they come from, and the institutions with which they are affiliated. We aimed to shed light on the distribution of research power among different provinces through analysis of the number of publications. We used Google Fusion Table, a widely used GIS tool developed by Google, to geocode the addresses of Chinese authors and visualize their locations on maps. By geocode, we mean that Google Fusion Table uses its state-of-the-art cloud-computing service to transform physical locations into KML format, which enables mapping a location on maps (Google, 2016). Google Fusion Table also allowed us to visualize the numbers of papers published by authors coming from each province using the heatmap function.

We classified three types of institutions in China: universities, Chinese Academy of Sciences (CAS) institutions, and other institutions such as institutions of the Chinese Academy of Medical Sciences or hospitals. As of 2010, CAS governed 97 research institutions in over 20 provinces around the country and has top-tier researchers across all of China, many of whom are recruited from abroad (Liu & Zhi, 2010). CAS is the fourth largest funding agency in the country, second only to the National Natural Science Foundation of China, which is an equivalent of National Science Foundation of the US, the Ministry of Science and Technology, and the Ministry of Education (NPG, 2015). We examined the percentage of Chinese authors from those three types of institutions.

4.1.4. Comparing the keywords of Chinese authors and authors from other countries

Because different countries have different research strategies and traditions, we wanted to know whether Chinese authors have the same research topics, whether they lag

behind, or whether they have totally different topics than their western counterparts. Keywords, which are identified by authors, are indications of the research topics, and many previous bibliometric studies have examined the keywords of literature to understand the topics of different disciplines (Su & Lee, 2010). We retrieved the keywords as formally defined in the literature, and ranked them according to how many times they appear in the publications for each country. We then compared the keywords of Chinese authors with those of publications from the US, Japanese, and German authors.

By comparing the ranks of the top 30 keywords for the four countries, we aimed to investigate the difference in the research interests of each country. We also used a network approach to visualize the connections between keywords. If two keywords co-occur in the Keywords section of a paper, it indicates a relationship between those two keywords. We visualized the co-word network of keywords in Cytoscape (Shannon et al., 2003). The reason that we look at the co-word network because network can highlight words with high betweenness centrality, which will be explained later when discussing results.

4.1.5. Analyzing the cooperation of Chinese institutions with foreign institutions

For more than 30 years, China has adopted a policy of opening up and learning from the West after Mao Zedong's reign, in which international cooperation was not encouraged (Zhou & Glänzel, 2010). Chinese authors not only needed to overcome the difficulties of using English as a second language, but also needed to keep up with the latest trends in areas of study in the English-speaking community. The best way to learn from the West is cooperating with the West. According to a report by the British Royal

Society, over 35% of papers that were published in international journals in the year 2008 for the whole world were a result of international cooperation; that number was just 25% in 1996 (The Royal Society, 2011). Another study examines the percentage of internationally co-authored publications among all international publications in several countries between 1997 and 2007 (Zhou & Glänzel, 2010). From 1997 to 2007, the percentage of internationally co-authored papers in the US increased from 18% to 28.9%; for the UK, the number increased from 27.7% to 45.5%; China's number decreased from 24.4% to 21.9%. According to the authors of that study, it was because the denominator, the number of international publications increased. Therefore, it is interesting to see whether for systems biologists in China international cooperation increased or decreased.

Coauthoring a paper is an indication of cooperation. Although analyzing coauthor information is not a comprehensive indication of all types of cooperation happening between scholars—others include email exchanges, communicating through conferences, or inviting foreign scholars to give guest lectures—it is used as a proxy for evaluating cooperation in many previous studies and coauthor information can be easily retrieved compared with documenting other forms of cooperation (Wang, Wu, Pan, Ma, & Rousseau, 2005). We retrieved the information of coauthors with Chinese authors and identified the nationality of their coauthors. Next we identified the highest-ranking international countries in terms of the number of co-authored papers.

4.1.6. Analyzing the quality of journals of Chinese authors

This study explores the quality of journals that Chinese researchers publish in compared with their US counterparts. Evaluating the quality of research is a difficult task and we chose to use IF to do so. Eugene Garfield (1995) first proposed the use of IF. A

journal's IF for a specific year is the average number of citations of all articles published in that journal in a certain period, usually two years before that year (Garfield, 2006). Impact factors from Thomson Reuters' Journal of Citation Report (JCR) and SCImago Journal Ranking are the most widely used ones; the former is based on the WoS database, and the latter based on Scopus database.

There are some debates about using IF to assess the quality of a research study (Saha, Saint, & Christakis, 2003). For example, some argue that the IF of a journal is not representative of an individual article because when authors choose a journal for submitting their work, they do not just consider the IF of that journal alone, but also other factors (Seglen, 1997). On the other hand, some claim that citation count and IF are the most commonly used approach to measure the quality of papers (Wang, 2016).

We concede that the IF of a journal is not a predictor of the actual citations for a paper in that journal, but our research does not aim to examine the quality of one article or one author, but rather to examine many papers altogether. Therefore, we think that the average IF can be used to assess the quality of many publications for a country. Another reason is that IF are conveniently obtained compared to other factors such as the H index for all authors, which would require enormous amount of work. Actually, funding agencies in China often use IF to assess the quality of Chinese scientists' work for promotion, for instance, using the number of articles published in journals included in the JCR with a cut-off IF as a method of evaluation (Xue, 2008). This is not unique to China; the evaluation scheme was reported to be used in Italy and some Nordic countries as well (Seglen, 1997).

We obtained the IFs of more than 10,000 journals from JCR for the years 2011, 2012, and 2013. We analyzed the IF of a journal that an article was published in, which is the IF of the journal for the publication year of the article. We then compared the average IF of journals in which Chinese scholars published with that of their US counterparts. We also counted the number of articles that were published in journals with IF higher than 8 and journals with IF smaller than 8. What is considered a high-impact journal depends on the field (Leydesdorff, 2007). In some fields, for example, in medical field, impact factor of 10 might not be a high impact journal. However, in other fields, 5 could be considered a high impact factor. In our case, we picked the impact factor of 8 as a threshold because that way around 20% of article authored by the US authors are high-impact articles.

4.2. Results

The results are organized into five sub-sections, each of which corresponds to our five driving questions and methods.

4.2.1. The numbers of publications for various countries

We compared China to four countries: the US, England, Germany, and Japan. We selected these countries because they are among the top ten countries with the highest scientific impacts according to Scimago Country Rank, which includes the USA, China, Japan, Germany, South Korea, India, France, England, Russia, and Canada.

Figure 25 shows pie charts listing the total percentages of research papers that have reprint authors from the four countries as well as all other other countries grouped together in 2005 and 2013, respectively. For the exact numbers for each country, see Appendix G. Because in different years, the number of publications is different, what we

compared here is the percentage. It suggests that China contributed only a small fraction of the pie chart in 2005, but contributed a significant portion in 2013.

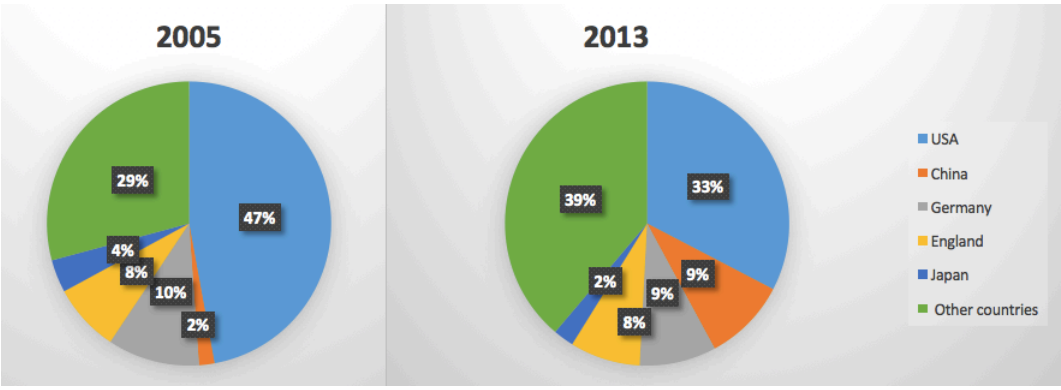


Figure 25. The comparison of the percentage of papers for each country.

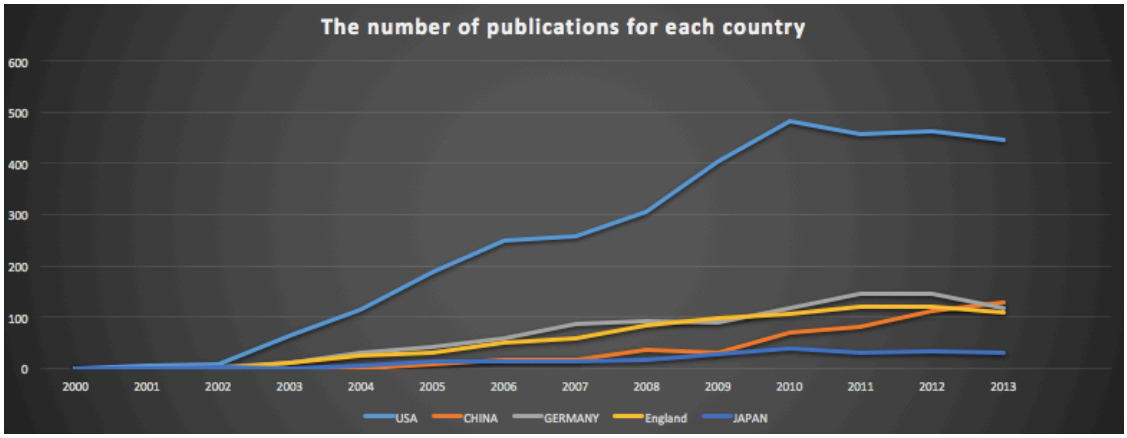


Figure 26. The number of the papers from the US, China, Germany, England and Japan from 2000 to 2013. The *x* axis stands for the year and the *y* axis stands for the number of articles published in that year for a country.

Figure 26 shows that the red line representing China has a steep slope and has exceeded that of Japan, England, and Germany in 2013. In 2005, only 1.54% of papers (7 papers) have reprint authors from China; in 2013, this number has jumped to 9.03% (129 papers), showing steady growth (The annual growth rate from 2005 to 2013 is 43.94%).

For the US, in 2005, it has 189 articles and that number for 2013 is 445, and the annual growth rate is 11.30%.

4.2.2. The geographical and institutional analysis of Chinese authors

To give an example of the distribution of research power across different regions of China, we mapped the number of papers published in 2013 onto a map of China. Figure 27 shows that Shanghai (33 papers), Beijing (23 papers), and Heilongjiang province (12 papers) have the darkest colors because highest numbers of publications were from these three places. It is not surprising that Beijing and Shanghai are two hot spots because they are the most developed regions in China, but Heilongjiang province caught our attention because it was not an economically prosperous area. We discovered that many papers were published by scholars from the Heilongjiang University of Chinese Medicine and some of those articles argue that traditional Chinese medicine is, in essence, systems medicine. For example, there is a review article in *Complementary Therapies in Medicine* arguing that traditional Chinese medicine values a holistic approach, just like systems medicine. One essence of traditional Chinese medicine is treating the body as a whole, instead of just treating a body part. That article also discusses how traditional medicine incorporated modern systems biology platforms to reform itself (Zhang, Sun, Wang, Han, & Wang, 2012).

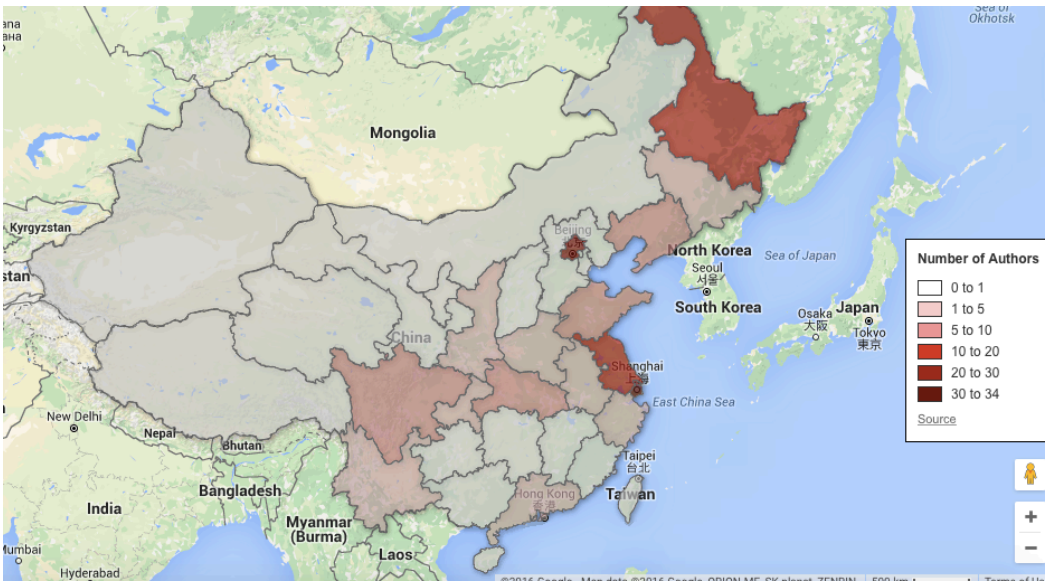


Figure 27. The numbers of papers produced by each province in 2013. As shown in the legend, the darker the red color is, the more publications a province has. Grey color means no publication at all.

We noticed the unequal distribution of publications: in many provinces, not a single paper was produced, especially cities in the northern and western regions, where the economy is not as advanced as other parts of China. It could be that systems biology is still a new discipline, so no paper was published. However, it might be more likely due to the unequal distribution of research resources. We explored the research resource distribution in China. A research paper in 2015 reports that in the life sciences sectors, resources are distributed unevenly in China, mainly favoring the eastern coastal areas and big cities like Beijing and Shanghai (Zhi & Meng, 2015). The areas with zero publications are areas that are economically less developed regions in China. We reproduced a figure from that article, as shown in Figure 28, and we found that the unequal distribution of resources matches the unequal distribution of publications on systems biology.

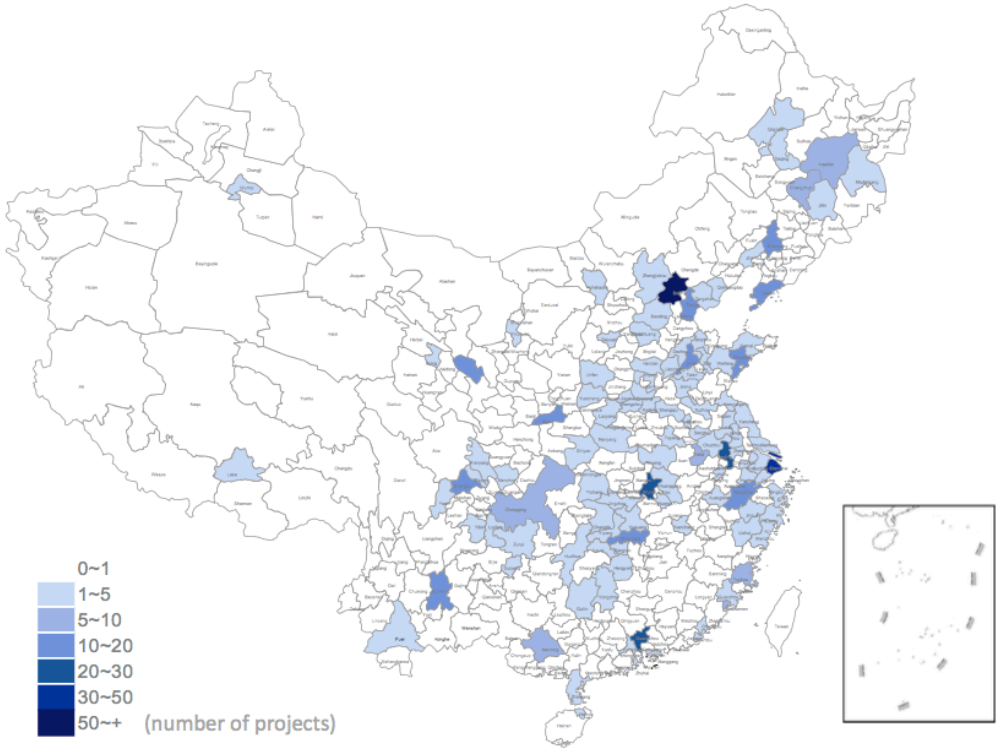


Figure 28. National Natural Science Foundation of China funding allocation in the life sciences at the city level (2006–2010). Reproduced from Zhi & Meng (2015). The legend shows the number of projects supported by the foundation. The darker the blue color is for a city, the more projects that city has.

The analysis of the types of Chinese institutions shows that from 2005 to 2013, most papers on systems biology were produced by universities (75.30%), followed by CAS institutions (15.26%), and then by other institutions (9.44%) (See Table 9).

According to the China National Bureau of Statistics, in 2013, the national spending in R&D was 1184.66 billion Chinese Yuan (roughly 189.54 billion US dollars), and CAS institutions got 43.80 billion Yuan from the government (roughly 7 billion US dollars), which accounted for 3.69 % of the government's total funding (Ministry of

Science and Technology of People's Republic of China, 2014; Chinese Academy of Science, 2014). In the field of systems biology, CAS institutions produced on average 15% of the papers published from 2005 to 2013, which exceeds the expectations if we assume that the amount of overall funding is proportional to the amount of funding for systems biology alone. One of the reasons might be the human resources reform of the CAS, which gives it an advantage in terms of recruiting researchers from abroad over other universities through its "One Hundred Talents Program" that offers more competitive salaries than universities and other institutions (Liu, & Zhi, 2010).

Table 9: The number of articles produced by three types of institutions.

	CAS Institutions	Other Research Institutions	Universities
2005	1	0	6
2006	2	1	13
2007	3	3	12
2008	9	5	21
2009	8	6	18
2010	11	0	58
2011	14	0	66
2012	19	14	79
2013	9	18	102
Total	76	46	375
Percentage	15.26%	9.44%	75.30%

Although CAS institutes have been outperforming average universities and other institutes in China, its input-output efficiency still lags behind many of its counterparts in the developed countries, for example the Max Planck Society. According to its official

website, the Max Planck Society has 83 research institutions and 17,284 employees as of January 2015, and the annual spending of the society is 1.7 billion euros[26]. The society has fewer researchers than the CAS, and less R&D spending, but produced nearly twice the numbers of publications than CAS in the period from 2005 to 2013 (153 for Max Planck Society versus 76 for CAS).

4.2.3. Keywords differences between countries

We compared the keywords ranked according to their frequencies in articles for four countries, China, the US, Japan, and Germany, from 2005 to 2013. The top 30 keywords in Table 10 shows that all four countries share many similar words with slightly different rankings: bioinformatics, proteomics, metabolomics, genomics, which are the foundational disciplines for systems biology; cancer, which is one of the most important application of systems biology to medical research; network and systems, which are key concepts of systems biology.

However, keywords like traditional Chinese medicine, herbal medicine, review, liver regeneration, rat genome 230 2.0 array, tuberculosis, chemometrics, and gc-ms are unique keywords of China or have higher rankings in Chinese publications than in publications from other countries. Keywords like modeling, drug discovery, synthetic biology, and inflammation have better rankings in US publications than in Chinese publications.

[26] For more information about the personnel and finances of the Max Planck Institute, see https://www.mpg.de/facts-and-figures

Table 10: Comparing the keywords of four countries.

US	China	Japan	Germany
systems biology	systems biology	systems biology	systems biology
proteomics	metabolomics	metabolomics	metabolomics
biology	metabonomics	database	proteomics
genomics	traditional chinese medicine [a]	bioinformatics	biology
systems	network	microarray	systems
bioinformatics	proteomics	transcriptome	bioinformatics
metabolomics	biomarkers	simulation	mathematical modeling
microarray	biology	analysis	apoptosis
mass spectrometry	bioinformatics	metabolome	transcriptomics
modeling	metabolic network	omics	cancer
computational biology	systems	systems	transcriptome
gene expression	mass spectrometry	arabidopsis thaliana	modeling
networks	genomics	feedback loop	analysis
biomarkers	networks	gastric cancer	mass spectrometry
cancer	review	synthetic biology	signal transduction
synthetic biology	omics	biology	gene expression
metabolism	cancer	cell cycle	protein
inflammation	proteome	notch	parameter estimation
genetics	nmr	toxicogenomics	mathematical model
mathematical modeling	liver regeneration	cancer	genomics
signal transduction	system biology	escherichia coli	network
protein	stability	stochasticity	metabolic networks
metabolic engineering	rat genome 230 2.0 array	computer simulation	arabidopsis thaliana
transcriptomics	time delay	metabolic engineering	metabolism
evolution	metabolites	personalized medicine	mathematical modelling
biomarker	regulatory network	reaction	gene regulation
drug discovery	gc-ms	biomarker	arabidopsis
gene regulation	tuberculosis	network	simulation
simulation	chemometrics	drug discovery	computational biology
regulation	herbal medicine	wnt	microarray

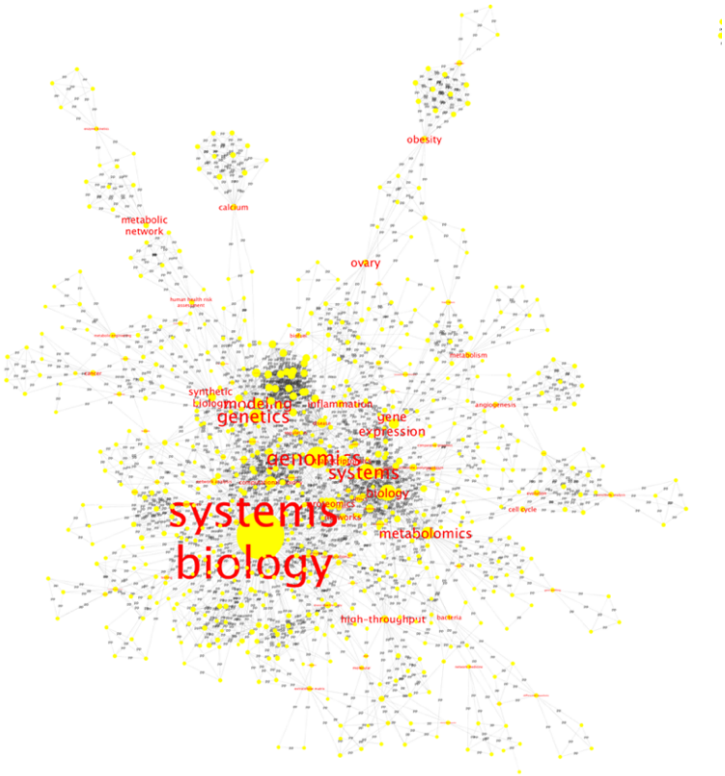
Note: [a] The words in red are words that are unique to systems biology in China.

For example, traditional Chinese medicine ranks third in Chinese publications, but is not mentioned by authors from USA, Japan and Germany at all. It is not surprising that Chinese herbal medicine does not show up in the research keywords of other countries, but it is surprising that it shows up in the keywords of systems biology literature by Chinese authors. Traditional Chinese medicine was sometimes criticized by some as pseudoscience in China, and systems biology, as a sub-branch of biology, is usually believed to be hard science (Qiu, 2007). Our research shows that the two have intersection. Traditional Chinese doctors treat patients mostly through herbal medicine, which was built on more than two thousand years' history of Chinese doctors using a trial-and-error method to test a wide range of herbals. The toxicology study of herbal medicine is modernizing through metabolomic techniques (Lao, Jiang, & Yan, 2009).

Another interesting thing we noticed is that technologies are high-ranking keywords for Chinese authors. To give two another examples, rat genome 230 2.0 array is a microarray tool for analyzing many transcripts at a time, widely used for toxicology, neurobiology, and other applications using the rat as a model organism[27]. GC-MS stands for gas chromatography-mass spectrometry, which can detect trace elements. In recent years, a majority of the funding in China has gone toward the purchase of the latest equipment, research materials, and kits from bio-companies, so Chinese authors are now equipped with the latest technologies (NPG, 2015).

[27] For more information about this tool, see http://www.affymetrix.com/catalog/131492/AFFY/Rat+Genome+230+2.0+Array#1_1

We compared the keyword co-word networks of systems biology in China and the US using the year 2013 as an example, as shown in Figure 29. Co-word network shows how words are connected instead of the rankings of words. We found that for the US, the network has more nodes than China; the network for the US has 895 keywords, and China has only 347 keywords. This can be explained by the higher number of publications by the US authors than that of Chinese authors. The US network has more clusters (51 clusters) compared with that number in China (34 clusters). A cluster consists of nodes that are closely connected with other nodes in the cluster and have loose or no connection with nodes outside of the cluster. A cluster contains keywords that usually are related to a sub-area of a discipline, so a cluster can be interpreted as a sub-area of research (He, 1999). The results suggest that research in the US is more mature and diversified.



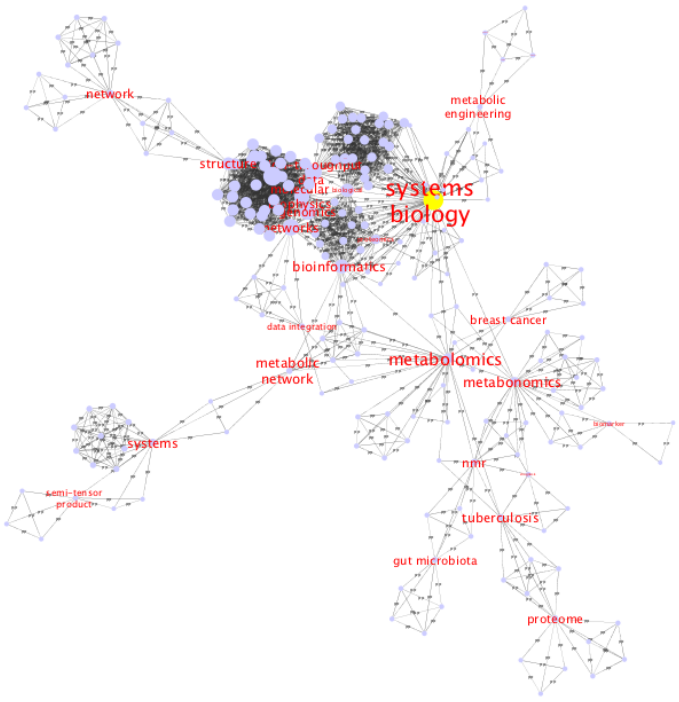


Figure 29. The keyword co-word network of the US (top figure) and China (bottom figure) in 2013. The highlighted red words are words with high betweeness centrality, meaning that they connect two clusters.

We highlighted keywords that connect different sub-areas of research by measuring their betweenness centrality, which is the number of shortest paths from all nodes to all others that pass through that node (Leydesdorff, 2007). Nodes with high betweenness centrality serves as "bridges" that connect different clusters together, and the implication is that those keywords connects different research topics or sub-area together. According to Chen (2006), nodes with high betweenness centrality can also be used to predict emerging trends in scientific literature. Figure 29 and Table 10 show two different aspects of keywords analysis, one focusing on the keywords that have higher frequency and the other focusing on keywords with higher betweeness centrality.

Comparing the nodes with high betweenness centrality, we found that some of such keywords are unique to the US or China. In the network for the US, we can see keywords like "obesity," "ovary," "inflammation," and "calcium" highlighted. In the network for China, we can see keywords like "gut microbiota," "tuberculosis," and "breast cancer" highlighted. For example, the keyword "obesity" is unique to the network of the US, and "tuberculosis" is unique to the network of China. This is interesting because obesity is a big issue in the US, and although China's rate of obesity is increasing, but it does not show up in the co-word network of China (Levine, 2011). As for tuberculosis, it is a major public health problem in China but not so much in the US because drug-resistant tuberculosis has led to increasing number of cases in China (Hu & Sun, 2013). It has been argued that a systems biology approach is better than the traditional antibiotic prescriptions in treating drug-resistant forms of TB (Young, Stark, & Kirschner, 2008).

4.2.4. The international cooperation of Chinese systems biologists with other countries

Among articles published from 2005 to 2013, on average, 25.70% of papers arose from international cooperation, and 74.30% are independent Chinese studies, as shown in Table 11. By independent, we mean that the publication has only Chinese authors, without authors from another country. The sheer number of internationally co-authored papers has increased over time, but there was not a clear trend of increase or decrease for the percentage of internationally co-authored papers.

Table 11: The number of papers produced by independent study and international cooperation

Year	Total	Independent study	International cooperation	Percentage of cooperation
2005	7	5	2	28.57%
2006	16	11	5	31.25%
2007	18	13	5	27.78%
2008	35	24	11	31.43%
2009	32	26	6	18.75%
2010	69	41	28	40.58%
2011	80	63	17	21.25%
2012	112	92	20	17.86%
2013	129	95	34	26.36%
total	498	370	128	25.70%

Chinese systems biologists have developed cooperation with authors from a total of 19 countries from 2005 to 2013. Chinese authors did not only cooperate with authors from developed countries in the North America and Europe, but also with scholars from developing countries in Asia and Africa. Table 12 shows that from 2005 to 2013, not surprisingly, the US is the biggest country where Chinese scientists' collaborators come from, followed by Japan and England. Note that Germany authors ranked second in producing the publications of systems biology until 2012, but in terms of cooperation with Chinese authors, it fell behind Japan, England, and Canada. We infer that Japan is second because it is a country that is near China geographically, and England and Canada have a language advantage over Germany because English is a universal language that many Chinese scholars speak compared with German, which makes these countries more attractive destinations for Chinese scientists. Other countries include Netherlands, Italy, Thailand, South Korea, Finland, Saudi Arabia, Ireland, South Africa, Philippines, Sweden, France, and Scotland.

Table 12: The top countries of cooperation with China

Ranking	Countries	Number of Coauthored Papers
1	USA	63
2	Japan	23
3	England	14
4	Canada	6
5	Australia	6
6	Germany	5
7	Singapore	5

4.2.5. The quality of journals in which Chinese authors published

Table 13 shows that the average IFs of the journals that Chinese authors publishing in is lower compared with US authors in three consecutive years, and not stable. The IFs for articles by US authors over the three years are relatively stable. It should be noted that the number of publications from China increased over time while the US did not.

Table 13: The comparison of IFs of journals between China and the US.

China				
	Average IF	Total	Journal with IF>=8	Ratio of high IF journals
2011	3.303	80	6	0.075
2012	2.784	112	1	0.009
2013	3.342	129	6	0.047
US				
	Average IF	Total	Journal with IF>=8	Ratio of high IF journals
2011	5.935	458	97	0.212
2012	5.701	464	102	0.220
2013	5.876	445	102	0.229

We also compared the situation for high impact journals. The percentage of articles published in high-impact journals (with IF>=8) for China is quite low and not stable, whereas that percentage for the USA stayed stable. Our data shows that it is difficult for Chinese authors to publish in high-impact journals like *Cell, Nature, or Lancet.*

4.3. Conclusions and Discussion

This study is a comparative study focusing on the publications on systems biology for Chinese authors. In this section, we discuss many suggestions that are helpful for the science policy-making in China because China is transforming from a manufacturing power into an innovation power, and from a labor-based economy into a knowledge-based economy (Dahlman & Aubert, 2001; Zhou & Leydesdorff, 2006).

Our research shows that although Chinese scholars did not have publications on systems biology in our metadata until 2005, Chinese scholars have taken up about 10 percent of all articles by 2013, and that number has grown rapidly. China has become the second-largest publisher of scientific articles on systems biology after the US by 2013. If China continues to put systems biology on its priority list as laid out in the *National Guidelines* mentioned earlier, this increase in output is likely to continue.

There has been an inequality in China's research strength, and Chinese funding agencies should pay attention not just to its coastal cities and big cities, but also to other inner provinces. Given enough resources, a province usually considered not as affluent as costal cities such as Heilongjiang was able to produce good publications. Also, the type of institutions in China can influence the input-output efficiency, with CAS being more

efficient than the average. Nonetheless, CAS institutes still lag behind its counterparts in developed worlds like the Max Planck Society.

In terms of research types, this study suggests that Chinese systems biologists share many of the same interests with Western systems biologists, but they have their unique type of study, for example research related to traditional Chinese medicine. Our study suggests that traditional Chinese medicine may not be as "traditional" as we used to think because of its incorporation of latest technologies used by systems biologists. Also, our research shows that systems biology has potential for the treating of complex diseases such as drug resistant tuberculosis, and for doctors and researchers in China, it might make sense to incorporate more systems biology approach in their research.

Over the years that we examined, around 25% of papers that had a Chinese corresponding author were a result of international cooperation, still lower than the world average. The Chinese government should continue to support the collaboration of researchers, either by sending out more visiting scholars, or allocating funding to inviting more foreign scholars to come to China to work or study. Previous literature suggests that China spent a small percentage of human resource expenditures (less than 15%) compared with those of developed countries (usually 40%) (NPG, 2015). This research suggests that the Chinese institutions should increase its dedication to recruiting and retaining researchers, as shown in the case of CAS because, as mentioned earlier, CAS is famous for its recruiting scholars from abroad.

At the same time, we show that from 2011 to 2013, despite the increase of numbers of publications, Chinese authors generally published in journals that have lower average IFs than their US counterparts. However, the Chinese government should

increase the incentives for Chinese scientists to publish in high-impact journals, because articles published in those journals often receive more scrutiny throughout the publishing process, and can raise the profile of an author and the country that author belongs to.

Our research examines publications in the WoS database, and only those published in English. It does not examine the publications in journals published in Chinese, so this research is mainly about systems biology in China perceived from abroad. In the future, we would like to examine a Chinese database such as the China Scientific and Technical Papers and Citation Database to see if there is a similar pattern.

CHAPTER 5: CONCLUSIONS

In March 2016, the ending stage of writing this dissertation, a Go game caught international attention of the whole world. The game is between AlphaGo developed by Google DeepMind Lab and South Korean professional Go player Lee Sedol, ranked No.4 in the world at the time. AlphaGo won four out of five games, and amazed the whole world with its power of computation. In an article titled "What AlphaGo's Win Means for Your Job," the author, a professor of strategic management and innovation, comments that "In the direction the world is headed, everyone will need to rethink their professional existence to ensure they have a broad prospective of where they could integrate different domain knowledge in their career track in a creative way" (Yu, 2016, para. 12) AlphaGo's victory confirms my belief that historians might want to master more computational skills to answer historical questions.

In this dissertation, I examine the development of systems biology from 1992 to 2013, and explore how biology-oriented and more systems-oriented contributions shaped its history. I studied systems biology's institutional context, research topics, knowledge structure, and regional differences using various computational tools to analyze the metadata and data of 9923 publications. I demonstrated that the computational analysis of big data embedded in the metadata of scientific literature can shed light on the historical trends of a scientific field in both qualitative and quantitative ways. My research has lead to a better understanding of systems biology's scope, institutions, epistemology, methodology, and goals using vivid representations of networks, tables, and graphs.

Equally important, other historians can use the methodology developed in my dissertation to trace the history of other disciplines especially those that have their

publications curated in online databases. Today this includes most scientific and many humanities disciplines. My computational workflow includes steps for the extraction of the metadata, analyzing the units of metadata using citation analysis, corpus linguistics analysis, co-word network analysis, GIS analysis and comparative analysis, and finally interpreting the results from historical and philosophical perspectives.

In the following sections, I will discuss a summary of my results, followed by reflections on what these results mean, both for our understanding of the history of systems biology and for the future development of computational HPS. Finally, I explain the future directions of this research and its limitations.

5.1. Summary of my Research Findings

My hypothesis was that systems biology was a discipline dominated by engineers in early 1990s, but has become more focused on empirical biological problems and biologists have become more dominant in more recent years. I used bibliographic analytical methods to pick out important authors and papers. From there, I explored the research topics and institutional backgrounds of authors through both close and distant reading. The findings of this dissertation support my hypothesis. Among the most highly cited authors, in the early 1990s most of them were from systems-oriented institutions, but in more recent years those authors came from more biology-oriented. Among the most highly cited publications, one category increased most significantly, namely, systems biology's application in medicine and bioengineering. The result of topic modeling echoed this observation.

I combined computational linguistic analysis and co-word network analysis to explore the evolution of concepts in systems biology. I examined the relationship

between different concepts and visualized those in the form of networks over time. I assessed the centrality of concepts using the SDC values. I discovered that more than half of biological concepts have increased SDC, and more than half of systems-oriented concepts have decreased SDC. My research showed how we can zoom into a big conceptual network and focus on the use of a single concept by using "cancer" as an example. Systems biologists have expanded the understanding of "cancer" as demonstrated by the increasing number of connections of this concept over time.

To explore the development of systems biology in different countries, I focused on the development in China as a case study to show that the development of systems biology varies by country, region, and different types of institutions. Through comparative study, I explored the difference in topics and publication quality between countries. For example, I found that traditional Chinese medicine, which modernizes itself through a systems biology approach and this highlights the broader appeal of a systems biology approach, is a unique topic in China.

5.2. Reflections on Trends in Systems Biology

5.2.1. A new turn in biology toward complexity

Chapter 2 shows that in systems biology, there has been a shift from focusing on abstract systems to focusing more on real biological systems and empirical problems. For example, among the most highly cited references, the category related to the -omics studies, which are about real biological systems, increased significantly after the mid-1990s. Another example, discussed at the end of Chapter two, the result of topic modeling suggests that topics related to drug, cancer, and immune system have been appearing in increasing numbers of abstracts.

This suggests that systems biology is part of a new turn in biology that focuses on complexity. First, real biological systems are inherently complex. Second, the empirical research on vaccines, complex diseases, or -omics research all need to deal with the problem of complexity at multiple scales from the network structures governing biological systems to the integrated biological, social, and economical networks that make up our healthcare system (Csete & Doyle, 2002). This focus on complexity represents a new turn in biology.

5.2.2. Systems biology's application in medicine and bioengineering

In the Chapter 2, I divided the 330 most-cited references into nine categories, and found that the applications in medicine and bioengineering have been increasing over time. My research in Chapter 3 highlights the increasing centrality of many biology-oriented concepts that are related to the application of systems biology in the medicine, like "cancer," "therapy," and "clinical."

My analysis of the literature suggests that systems biology has expanded our understanding of cancer and other complex diseases, and has potentials in the process of drug discovery or clinical use (Aderem, 2005; Hood et al., 2004). This application of systems biology approach in medicine triggered the invention of a new term called "systems medicine," as opposed to "spirit medicine," "herbal medicine," "modern medicine," and "biomedicine" (Bynum & Porter, 2013; Wood, 1997). Spirit medicine was used mostly in prehistory eras and has came from pseudoscience. Except for spirit medicine, each of the above medicine types mirrors the advancements of one or several biological disciplines. Herbal medicine originated from naturalists' understanding of the natural world, mostly plants, and it is still practiced in many parts of the world as an

alternate medicine. Since the nineteenth century, modern medicine has benefited from biochemistry, virology, microbiology, etc., and is still dominant in today's medicine. The modern medicine that is influenced by biological sciences is also called biomedicine. From these different names of medicine, we can find that the emergence of new biological disciplines is often associated with new forms of medicine. Systems biology is likely to bring a paradigm change in the history of medicine.

5.2.3. The interdisciplinarity of systems biology

My dissertation also sheds light on the relationship of systems biology with other math-heavy disciplines such as physics, math, and computer science. My research shows that among the most highly cited authors, systems-oriented scientists were more dominant than the biologists in the early days. Scientists from non-biology disciplines contributed to the foundation of systems biology by bringing new methodologies such as simulation algorithms and new statistical measurements that helped to understand big data in biology. However, among all the authors, after 2000, around 80% of scientists are from a biological institution and 20% from either a systems-oriented institute, an interdisciplinary institution or a systems biology institution. This suggests that when biologists realized the power of the systems biology approach, biologists began to take over systems biology by being the majority of systems biologists.

It is interesting to observe the relationship between non-biology disciplines and biology at the beginning of a discipline as compared to the time when a discipline has fully developed. One can find a similar case in molecular biology. Francis Crick, who contributed to the beginning of molecular biology, was a physicist when he started to work on the DNA structure with James Watson. Another important figure is Rosalind

Franklin, a chemist and crystallographer who contributed to the discovery of the double-helix structure of DNA (Elkin, 2003). The discovery of the DNA as the basis of heredity marks the beginning of molecular biology. According to Keller (1990), other contributions of non-biologists to molecular biology include those of Warren Weaver, H. J. Muller, and Erwin Schrodinger, who were all physicists. The article argues that the contribution of them was not merely technical and cognitive, but also the political resources and authority of physics. Similarly, I argue that the contribution of non-biology disciplines to systems biology is not merely methodological, technical, or political, but also in terms of resources. For example, I discovered that many researchers who were from non-biology institutions in early days migrated to biology-oriented, interdisciplinary, or systems biology institutions. Non-biology institutions provided a space for them to carry out research on systems biology while systems biology has not established itself as a legitimate discipline.

One clarification that I want to emphasize is that my research does not reject the idea that integration happens in systems biology; in any interdisciplinary field of study integration happens. However, the word 'integration' itself does not indicate the direction of the flow of information. I prefer to use the word 'incorporate' to indicate the direction of information flow is mainly from non-biology side to biology's side.

5.2.4. The relationship between systems biology and systems science

Systems science, physics, chemistry, and computational science are what I called systems-oriented science, but they are different. Physics, chemistry, and computational science are all well-established disciplines; almost every university has such departments and many practitioners in those departments. Therefore, these disciplines can contribute

to systems biology by contributing their researchers and methodologies. However, systems science itself is newer, and does not have a clear discipline boundary (Flood & Carson, 2013). After decades of development, systems science has become a discipline that has encompassed many areas of research in science, engineering, and social sciences, united under the epistemology of "systems thinking." One might ask what the contribution of systems science is?

In my results about the institutional background of the most cited authors and all authors in Chapter Two, very few researchers were from an institution that is related to systems science. Systems science *per se* contributes less direct than systems-oriented sciences like physics, engineering, which has technological and methodological contributions. This suggests that systems science's contribution may be more epistemological, meaning that it contributes mostly general ideas and concepts, instead of tools and methods. The fundamental concepts of systems science include "complexity," "robustness," "negative feedback," and "positive feedback" (Flood & Carson, 2013). These concepts also appear in keywords identified in Chapter 3.

5.3. Computational History of Science

How will the history of science be told differently in the future? There is clearly a trend that history is going digital. In 2006, Daniel J. Cohen and Roy Rosenzweig published a book on digital history, surveying a wide range of websites that tell digital history. They found that Yahoo's web directory listed 32,959 history websites by 2006 (Cohen & Rosenzweig, 2006). Not just history, but other disciplines in the humanities are experiencing a digital turn, with funding poured into this field and new centers of digital humanities being established.

Georg G. Iggers wrote that there were two very different orientations of historiography in the twentieth century: one is the traditional narrative, event-oriented history, and the other is social science-oriented history; for example, the quantitative sociological approach, economic approach, or Marxist class analysis (Iggers, 2005). If Iggers were writing now, he might argue that there is a third type of orientation for history based on big data and computation.

The history of science as a discipline studies the development of science and knowledge, and the intended readers are scientists and historians of science. The way scientists think is typically different from historians, in that scientists tend to emphasize quantitative results that are statistically significant, insights from a large sample size, and repeatability of the result. Digital history of science depends on the assembly of big data, and offers a new way to represent knowledge in science and new tools to study the big data. Thus, digital history of science can blur the distinctive line between the way scientists and historians work.

That is why scientometrics, which emerged in the 1920s, is instrumental for writing the history of science (Garfield, 2009). Historians have been enjoying the convenience brought by using computers, the Internet, etc., to do research, and I predict that more and more historians will benefit from techniques of scientometrics. There are many similarities between the history of science and scientometrics; for example, one common goal is to shed light on the policy-making process. Scientometric analysis often can help policy makers, stakeholders, and funding agencies to make decisions about where their funding should go, which priority should they give to, etc (Garousi, 2015). In this dissertation I offered some suggestions for science policy in China.

With that being said, the digital history of science should not be scientometrics alone. Nor do I mean that a historian only needs to master the technical side. Of equal importance are the hypotheses and questions behind the historical research, and the interpretation of the quantitative results from historical or philosophical perspectives.

The computational workflow presented in Chapter One and explained in detail in Chapter Two, Three, and Four is an innovative, and easy-to-replicate methodology for computational history of science. Many similar studies often focus on one type of analysis, for example co-word network study alone, or use one computational tool in their analysis. However, as interdisciplinary research is increasing, I argue that multiple approaches can give a more comprehensive picture of a scientific field. Also, the computational workflow presented here is not just about selecting tools that are available online, but also includes dozens of self-developed Python codes to facilitate the extraction and analysis of results, for example code that extracts abstracts from metadata of WoS that are later analyzed by MALLET.

5.3.1. A new form of data to examine

In my study, I extracted big data from metadata of publications downloaded from online scholarly database. As early as 1955, Eugene Garfield, suggested that citation data can be used for historical research (Garfield, 1955). Garfield and his colleagues asked the question in 1964, “Can a computer write the history of science?” (Garfield et al., 1964) He argued that the citation network might significantly alter future historiography. No scientist can carry out research on their own. Science progresses with the publications of new research based on existing publications of previous research. History involves the chronology and relationship of events and people, and metadata records people as the

authors of articles, events as the publication of a paper, and the relationship of publications as citing and being cited. Also, the importance of a scholar and a paper may be inferred from the analysis of the metadata because they tend to have a higher number of citations, which can be calculated by analyzing metadata, instead of basing on the evaluation of historians, which is inevitably subjective (Börner, Chen, & Boyack, 2003).

Computational history of science is suitable to reveal the history of recent scientific developments (Hessenbruch, 2006). In the case of systems biology, most of its publications are stored in online databases. This gives me the advantage of being able to trace things from the beginning. I examined the metadata from the WoS database because of the easy access to a computer-generated large data file. The WoS database dedicated many years of research to building the metadata database, but that does not mean that to write digital history of a field, that field must have its research output stored in such a database and having abundant citation data. Even when studying poetry for example, which does not have abundant metadata, the historian can still use some computational methods. Some of the methods showcased in this dissertation can be applied to study other types of big data, like Twitter message, or poetry, using methods like topic modeling or co-word analysis. For example, Twitter has APIs that enable researchers to download Twitter messages as big data for research purpose.

5.3.2. A new way of representing the evolution of knowledge

The evolution of knowledge can be represented in different ways, such as through narrative, rhetoric, and reasoning. One way to represent the knowledge of a scientific domain is to visualize it, and this field is called "scientography" (Börner, Chen, &

Boyack, 2003). In my research, I visualized knowledge with all sorts of networks and maps, etc.

I used network analysis extensively in my research. For example, I used co-citation network analysis in Chapter 2 and co-word network analysis in Chapter 3. Networks are powerful visualizations in both systems biology and my research, because in essence biological systems, as well as knowledge or scientific discourse are all complex adaptive systems. A network not only shows the vertices, whether it is a person, an article, or a concept, but also the relationships between them. Networks offers a perspective from a bird's eye view and show the whole picture of a knowledge domain. We can also zoom into networks to focus on a sub-network. A network can also be analyzed using clustering techniques, which identifies a closely connected group of vertices using algorithms (Börner, et al., 2003). A cluster could be interpreted as a research sub-field or a research group depending on the type of networks.

I used the geographical map in Chapter 4 to represent where systems biologists in China came from. Without GIS tools such as Google Fusion Table, it would take many more hours to finish it. Google Fusion Table enabled me to visualize the unequal research power of China, to question the funding imbalance, and to shed light on the research policy of China. It is a new way of story telling.

5.3.3. New tools to analyze the history

This project is also an exploration of the new computational tools that can be used by historians of science to look at the development of a scientific discipline. Many digital humanities projects utilize not just a single tool, but combine many tools to suit the purposes of the humanists (Gardiner & Musto, 2015).

When the first historical research based on citation data was done in the early 1960s, the citation data of DNA research was analyzed manually, which made the research time-consuming and difficult (Garfield et al., 1964). Computer scientists have developed many software packages by cooperating with humanists over the years. Take the tools used in my dissertation, for example. Some are easy to use, like Citespace, ConText, Wordsmith, and Google Fusion Table, some require the use of command line like MALLET, and some require one to master programming skills like Tethne. These are tools that work great with texts. Another important skill that facilitated my research is Python programming; for example, it was used to retrieve information from the metadata like abstracts or author information to be used by another tool like MALLET or Google Fusion Table, or to be analyzed statistically.

These tools involve computation, mathematical algorithms, and statistics, which allow us to implement visualizing, text-mining, and machine learning. These would be difficult if humanities scholars have to learn them from scratch, but are much easier when they were implemented as computational tools. For example, in Chapter 2, I used topic modeling to study the topics in the abstracts, which is essentially a machine learning approach that teaches computers to study the topics instead of having humans do it. The automatic labeling of clusters relies on LLR algorithm, and the determination of the keywords in Chapter 3 also use log likelihood test to calculate keyness.

These tools offer an opportunity and a challenge to historians. The opportunity is that these tools, given the availability of metadata, can enhance the ability of historians to examine a large research area and work in an interdisciplinary environment. The challenge lies in navigating a plethora of tools with different functions. For example, a

primer book on digital humanities lists 35 pages of websites for digital tools, which fall into 29 areas including data analysis, database management systems, data collection, data management, data visualization, etc. (Gardiner & Musto, 2015).

There are many online resources that help historians to learn these tools. For example, there is a website teaching historians to learn programming[28]. To learn how to use specific tools, the tutorial of a tool's website and videos on YouTube can also be helpful. I also attended a conference in the digital humanities and found that currently many digital humanists are doing the same things as I am such as programming, extracting information from metadata, etc. From the perspective of pedagogy, there is a need to systematically teach historians how to use computational tools and include such courses in the curriculum for historians.

5.4. Directions of Future Research and Limitations of This Research

The research reported in this dissertation points towards some promising lines of inquiry for further research. First of all, more work can be done to investigate systems medicine because my research indicates that it is a very promising direction of systems biology. That involves downloading big data specifically about systems medicine and analyze its the research topics, major contributors, key concepts, etc.

Another opportunity for future research is to expand our initial dataset to include metadata from other databases, such as metadata from Scopus, to see if there are similar patterns to observe. Although WoS has been the No. 1 database for citation analysis from

[28] To know more programming skills that historians can benefit from, see http://programminghistorian.org

the 1960s to the early 2000s, other citation databases have been developed and contain articles that WoS does not have (Meho & Yang, 2007).

A limitation of this research is that the accuracy of some current machine learning techniques can still be improved. For example, despite its fast speed to analyze millions of words, the LDA model for topic modeling is based on the assumption that a topic is a distribution of a set of words, which is a simplistic assumption and we know that a topic is far more advanced than that. That is why computer scientists are still trying to develop many other different models for topic modeling (Nallapati & Cohen, 2008).

Another limitation of this research is that there are many dynamic processes, for example a video showing the evolution of networks, that could not be shown in a paper-version of a dissertation. In many cases in my dissertation, the evolution of knowledge is shown only with a figure for the beginning year and for the ending year. The results of computational tools thus challenge the traditional way of publication on paper.

This also brings up another a new line of research: mastering the skill of building an online website that showcases the research for this dissertation and to tell an online history, as many have already done. The benefit of an online website is its wider accessibility with no fees, compared with dissertation databases such as ProQuest, which are not open to all citizens. Many scholars in the digital humanities have thus advocated for the open access movement (Suber, 2005).

To sum up, this study contributes to the scholarship on history of systems biology. My findings are illuminating for biologists, especially systems biologists, for scholars who work in an interdisciplinary field, and for historians. But more importantly, the results from this study demonstrate the power of computational tools. There are many

other questions related to systems biology that worth exploring, and the methodology laid out in this dissertation can be useful in studying those questions and questions in other disciplines. Jane Maienschein and her colleagues asked the question of how history of science can have a bigger impact on scientists, and they noted the examples of historians working with scientists in the same lab, historians reproducing the scientific findings, and a former scientist turning into a historian of science (Maienschein, Laubichler, & Loettgers, 2008). The computational approach presented in this dissertation can add to historian's repertoire to make their findings directly appealing for scientists because scientists will be interested to see the "big picture" of their scientific fields through a combination of quantitative and quantitative results.

REFERENCES

Alon, U. (2006). *Introduction to systems biology: design principles of biological networks.* Roca Raton, FL: CRC press.

Aderem, A. (2005). Systems biology: Its practice and challenges. *Cell,* 121, 511-513.

Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews. Genetics*, *5*(2), 101-13.

Barabási, A. L. (2011). The network takeover. *Nature Physics*, *8*(1), 14.

Barkal, N., & Leibler, S. (1997). Robustness in simple biochemical networks. *Nature*, 387(6636), 913-917.

Berry, D. M. (2011). The computational turn: Thinking about the digital humanities. *Culture Machine*, 12(0), 1-22.

Bianchi, M., Boyle, M., & Hollingsworth, D. (1999). A comparison of methods for trend estimation. *Applied Economics Letters*, 6(2), 103-109.

Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic Computing*, 8(4), 243-257.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Sebastopol, CA: O'Reilly Media, Inc.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the Association for Computer Machinery*, 55(4), 77-84.

Boogerd, F. C., Bruggeman, F. J., Hofmeyr, J. H. S., & Westerhoff, H. V. (2007). Towards philosophical foundations of Systems Biology: Introduction. In Boogerd, F. C., Bruggeman, F. J., Hofmeyr, J. H. S., & Westerhoff, H. V (Eds.). *Systems Biology: Philosophical foundations* (pp. 3-19). Amsterdam, The Netherlands: Elsevier Science.

Borgatti, S. P. (1995). Centrality and AIDS. *Connections*, 18(1), 112-114.

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37(1), 179-255.

Braisford, S. C., Harper, P. R., Patel, B., & Pitt, M. (2009). An analysis of the academic literature on simulation and modelling in health care. *Journal of Simulation*, 3, 130-140.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139-173.

Brigandt, I. (2013). Systems biology and the integration of mechanistic explanation and mathematical explanation. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences,* 44(4), 477-492.

British National Corpus Consortium. (2007). The *British National Corpus, version 3* (BNC XML ed.). Retrieved from http://www.natcorp.ox.ac.uk/

Bynum, W. F., & Porter, R. (2013). *Companion encyclopedia of the history of medicine*. London, England: Routledge.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3), 510-526.

Callebaut, W. (2012). Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1), 69-80.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22(2), 191-235.

Calvert, J., & Fujimura, J. H. (2011). Calculating life? Duelling discourses in interdisciplinary systems biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 42(2), 155-163.

Carnap, R. (1991). Empiricism, semantics, and ontology. In Boyd, R., Gasper, P., & Trout, J. D. (Eds.) *The Philosophy of Science* (pp. 85-98). Cambridge, MA: The MIT Press.

Chein, M., & Mugnier, M. L. (2008). *Graph-based knowledge representation: computational foundations of conceptual graphs*. London, England: Springer Science & Business Media.

Chen, C. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences*, 101(suppl 1), 5303-5310.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology,* 57(3), 359-377.

Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis. *Journal of the American Society for information Science and Technology,* 61(7), 1386-1409.

Chen, C., Hu, Z., Liu, S., & Tseng, H. (2012). Emerging trends in regenerative medicine: a scientometric analysis in CiteSpace. *Expert Opinion on Biological Therapy*, 12(5), 593-608.

Chinese Academy of Science. (2013). *The Chinese Academy of Sciences Budget in 2013.* Retrieved from http://www.cas.cn/xx/yb/cj/201304/P020130419368714220231.pdf

Church, G. M. (2005). From systems biology to synthetic biology. *Molecular Systems Biology*, 1(1).

Cobbs, J. B. (2011). The dynamics of relationship marketing in international sponsorship networks. *Journal of Business & Industrial Marketing,* 26(8), 590-601.

Cohen, D. J., & Rosenzweig, R. (2006). *Digital history: a guide to gathering, preserving, and presenting the past on the web* (Vol. 28). Philadelphia, PA: University of Pennsylvania Press.

Csete, M. E., & Doyle, J. C. (2002). Reverse engineering of biological complexity. *Science*, 295, 1664-1669.

Courtial, J-P. (1998). Comments on Leydesdorff's article. *Journal of the American Society for Information Science*, 49 (1), 98.

Dahlman, C. J., & Aubert, J. E. (2001). *China and the knowledge economy: Seizing the 21st century*. Washington, DC: World Bank Publications.

Damerow, Julia. (2014). *A Quadruple-Based Text Analysis System for History and Philosophy of Science* (Doctoral dissertation). Arizona State University, Tempe.

Darvish, H., & Tonta, Y. (2016). Diffusion of nanotechnology knowledge in Turkey and its network structure. *Scientometrics*, 107(2), 569-592.

Davies, M. (2009). The 385+ million-word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159-190.

De Bellis, N. (2009). *Bibliometrics and citation analysis: from the science citation index to cybermetrics*. Lanham, MD: Scarecrow Press.

Diesner, J. (2014). ConText: Software for the integrated analysis of text data and network data. In *Social and Semantic Networks in Communication Research.*

Preconference at Conference of International Communication Association, Seattle, WA: International Communication Association.

Dunne, C., Shneiderman, B, Gove, R., Klavans, J., Dorr, B. (2012). Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology*, 63 (12), 2351-2369.

Edmonds, P. (1997, July). Choosing the word most typical in context using a lexical co-occurrence network. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 507-509). Stroudsburg, PA: Association for Computational Linguistics.

Elkin, L. O. (2003). Rosalind Franklin and the double helix. *Physics Today*, 56(3), 42-48.

Fairweather-Tait, S. J. (2003). Human nutrition and food research: opportunities and challenges in the post-genomic era. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*,358(1438), 1709-1727.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google scholar: strengths and weaknesses. *The FASEB Biology Journal*, 22(2), 338-342.

Faust K. (2006). Comparing social networks: size, density, and local structure. *Metodoloski zvezki*, 3, 185-216.

Fell, David A. (1992). Metabolic control analysis: a survey of its theoretical and experimental development. *Biochemical Journal* 286, 2: 313-330.

Ferrer, I. C. R., & Solé, R. V. (2001, November). The small world of human language. *Proceedings of the Royal Society B: Biological sciences,* 268, (1482), 2261-2265.

Flood, R. L., & Carson, E. (2013). *Dealing with complexity: an introduction to the theory and application of systems science*. London, England: Springer Science & Business Media.

Francis, W. N., & Kucera, H. (1979). *Brown corpus manual*. Providence, RI: Brown University.

Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., & Kitano, H. (2008). CellDesigner 3.5: a versatile modeling tool for biochemical networks. *Proceedings of the IEEE*, 96(8), 1254-1265.

Gardiner, E., & Musto, R. G. (2015). *The digital humanities: A primer for students and scholars.* Cambridge, England: Cambridge University Press.

Garfield, E. (1955). Citation Indexes for Science. *Science*, 122(3159), 108-111.

Garfield, E., Sher, I. H., & Torpie, R. J. (1964). *The use of citation data in writing the history of science.* Philadelphia, PA: Institute for Scientific Information Inc.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA*, 295(1), 90-93.

Garfield, E. (2009). From the science of science to Scientometrics visualizing the history of science with HistCite software. *Journal of Informetrics*, 3(3), 173-179.

Garousi, V. (2015). A bibliometric analysis of the Turkish software engineering research community. *Scientometrics*, *105*(1), 23-49.

Gillespie, D. T. (1992). A rigorous derivation of the chemical master equation. *Physics A: Statistical Mechanics and its Applications*, 188(1), 404-425.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of National Academy of Sciences of United States of America,* 99, 8271-8276.

Google, 2012. Google Drive, Retrieved from http://www.google.com/drive/start/apps.html.

Gutiérrez, R. A. (2012). Systems biology for enhanced plant nitrogen nutrition. *Science*, 336 (6089), 1673-1675.

Hall, D., Jurafsky, D., & Manning, C. D. (2008, October). Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 363-371). Stroudsburg, PA: Association for Computational Linguistics.

Harrold, J. M., Ramanathan, M., & Mager, D. E. (2013). Network-based approaches in drug discovery and early development. *Clinical Pharmacology & Therapeutics*, 94(6), 651-658.

He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends*, 48(1), 133-159.

He, T., Zhang, J., & Teng, L. (2005). Basic research in biochemistry and molecular biology in China: A bibliometric analysis. *Scientometrics*, 62(2), 249-259.

Hessenbruch, A. (2006). "The mutt historian": The perils and opportunities of doing history of science on-line. In Doel, R. & Soderqvist, T. (Eds.) *The historiography of comtemporary science, technology, and medicine: writing recent science.* Longdon, England: Routledge.

Hettel, J. M. (2013). *Harnessing the power of context: a corpus-based analysis of variation in the language of the regulated nuclear industry* (Doctoral dissertation). Retrieved from https://getd.libs.uga.edu/pdfs/hettel_jacqueline_m_201305_phd.pdf

Hlavacek, W. (2011). Two challenges of systems biology. In Stumpt, M.P., Balding, D. J., & Girolami, M. (Eds.) *Handbook of Statistical Systems Biology*. Chichester, UK: John Wiley & Sons.

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology,* 117(4), 500-544.

Hong, L., & Davison, B. D. (2010, July). Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics* (pp. 80-88). New York, NY: *Association for Computer Machinery*.

Hood, L. (1998, August). Systems biology: new opportunities arising from genomics, proteomics and beyond. *Experimental Hematology,* 26(8), 681-681.

Hood, L. (2002). A personal view of molecular technology and how it has changed biology. *Journal of Proteome Research,* 1(5), 399-409.

Hood, L. (2003). Systems biology: integrating technology, biology, and computation. *Mechanisms of Ageing and Development*, 124(1), 9-16.

Hood, L., Heath, J. R., Phelps, M. E., & Lin, B. (2004). Systems biology and new technologies enable predictive and preventative medicine. *Science*, 306 (5696), 640-643.

Hood, L., Rowen, L., Galas, D. J., & Aitchison, J. D. (2008). Systems biology at the Institute for Systems Biology. *Briefings in Functional Genomics & Proteomics*, 7(4), 239-248.

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., ... & Kummer, U. (2006). COPASI—a complex pathway simulator. *Bioinformatics*, *22*(24), 3067-3074.

Hu, T., & Sun, W. (2013). Tuberculosis in China. *Journal of Tuberculosis Research*, 1(02), 9.

Huang, J., Rozelle, S., Pray, C., & Wang, Q. (2002). Plant biotechnology in China. *Science*, 295(5555), 674-676.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., ... & Cuellar, A. A. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4), 524-531.

Ide, N., & Macleod, C. (2001). The american national corpus: A standardized resource of american english. *Proceedings of Corpus Linguistics,* 3.

Ideker, T., Galitski, T., & Hood, L. (2001). A new approach to decoding life: systems biology. *Annual Review of Genomics and Human Genetics,* 2(1), 343-372.

Iggers, G. G. (2005). *Historiography in the twentieth century: From scientific objectivity to the postmodern challenge.* Middletown, CT: Wesleyan University Press.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860-921.

Jang, H. L., Lee, Y. S., & An, J. Y. (2012). Application of social network analysis to health care sectors. *Healthcare Informatics Research*, 18(1), 44-56.

Jong, H. D. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*, 9(1), 67-103.

Joyce, A. R., & Palsson, B. Ø. (2006). The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology*, *7*(3), 198-210.

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27-30.

Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3), 437-467.

Kauffman, S. A. (1993). *The origin of orders: Self organization and selection in evolution.* Oxford, England: Oxford University Press.

Keller, E. F. (1990). Physics and the emergence of molecular biology: A history of cognitive and political synergy. *Journal of the History of Biology*, 23(3), 389-409.

Kevles D. J., Hood L. (Eds.). (1992). *The code of codes.* Cambridge, MA: Harvard University Press.

Khalil, A., & Collins, J. (2010). Synthetic biology: applications come of age. *Nature Reviews Genetics*, 11(5), 367-379.

Kirschner, M. W. (2005). The meaning of systems biology. *Cell,* 121(4), 503-504.

Kitano, H. (Eds.). (2001). *Foundations of systems biology.* Cambridge, MA: The MIT press.

Kitano, H. (2002). Systems biology: a brief overview. *Science*, 295(5560), 1662-1664.

Kitano, H. (2002). Computational systems biology. *Nature*, 420(6912), 206-210.

Knuf, C., & Nielsen, J. (2012). Aspergilli: systems biology and industrial applications. *Biotechnology Journal*, 7(9), 1147-1155.

Kreeger, P. K., & Lauffenburger, D. A. (2010). Cancer systems biology: a network modeling perspective. *Carcinogenesis*, *31*(1), 2-8.

Krohs, U., & Callebaut, W. (2007). Data without models merging with models without data. In Boogerd, F. C., Bruggeman, F. J., Hofmeyr, J. H. S., & Westerhoff, H. V (Eds.). *Systems Biology: Philosophical foundations* (pp. 181-213). Amsterdam, The Netherlands: Elsevier Science.

Lahiri, S., Choudhury, S. R., & Caragea, C. (2014). Keyword and keyphrase extraction using centrality measures on collocation networks. *ArXiv preprint* arXiv:1401.6571.

Lattuca, L. R. (2001). *Creating interdisciplinarity: Interdisciplinary research and teaching among college and university faculty*. Nashville, TN: Vanderbilt University Press.

Lao, Y. M., Jiang, J. G., & Yan, L. (2009). Application of metabonomic analytical techniques in the modernization and toxicology research of traditional Chinese medicine. *British Journal of Pharmacology*, 157(7), 1128-1141.

Laubichler, M. D., Maienschein, J., & Renn, J. (2013). Computational perspectives in the history of science: To the memory of Peter Damerow. *Isis*, 104(1), 119-130.

Laubichler, M. L., Peirson, B. R., & Damerow, J. (2013). *Don't Panic! A research system for network-based digital history of science.* Retrieved from https://hpsrepository.asu.edu/handle/10776/6268.

Leech, G. (2002). The importance of reference corpora. *Hizkuntza-corpusak*: *Oraina eta geroa* [Corpus, now and the future]. Donostia, Spain: UZEI.

Leonelli, S., & Ankeny, R. A. (2012). Re-thinking organisms: The impact of databases on model organism biology. *Studies in History and Philosophy of Biological and Biomedical Sciences,* 43(1), 29-36.

Leonidou, L. C., Katsikeas, C. S., & Coudounaris, D. N. (2010). Five decades of business research into exporting: A bibliographic analysis. *Journal of International Management*, 16(1), 78-91.

Levesque, M. P., & Benfey, P. N. (2004). Systems biology. *Current Biology*, 14(5), R179-R180.

Levine, J. A. (2011). Poverty and obesity in the US. *Diabetes*, *60*(11), 2667-2668.

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology*, 58(9), 1303-1319.

Liu, E. T. (2005). Systems biology, integrative biology, predictive biology. *Cell*, 121(4), 505-506.

Liu, X., & Zhi, T. (2010). China is catching up in science and innovation: the experience of the Chinese Academy of Sciences. *Science and Public Policy*, 37(5), 331-342.

Machado, D., Costa, R. S., Rocha, M., Ferreira, E. C., Tidor, B., & Rocha. (2011). Modeling formalisms in systems biology. *AMB Express,* 1(1), 45-60.

Maienschein, J., Laubichler, M., & Loettgers, A. (2008). How can history of science matter to scientists? *Isis,* 99, 341-349.

Martin, M. K., Pfeffer, J., & Carley, K. M. (2013). Network text analysis of conceptual overlap in interviews, newspaper articles and keywords. *Social Network Analysis and Mining,* 3(4), 1165-1177.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., ... & Kanapin, A. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research*, *37*(suppl 1), D619-D622.

McCallum AK. 2002. *MALLET: A Machine Learning for Language Toolkit.* http://mallet.cs.umass.edu.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction.* Edinburgh, England: Edinburgh UP.

Mehler, A., & Gleim, R. (2016). Linguistic Networks–An Online Platform for Deriving Collocation Networks from Natural Language Texts. In *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks* (pp. 331-342). Springer Berlin Heidelberg.

Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American society for Information Science and Technology*, 58(13), 2105-2125.

Ministry of Science and Technology of the People's Republic of China (MSTPRC). 2013 *National Science and Technology Funding Statistics Bulletin.* Accessed online at http://www.most.gov.cn/tztg/201410/t20141030_116370.htm.

Moed, H. F. (2006). *Citation analysis in research evaluation* (Vol. 9). Springer Science & Business Media.

Nallapati, R., & Cohen, W. W. (2008-March). Link-PLSA-LDA: A New Unsupervised Model for Topics and Influence of Blogs. In *Proceedings of International Conference on Web and Social Media* (pp. 84-92.). Seattle, WA: ICWSM.

Nature Publishing Group. (2015). *Turning point: Chinese science in transition*. London: England: Nature Publishing Group.

Newman, D. J. & Block, S. (2006). Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper. *Journal of the American Society for Information Science and Technology,* 57(6):753–767.

Noble, D. (1960). Cardiac action and pacemaker potentials based on the Hodgkin-Huxley equations. *Nature,* 188, 495-497.

O'Malley, M. A., & Dupré, J. (2005). Fundamental issues in systems biology. *BioEssays,* 27(12), 1270-1276.

O'Malley, M. A., Elliott, K. C., & Burian, R. M. (2010). From genetic to genomic regulation: iterativity in microRNA research. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical* Sciences, 41(4), 407-417.

O'Malley, M. A., & Soyer, O. S. (2012). The roles of integration in molecular systems biology. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43 (1), 58-68.

Ohta, T., Tateisi, Y., & Kim, J. D. (2002). The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the Second International Conference on Human Language Technology Research* (pp. 82-86). Burlington, MA: Morgan Kaufmann Publishers Inc.

Palsson, B. O. *Systems biology: Properties of reconstructed networks*. 2006. Cambrige, England: Cambridge University Press.

Peter, M., Gartner, A., Horecka, J., Ammerer, G., & Herskowitz, I. (1993). FAR1 links the signal transduction pathway to the cell cycle machinery in yeast. *Cell*, *73*(4), 747-760.

Popping, R. (2000). *Computer-assisted text analysis*. New York, NY: Sage Publishings.

Powell, A., O'Malley, M. A., Müller-Wille, S., Calvert, J., & Dupré, J. (2007). Disciplinary baptisms: a comparison of the naming stories of genetics, molecular biology, genomics, and systems biology. *History and Philosophy of the Life Sciences*, 5-32.

Powell, A., & Dupré, J. (2009). From molecules to systems: the importance of looking both ways. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 40(1), 54–64.

Pumfrey, S., Rayson, P., & Mariani, J. (2012). Experiments in 17th century English: manual versus automatic conceptual history. *Literary and Linguistic Computing*, 27(4), 395-408.

Pulendran, B., Li, S., & Nakaya, H. I. (2010). Systems vaccinology. *Immunity*, *33*(4), 516-529.

Qiu, J. (2007). China plans to modernize traditional medicine. *Nature*, 446(7136), 590-591.

Ramage, D., & Rosen, E. (2011). Stanford topic modeling toolbox. Retrieved from http://nlp.stanford.edu/software.tmt.tmt-0.4/

Ramage, M., & Shipp, K. (2009). *Systems Thinkers.* London, England: Springer Science & Business Media.

Rea, C. (2010). Getting on with corpus compilation: from theory to practice. *ESP World*, 1(27), 1-23.

Rizzo, C. R., & Pérez, M. J. M. (2015). A Key Perspective on Specialized Lexis: Keywords in Telecommunication Engineering for CLIL. *Procedia - Social and Behavioral Sciences,* 198, 386-396.

Ronda-Pupo, G. A., & Guerras-Martin, L. Á. (2012). Dynamics of the evolution of the strategy concept 1962–2008: a co-word analysis. *Strategic Management Journal,* 33(2), 162-188.

Russ-Eft, D. (2008). SSCI, ISI, JCR, JIF, IF, and journal quality. *Human Resource Development Quarterly*, *19*(3), 185-189.

Saha, S., Saint, S., & Christakis, D. (2003). Impact factor: A valid measure of journal quality? *Journal of the Medical Library Association*, 91, 42-46.

Scharnhorst, A., & Garfield, E. (2010). Tracing scientific influence. *arXiv preprint* arXiv:1010.3525.

Schilling, D. R. (2013, April 19). Knowledge doubling every 12 months, soon to be every 12 hours. *Industry Tap into News*. Retrieved from http://www.industrytap.com/knowledge-doubling-every-12 months-soon-to-be-every-12 hours/3950

Schreibman, S., Siemens, R., & Unsworth, J. (Eds.). (2008). *A companion to digital humanities.* Chichester, UK: John Wiley & Sons.

Schreibman, S., & Hanlon, A. (2010). Determining value for digital humanities tools: report on a survey of tool developers. *Digital Humanities Quarterly,* 4(2). Retrieved from http://hdl.handle.net/2262/67330.

Scott, M. (1997). PC analysis of key words –and key key words. *System* 25(2), 233 - 245.

Scott, M. (1996). *WordSmith tools*. Oxford, England: Oxford University Press.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ: British Medical Journal*, 314(7079), 498.

Sengupta, I. N. (1992). Bibliometrics, informetrics, scientometrics and librametrics: an overview. *Libri*, 42(2), 97-98.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research,* 13(11), 2498-2504.

Sowa, J. F. (1983). *Conceptual structures: information processing in mind and machine*. Reading, MA: Addison-Wesley Pub.

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, *34*(suppl 1), D535-D539.

State Council of China. (2006). *National guidelines on the planning of midterm and long term development of science and technology (2006 to 2020).* Retrieved from http://www.gov.cn/gongbao/content/2006/content_240244.htm.

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences*. London, England: Routledge.

Su, H. N., & Lee, P. C. (2010). Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in Technology Foresight. *Scientometrics*, 85(1), 65-79.

Suber, P. (2005). Promoting open access in the humanities. *Syllecta Classica,* 16(1), 231-246.

Tessuto, G. (2015). Generic structure and rhetorical moves in English-language empirical law research articles: Sites of interdisciplinary and interdiscursive cross-over. *English for Specific Purposes*, 37, 13-26.

The Royal Society. (2011). *Knowledge, networks and nations: Global scientific collaboration in the 21st century.* London, England: The Royal Society.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., ... & Qureshi-Emili, A. (2000). A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770), 623-627.

van Atteveldt, W. H. (2008). *Semantic network analysis: Techniques for extracting, representing, and querying media content* (Doctoral dissertation). Retrieved from http://hdl.handle.net/1871/15964.

van Eck, N., & Waltman, L. (2009). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538.

van Wyhe, J. (2006). The complete work of Charles Darwin online. *Notes and Records of the Royal Society*, 60(1), 87-89.

Veling, A., & Van Der Weerd, P. (1999). Conceptual grouping in word co-occurrence networks. *International Joint Conference on Artificial Intelligence,* 9, 694-701.

Voit, E. O. (2000). *Computational analysis of biochemical systems: A practical guide for biochemists and molecular biologists*. Cambridge, England: Cambridge University Press.

Von Bertalanffy, L., & Rapoport, A. (Eds.). (1963). *General systems*. Washington, DC: Society for General Systems Research.

Wang, L. (2016). The structure and comparative advantages of China's scientific research: quantitative and qualitative perspectives. *Scientometrics,* 106(1), 435-452.

Wang, L., Zhang, Q., & Qiu, M. (2012). Evolution and Trends in Decision-Making under Uncertainty on Neuroscience and Psychology: a Scientometric Analysis in CiteSpace. *International Journal of Digital Content Technology and Its Applications*, 6(19), 181.

Wang, Y., Wu, Y., Pan, Y., Ma, Z., & Rousseau, R. (2005). Scientific collaboration in China as reflected in co-authorship. *Scientometrics*, 62(2), 183-198.

Weingart, S., & Jorgensen, J. (2013). Computational analysis of the body in European fairy tales. *Literary and Linguistic Computing,* 28(3), 404-416.

Weston, A., & Hood, L. (2004). Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *Journal of Proteome Research*, 3(2), 179–196.

Wiener, Norbert (1948). *Cybernetics, or Control and Communication in the Animal and the Machine*. Cambridge, MA: MIT Press.

Wimsatt, W. C. (2007). On building reliable pictures with unreliable data: An

evolutionary and developmental coda for the new systems biology. In Boogerd, F. C., Bruggeman, F. J., Hofmeyr, J. H. S., & Westerhoff, H. V (Eds.). *Systems Biology: Philosophical foundations* (pp. 103-120). Amsterdam, The Netherlands: Elsevier Science.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann Publishers Inc.

Wood, M. (1997). *The book of herbal wisdom: Using plants as medicine*. Berkeley, CA: North Atlantic Books.

Wu, J., Xiao, J., Zhang, R., & Yu, J. (2011). DNA sequencing leads to genomics progress in China. *Science China Life Sciences*, 54(3), 290-292.

Xue, L. (2008). China: The prizes and pitfalls of progress. *Nature*, 454(7203), 398-401.

Young, D., Stark, J., & Kirschner, D. (2008). Systems biology of persistent infection: tuberculosis as a case study. *Nature Review Microbiology*, 6(7), 520-528.

Yu, E. (2016, March 21). What AlphaGo's win means for your job. *Fotune Magazine.* Retrieved from http://fortune.com/2016/03/21/google-alphago-win-artificial-intelligence/

Zhang, A., Sun, H., Wang, P., Han, Y., & Wang, X. (2012). Future perspectives of personalized medicine in traditional Chinese medicine: a systems biology approach. *Complementary Therapies in Medicine*, 20(1), 93-99.

Zhi, Q., & Meng, T. (2015). Funding allocation, inequality, and scientific research output: an empirical study based on the life science sector of Natural Science Foundation of China. *Scientometrics*, 106(2), 1-26.

Zhou, P., & Leydesdorff, L. (2006). The emergence of China as a leading nation in science. *Research Policy*, 35(1), 83-104.

Zhou, P., & Glänzel, W. (2010). In-depth analysis on China's international cooperation in science. *Scientometrics*, 82(3), 597-612.

APPENDIX A

WOS BIBLIOGRAPHIC DATA FORMAT AND THE NINE CATEGORIES OF SYSTEMS BIOLOGY RESEARCH

In this appendix, I first introduce the types of bibliographic information that WoS can export. Then, I introduce the nine categories of systems biology research, and then show the number of articles in each category over time.

A-1 WoS Bibliographic Data Format

This appendix shows the bibliographic data format that WoS can export in bulk. A field tag is a two-character code that appears in the data. The computational programs and Python code can retrieve information according to the field tag for its corresponding information.

Field Tag	Corresponding bibliometric information
FN	File Name
VR	Version Number
PT	Publication Type (J=Journal; B=Book; S=Series)
AU	Authors
AF	Author Full Name
BA	Book Authors
CA	Group Authors
GP	Book Group Authors
TI	Document Title
ED	Editors
SO	Publication Name
SE	Book Series Title
BS	Book Series Subtitle
LA	Language
DT	Document Type
CT	Conference Title
CY	Conference Date
HO	Conference Host
CL	Conference Location
SP	Conference Sponsors
DE	Author Keywords
ID	Keywords Plus
AB	Abstract
C1	Author Address
RP	Reprint Address
EM	E-mail Address
CR	Cited References

NR	Cited Reference Count
TC	Times Cited
PU	Publisher
PI	Publisher City
PA	Publisher Address
SN	ISSN
BN	ISBN
J9	29-Character Source Abbreviation
JI	ISO Source Abbreviation
PD	Publication Date
PY	Year Published
VL	Volume
IS	Issue
PN	Part Number
SU	Supplement
SI	Special Issue
BP	Beginning Page
EP	Ending Page
AR	Article Number
PG	Page Count
DI	Digital Object Identifier (DOI)
SC	Subject Category
GA	Document Delivery Number
UT	Unique Article Identifier
ER	End of Record
EF	End of File

A-2 Description of nine categories of systems biology research

1): Metabolic Flux Analysis

Metabolic Flux Analysis measures the stoichiometric data of metabolites, and it relies on modeling using non-differential equations and a few parameters (Fell, 1992). Therefore, I classify these articles as systems-oriented. If one searches the term "systems biology" in these articles, one will not find any matching, but many historians have remarked that Metabolic Flux Analysis is the early precursors of systems biology (need a citation).

In early 1990s, many articles fall into this category. However, metabolic flux analysis become part of metabolomics which is classified into the category of "omics" research, and the name Metabolic Flux Analysis is less frequently mentioned, so a decline of the slope for metabolic flux analysis is observed.

2): Development of high-throughput technologies

The technologies include but are not limited to sequencing technologies, protein chips, DNA arrays, and biological measurements using Mass Spectrometry (Hood 2003). The focus of these articles is on technologies *per se*. The preparation of technologies contributed greatly to the emergence of systems biology.

Most articles about the development of high-throughput technologies were published in early 1990s. Such study declines in the latter half of the slope, mainly because the main technologies had been developed before 2005.

3): Algorithms, equations, modeling and simulation.

Without mathematical modeling, the data produced by high-throughput technologies would be meaningless. Mostly, it is mathematicians and engineers developing algorithms, equations, and modeling and simulation to infer, or reconstruct metabolic pathways, signal transduction pathways, or gene regulatory networks (Brigandt, 2013). Articles which fall into this category are those focused solely on algorithms and modeling per se.

4): Omics research characterizing a real biological system.

Omics research relies on the data produced by high-throughput technologies and modeling, but the ultimate goal is offering a system-level characterization of a model organism (Joyce & Palsson, 2006). Metabolomics is part of omics that has its precursor

as Metabolic Flux Analysis. However, metabolomics measures the metabolites using Mass Spectrometry and other more advanced equipment; therefore, its goal is to systematically study all the metabolites and how they interact. This type of research began to emerge with the sequencing of several important genomes of model organisms; for example, the flu genome and the yeast genome were sequenced in 1995 and1996 respectively. Omics research gradually takes up a large percentage in early 2000s.

5): Database building and curation.

This category of research involves the launch of databases storing genes, pathways, proteins, etc. It also involves the standardization of data and procedures, such as the SBML (Systems Biology Markup Language) and KEGG, which were introduced earlier. Since then, more and more other databases were developed, such as MINT (Molecular Interaction Database).

6): Software development.

Software is developed to process, analyze and visualize large data and this category is straightforward. For example, Cytoscape is especially useful for mapping various biochemical networks and was released in 2002 (Shannon et al., 2002). The developers of Cytoscape include a group of computer scientists in the Institute of Systems Biology in Seattle and engineers in the Department of Bioengineering at UCSD, and biologists from Whitehead Institute for Biomedical Research. Cytoscape can be used to visualize according to different algorithms, and analyze the network, for example, giving measurements about the centrality of the nodes. Other software includes the OpenCOBRA project and COPASI (Hoops et al., 2006).

7): Theoretical and mathematical work on network properties

These properties include robustness, dynamics, stochasticity, and emergent properties of networks. These properties can be applied to every system, not just biological systems. These studies usually deploy mathematical models to study the network properties; therefore, they are mostly systems-oriented. The study of these network properties is not a recent thing. For example, in 1997, Barkai and Leiber had already studied the robustness of simple biochemical networks (Barkal & Leiber, 1997).

8): The application of systems biology in the medical field.

Systems biology is especially useful in tackling cancer, because scientists have realized that cancer has multiple causes and involves multiple players. In the last ten years, cancer systems biology has made much progress such as in building cancer genome databases and uncovering the regulatory networks underlining cancer. Another example is that in 2004, Leroy Hood proposed that systems biology will lead to the 3P medicine, predictive, preventive and personalized medicine, which is different from traditional medicine in that it will be produced with a systems understanding of the causes of disease (Weston & Hood, 2004). The application of systems biology is the most prominent feature of the advancements of the field in the last several years.

9): Biological Mechanisms

This category deals with using systems approach to understand a specific biological mechanism, for example, how FAR1 functions in the signal transduction pathway to link to the cell cycle machinery in yeast. The main focus is on a biological mechanism (Peter, Gartner, Horecka, Ammerer, & Herskowitz, 1993).

A-3 The number of 330 most highly cited references in different categories over time

Categories	1	2	3	4	5	6	7	8	9	10	11
1992-1993	1	0	0	0	1	0	9	4	6	1	8
1994-1995	1	1	2	6	1	0	10	2	4	1	2
1996-1997	0	6	1	5	2	0	4	3	4	1	4
1998-1999	1	9	5	4	1	1	4	2	1	0	2
2000-2001	3	13	3	1	0	2	6	0	0	1	1
2002-2003	1	14	2	2	1	0	8	0	0	2	0
2004-2005	2	9	4	2	3	4	5	0	0	1	0
2006-2007	5	6	0	0	4	8	5	0	0	1	1
2008-2009	8	6	0	1	0	8	5	0	0	2	0
2010-2011	9	6	0	1	1	6	2	0	1	4	0
2012-2013	7	4	0	1	0	13	5	0	0	0	0

Note: Nine systems biology research categories include: 1: Database development; 2: Omics research; 3: Network properties; 4: Development of high-throughput technologies; 5: Software development; 6: The application of systems biology; 7: Algorithms, equations, modeling and simulation; 8: Biological Mechanisms; 9: Metabolic Flux Analysis. Other two categories are 10: hard to tell; 11: a book.

APPENDIX B

THE WORDS TO LABEL FOUR CATEGORIES OF INSTITUTIONS

Appendix B includes four categories of words that are used to automatically match the addresses of authors to identify four types of institutions: biology-oriented, systems-oriented, interdisciplinary, and systems biology institutes.

Four Categories	Words
Words for biology-oriented institutions (119 words)	allergy, anim, animal, anat, anatomy, anesthesiology, anesthesiol, arteriosclerosis, bioanalyt, biometry, biosci, brain, bacteriol, conservat, cattle, cardiol, canc, cancer, cardiac, cardiovasc, cell, clin, clinical, cytology, cytol, dermatol, dermatology, developmental, diseases, dna, drug, ecol, ecology, entomol, entomology, epidemiology, evolutionary, food, genet, genetics, genom, genome, genomics, glaxosmithkline, health, heart, hlth, hosp, hospital, Hepatol, Hepatology, human, immunology, immunol, Immunotechnol, immun, infect, insect, life, liver, livestock, lung, marine, med, medical, medicine, merck, microbiol, microbial, microbiology, mol, molecular, neurosci, nih, neurology, neurol, nutrit, nutrition, nutr, oncology, oncol, oral, organ, pfizer, pediat, pediatric, pathol, pathology, padiat, pediatrics, plant, pharm, pharma, pharmacol, pharmacology, pharmaceut, physiol, physiology, plant, proteomics, psychiat, reprod, reproduce, structrual, surgery, surg, syngenta, therapeut, therapeutical, toxicol, toxicology, vaccine, vaccines, vet, virol, virus, virology, zoology, zool
Words for systems-oriented institutions (47	artificial, artificial, astrophysics, aerosp, aerospace, automat, chem, chem, chemical, chemical, chemistry, commun, communication, comp, computat, computer, cs, control, data, database , dynam, dynamics,

words)	elect, elec, ee, energy, electrical, infocomm, informat, information, mat, math, mathematics, mech, mechanical, microsoft, modeling, phys, physics, sensors, siemens, signals, simulat, sony, statistics, stat, weapons
Words for Interdisciplinary institutions (24 words)	biochem, biochemistry, biocomputing, biodynamics, biodesign, bioenergy, bioenerget, bioenergetics, bioengn, bioengineering, bioinformat, bioinformatics, biomech, biomechanics, biomodeling, biophysics, biophys, biostatistics, biostat, biotech, biotechnol, biotechnology, ebi, interdisciplinary
Systems biology institutions (3 words)	systems biology, biosystems, biosyst

APPENDIX C

THE TOPICS AND THEIR TRENDS OVER TIME

The first part of appendix shows the words in the 20 topic bins that are based on the machine learning of the topics in abstracts of 8809 articles on systems biology. The second part shows the percentage of articles containing a topic for each topic over time.

C-1: Most likely words found in the 20 topics of systems biology.

Index	Most likely words in topic (Machine Assigned)	Description of the topic (Manually assigned)
0	plant species molecular biology systems development processes plants physiological environment developmental life physiology major importance arabidopsis organisms environmental increasing	Biology
1	model models modeling computational experimental mathematical process simulation systems modelling system quantitative biological biochemical framework processes complex dynamic hypotheses	models
2	metabolites mass metabolomics ms samples spectrometry metabolite profiling metabolic high sample quantitative nmr analytical identification profiles metabolomic detection quantification	metabolic studies
3	metabolic metabolism growth conditions flux enzymes acid enzyme energy glucose yeast production rate mitochondrial strains coli pathway carbon strain	metabolic studies
4	human medicine health effects environmental individual impact toxicity risk personalized potential major current assessment exposure development disease chemical animal	disease
5	protein proteins interactions interaction molecular functional function structural complex proteome specific human functions complexes cellular proteomic molecules proteomics large	proteomics
6	systems biology design engineering metabolic scale genome production synthetic microbial process natural strategies products applications potential efficient development interest	synthetic biology
7	data information pathway tools integration database biological databases pathways	database/software

	software developed literature tool web large integrated resources facilitate open	
8	cell cells single cellular high vivo quantitative individual molecules imaging time tissue spatial low intracellular small patterns tissues surface	Cell/tissue
9	systems biology biological complex molecular level system processes complexity view components context fundamental principles living concepts general theory perspective	systems theory
10	data parameters number experiments experimental set parameter time large algorithm sets statistical sensitivity applied algorithms prediction values inference predict	algorithms
11	response responses host immune mechanisms stress specific systems infection pathways biology cellular cells bacterial virus vaccine pathogen background understood	Immune systems
12	dynamics system control state time model reaction behavior stochastic differential reserved biochemical dynamic rate conditions cycle feedback kinetic reactions	dynamics and stochasticity
13	network networks regulatory biological interactions structure cellular complex scale components modules functions properties features robustness relationships multiple functional information	network
14	disease diseases tissue liver blood patients brain mice mechanisms disorders aging role heart human chronic normal induced increased tissues	disease
15	gene genes expression genetic genome functional identified microarray analyses data identify phenotypes genomic phenotype pathways expressed wide common specific	genomics
16	recent high technologies throughput field advances techniques biology proteomics omics genomics tools development years current future technology challenges applications	technologies and tools
17	drug cancer clinical targets discovery disease target treatment potential drugs molecular therapeutic development biomarkers tumor therapy diseases patients multiple	drug
18	regulation dna transcription regulatory gene transcriptional binding rna factors sequence	regulation

	expression mrna evolution genome msb specific sequences factor sites	
19	cell signaling cells pathways pathway signal activation receptor signalling kinase growth beta transduction stem factor phosphorylation activity differentiation alpha	pathway

C-2: the percentage of articles containing each topic over time.

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Topic 0	0.118	0.082	0.092	0.045	0.063	0.049	0.069	0.069	0.078	0.068	0.072
Topic 1	0.140	0.137	0.125	0.141	0.104	0.125	0.119	0.107	0.102	0.103	0.096
Topic 2	0.075	0.073	0.111	0.107	0.129	0.114	0.092	0.072	0.094	0.090	0.080
Topic 3	0.097	0.087	0.073	0.073	0.092	0.094	0.099	0.110	0.115	0.096	0.112
Topic 4	0.151	0.078	0.095	0.068	0.054	0.080	0.066	0.080	0.073	0.096	0.078
Topic 5	0.043	0.082	0.098	0.086	0.101	0.080	0.101	0.089	0.088	0.078	0.070
Topic 6	0.075	0.078	0.060	0.064	0.053	0.058	0.067	0.068	0.069	0.072	0.074
Topic 7	0.086	0.142	0.114	0.133	0.132	0.103	0.102	0.087	0.090	0.087	0.088
Topic 8	0.043	0.050	0.060	0.103	0.061	0.071	0.066	0.067	0.075	0.072	0.076
Topic 9	0.140	0.160	0.133	0.109	0.104	0.123	0.109	0.090	0.108	0.077	0.069
Topic 10	0.086	0.073	0.079	0.111	0.101	0.100	0.107	0.100	0.088	0.107	0.104
Topic 11	0.032	0.027	0.035	0.041	0.045	0.046	0.056	0.071	0.092	0.093	0.108

Topic 12	0.118	0.132	0.119	0.116	0.107	0.099	0.104	0.106	0.103	0.091	0.088
Topic 13	0.129	0.100	0.100	0.105	0.106	0.100	0.096	0.095	0.084	0.090	0.082
Topic 14	0.086	0.059	0.046	0.068	0.094	0.113	0.093	0.121	0.121	0.132	0.145
Topic 15	0.075	0.096	0.111	0.083	0.099	0.095	0.105	0.114	0.093	0.106	0.101
Topic 16	0.161	0.169	0.173	0.182	0.145	0.155	0.142	0.111	0.116	0.118	0.116
Topic 17	0.011	0.068	0.098	0.101	0.097	0.094	0.091	0.110	0.120	0.145	0.141
Topic 18	0.054	0.050	0.060	0.083	0.092	0.063	0.093	0.093	0.083	0.087	0.072
Topic 19	0.065	0.073	0.087	0.077	0.099	0.101	0.113	0.104	0.114	0.110	0.111

APPENDIX D

THE STOPWORDS USED IN WORDSMITH

This appendix shows the stop words that WordSmith tool ignores.

524 Common English Stop words.

a, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, all, allow, allows, almost, alone, along, already, also, although, always, am, among, amongst, an, and, another, any, anybody, anyhow, anyone, anything, anyway, anyways, anywhere, apart, appear, appreciate, appropriate, are, around, as, aside, ask, asking, associated, at, available, away, awfully, b, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, believe, below, beside, besides, best, better, between, beyond, both, brief, but, by, c, came, can, cannot, cant, cause, causes, certain, certainly, changes, clearly, co, com, come, comes, concerning, consequently, consider, considering, contain, containing, contains, corresponding, could, course, currently, d, definitely, described, despite, did, different, do, does, doing, done, down, downwards, during, e, each, edu, eg, eight, either, else, elsewhere, enough, entirely, especially, et, etc, even, ever, every, everybody, everyone, everything, everywhere, ex, exactly, example, except, f, far, few, fifth, first, five, followed, following, follows, for, former, formerly, forth, four, from, further, furthermore, g, get, gets, getting, given, gives, go, goes, going, gone, got, gotten, greetings, h, had, happens, hardly, has, have, having, he, hello, help, hence, her, here, hereafter, hereby, herein, hereupon, hers, herself, hi, him, himself, his, hither, hopefully, how, howbeit, however, i, ie, if, ignored, immediate, in, inasmuch, inc, indeed, indicate, indicated, indicates, inner, insofar, instead, into, inward, is, it, its, itself, j, just, k, keep, keeps, kept, know, knows, known, l, last, lately, later, latter, latterly, least, less, lest, let, like, liked, likely, little, look, looking, looks, ltd, m, mainly, many, may, maybe, me, mean, meanwhile, merely,

might, more, moreover, most, mostly, much, must, my, myself, n, name, namely, nd, near, nearly, necessary, need, needs, neither, never, nevertheless, new, next, nine, no, nobody, non, none, noone, nor, normally, not, nothing, novel, now, nowhere, o, obviously, of, off, often, oh, ok, okay, old, on, once, one, ones, only, onto, or, other, others, otherwise, ought, our, ours, ourselves, out, outside, over, overall, own, p, particular, particularly, per, perhaps, placed, please, plus, possible, presumably, probably, provides, q, que, quite, qv, r, rather, rd, re, really, reasonably, regarding, regardless, regards, relatively, respectively, right, s, said, same, saw, say, saying, says, second, secondly, see, seeing, seem, seemed, seeming, seems, seen, self, selves, sensible, sent, serious, seriously, seven, several, shall, she, should, since, six, so, some, somebody, somehow, someone, something, sometime, sometimes, somewhat, somewhere, soon, sorry, specified, specify, specifying, still, sub, such, sup, sure, t, take, taken, tell, tends, th, than, thank, thanks, thanx, that, thats, the, their, theirs, them, themselves, then, thence, there, thereafter, thereby, therefore, therein, theres, thereupon, these, they, think, third, this, thorough, thoroughly, those, though, three, through, throughout, thru, thus, to, together, too, took, toward, towards, tried, tries, truly, try, trying, twice, two, u, un, under, unfortunately, unless, unlikely, until, unto, up, upon, us, use, used, useful, uses, using, usually, uucp, v, value, various, very, via, viz, vs, w, want, wants, was, way, we, welcome, well, went, were, what, whatever, when, whence, whenever, where, whereafter, whereas, whereby, wherein, whereupon, wherever, whether, which, while, whither, who, whoever, whole, whom, whose, why, will, willing, wish, with, within, without, wonder, would, would, x, y, yes, yet, you, your, yours, yourself, yourselves, z, zero

137 research article stop words

abstract, advances, analyses, analysis, analyze, analyzed, analyzes, analyzing, approach, approaches, argue, argued, argues, article, articles, assume, assumed, assumes, background, based, biol, characterization, characterize, characterize, compared, conclusions, conclusions, content, context, current, demonstrate, demonstrated, describe, described, describes, dev, developed, different, discuss, discussed, discusses, discussing, doi, elsevier, elucidate, elucidating, exam, explain, explore, find, finding, findings, finds, focus, found, generated, good, highlight, http, hypotheses, identified, identify, identifying, ii, implied, implies, imply, important, including, infer, insight, insights, investigated, knowledge, large, levels, method, methodologies, methodology, methods, mol, number, observe, observed, observes, obtained, online, overview, paper, present, problem, problems, process, processes, propose, proposed, proposes, provide, provided, provides, published, question, questions, recent, related, relevant, res, research, reserved, result, resulting, results, revealed, review, reviewed, rights, science, show, showed, shows, significant, small, strategies, strategy, studies, study, technique, techniques, tool, tools, underlying, understand, understanding, view, viewed, views, wiley,

APPENDIX E

A FULL LIST OF THE 300 KEYWORDS WITH THEIR KEYNESS AND CATEGORIZATION

This appendix shows the key words that are identified by WordSmith, their frequency and keyness, and orientation that are determined manually.

	Keywords	Orientation	Frequency	Keyness
1	biology	biology-oriented	8335	61205.56641
2	systems	systems-oriented	10445	49064.72656
3	gene	biology-oriented	5372	34567.33203
4	data	neutral	7977	33467.54688
5	biological	biology-oriented	4867	31459.65039
6	protein	biology-oriented	5095	31102.29297
7	cell	biology-oriented	5272	28220.91992
8	metabolic	biology-oriented	3752	28147.19727
9	network	neutral	5246	26004.04688
10	molecular	biology-oriented	3775	24884.41016
11	model	systems-oriented	5362	21781.83984
12	expression	biology-oriented	4041	18439.69922
13	signaling	systems-oriented	2082	16897.99219
14	cellular	biology-oriented	2486	16835.04688
15	genome	biology-oriented	2048	15319.06543
16	pathway	biology-oriented	2152	15100.7041
17	computational	systems-oriented	1908	13685.21094
18	regulatory	biology-oriented	2177	13360.91113
19	modeling	systems-oriented	1633	13167.17285

20	experimental	neutral	2131	11239.30859
21	genetic	biology-oriented	1966	10814.58105
22	metabolism	biology-oriented	1606	10803.46289
23	functional	neutral	1987	10745.54785
24	complex	neutral	2911	10554.54297
25	disease	biology-oriented	2715	9787.826172
26	proteomics	biology-oriented	1161	9447.509766
27	drug	biology-oriented	2109	8685.150391
28	throughput	neutral	1141	8459.279297
29	cancer	biology-oriented	1968	8429.107422
30	quantitative	neutral	1400	8308.087891
31	metabolomics	biology-oriented	1001	8145.436035
32	regulation	neutral	1691	8141.724121
33	biochemical	biology-oriented	1206	8016.379883
34	interaction	neutral	1584	7733.947266
35	genomics	biology-oriented	919	7478.134277
36	dynamics	systems-oriented	1175	7281.458496
37	genomic	biology-oriented	975	7051.517578
38	multiple	neutral	1408	6476.132813
39	transcriptional	biology-oriented	895	6475.324219
40	mathematical	systems-oriented	1165	6187.255859
41	transcription	biology-oriented	1047	6103.785645

42	microarray	biology-oriented	720	5858.73877
43	response	biology-oriented	2138	5842.679688
44	omics	biology-oriented	709	5769.226074
45	behavior	biology-oriented	740	5513.919922
46	metabolite	biology-oriented	711	5492.158691
47	dynamic	systems-oriented	1137	5457.967773
48	profiling	biology-oriented	717	5403.707031
49	human	biology-oriented	2432	5226.474121
50	yeast	biology-oriented	819	5157.376465
51	bioinformatics	biology-oriented	626	5078.968262
52	simulation	systems-oriented	822	5007.046875
53	clinical	biology-oriented	1220	4960.95752
54	function	neutral	1692	4866.922363
55	identification	neutral	1104	4850.516113
56	proteomic	biology-oriented	594	4833.416504
57	spectrometry	biology-oriented	627	4701.719238
58	tumor	biology-oriented	594	4651.02832
59	discovery	neutral	1126	4567.30957
60	proteome	biology-oriented	559	4548.60791
61	immune	biology-oriented	836	4344.881348
62	phenotype	biology-oriented	619	4254.375977
63	activation	biology-oriented	728	4192.699219

64	tissue	biology-oriented	961	4124.571289
65	stochastic	systems-oriented	551	4010.531982
66	parameter	systems-oriented	709	3993.535889
67	integrated	neutral	1008	3992.280762
68	physiological	biology-oriented	711	3967.955078
69	complexity	biology-oriented	891	3944.255859
70	ms	biology-oriented	899	3924.736816
71	development	biology-oriented	2640	3921.371094
72	silico	systems-oriented	480	3905.759766
73	therapeutic	biology-oriented	703	3836.727539
74	integration	neutral	953	3826.568604
75	intracellular	biology-oriented	576	3808.160889
76	kinetic	systems-oriented	582	3793.374512
77	algorithm	systems-oriented	666	3755.89624
78	vivo	biology-oriented	628	3727.744385
79	receptor	biology-oriented	648	3649.760742
80	transduction	biology-oriented	481	3633.080078
81	dna	biology-oriented	998	3543.198975
82	modelling	systems-oriented	732	3534.00415
83	microbial	biology-oriented	493	3524.203125
84	system	systems-oriented	2951	3466.7854
85	flux	biology-oriented	592	3331.375977

86	kinase	biology-oriented	520	3328.661377
87	global	neutral	967	3310.481689
88	signalling	systems-oriented	598	3294.043945
89	binding	biology-oriented	863	3245.120117
90	differential	systems-oriented	651	3237.42627
91	regulated	neutral	638	3212.411377
92	robustness	systems-oriented	442	3206.404297
93	multi	neutral	498	3167.998535
94	integrative	neutral	448	3162.60
95	information	systems-oriented	2597	3108.911865
96	transcriptome	biology-oriented	380	3092.038086
97	medicine	biology-oriented	851	3085.066406
98	predictive	neutral	486	3059.057373
99	cerevisiae	biology-oriented	399	3051.877686
100	prediction	neutral	601	3051.863525
101	signal	systems-oriented	873	3012.80127
102	apoptosis	biology-oriented	416	2998.35083
103	mechanistic	neutral	437	2994.864258
104	framework	neutral	933	2989.05249
105	coli	biology-oriented	492	2898.485352
106	mrna	biology-oriented	471	2891.92749
107	synthetic	biology-oriented	563	2811.379883

108	growth	biology-oriented	1427	2785.716064
109	phosphorylation	biology-oriented	431	2772.445557
110	msb	biology-oriented	343	2747.444092
111	chemical	systems-oriented	910	2674.066406
112	tcm	biology-oriented	327	2660.770264
113	vitro	biology-oriented	498	2641.252686
114	optimization	systems-oriented	359	2631.76001
115	rna	biology-oriented	563	2616.562012
116	robust	systems-oriented	523	2615.502686
117	host	biology-oriented	792	2609.386963
118	arabidopsis	biology-oriented	324	2602.342041
119	engineering	systems-oriented	929	2596.895996
120	sequencing	biology-oriented	417	2532.962158
121	target	neutral	1025	2522.986084
122	metabolomic	biology-oriented	306	2489.891357
123	biomarker	biology-oriented	306	2465.825684
124	enzyme	biology-oriented	506	2458.734619
125	comprehensive	neutral	797	2453.54248
126	mass	biology-oriented	1026	2408.724854
127	translational	neutral	320	2404.999512
128	organism	biology-oriented	501	2399.558838
129	annotation	biology-oriented	327	2392.108887

130	toxicity	biology-oriented	378	2391.470215
131	stress	biology-oriented	872	2386.202393
132	personalized	neutral	335	2376.688232
133	phenotypic	biology-oriented	339	2342.850098
134	cycle	biology-oriented	749	2341.596191
135	mitochondrial	biology-oriented	349	2338.059326
136	web	systems-oriented	455	2312.831543
137	visualization	systems-oriented	335	2296.304932
138	sbml	biology-oriented	283	2289.479004
139	database	systems-oriented	739	2281.32251
140	transcriptomics	biology-oriented	279	2270.190674
141	sensitivity	neutral	579	2261.710938
142	analytical	neutral	480	2232.541504
143	saccharomyces	biology-oriented	287	2228.697754
144	validation	neutral	441	2227.490967
145	nmr	biology-oriented	339	2224.616211
146	metabolome	biology-oriented	273	2221.368408
147	mechanism	neutral	690	2194.114014
148	kinetics	biology-oriented	312	2184.016846
149	evolution	biology-oriented	650	2181.921631
150	biosynthesis	biology-oriented	288	2150.449463
151	physiology	biology-oriented	380	2122.677979

152	glucose	biology-oriented	429	2111.425781
153	homeostasis	biology-oriented	280	2111.231689
154	biomedical	biology-oriented	306	2089.451904
155	escherichia	biology-oriented	287	2073.372559
156	plant	biology-oriented	988	2065.236816
157	structural	neutral	640	2031.841309
158	aging	biology-oriented	299	2003.765259
159	dimensional	neutral	330	2003.117432
160	lipid	biology-oriented	348	1992.485962
161	species	biology-oriented	1035	1970.590332
162	statistical	systems-oriented	575	1969.111816
163	reaction	neutral	811	1964.19397
164	feedback	systems-oriented	494	1961.496582
165	inference	neutral	386	1947.89978
166	pathogen	biology-oriented	270	1936.298218
167	transcriptomic	biology-oriented	237	1928.435913
168	evolutionary	biology-oriented	463	1923.471436
169	progression	neutral	416	1904.287354
170	factor	neutral	840	1893.415894
171	emerging	neutral	511	1889.112793
172	dynamical	systems-oriented	284	1887.355103
173	mammalian	biology-oriented	337	1883.458252

174	ontology	systems-oriented	265	1875.108765
175	inflammatory	biology-oriented	403	1874.7323
176	application	neutral	1019	1853.779907
177	temporal	neutral	378	1841.334106
178	design	systems-oriented	1147	1837.918335
179	differentiation	biology-oriented	426	1837.61792
180	role	neutral	1354	1825.267944
181	nonlinear	systems-oriented	270	1820.528687
182	bacterial	biology-oriented	375	1813.317993
183	oxidative	biology-oriented	251	1799.133179
184	environmental	biology-oriented	927	1791.842407
185	bayesian	systems-oriented	223	1759.180176
186	liver	biology-oriented	491	1753.308594
187	dependent	neutral	655	1748.217407
188	developmental	biology-oriented	372	1724.11853
189	mirna	biology-oriented	210	1686.171631
190	imaging	systems-oriented	331	1654.041016
191	diverse	neutral	439	1645.897705
192	circadian	biology-oriented	224	1614.773438
193	systematic	systems-oriented	468	1613.158081
194	ppi	biology-oriented	203	1612.423096
195	stem	biology-oriented	426	1605.444824

196	epigenetic	biology-oriented	215	1599.022095
197	interactome	biology-oriented	195	1586.683105
198	chromatography	biology-oriented	271	1585.139038
199	inflammation	biology-oriented	322	1580.475098
200	synthesis	neutral	415	1578.711792
201	amino	biology-oriented	344	1572.781128
202	detection	neutral	378	1567.493286
203	systemic	systems-oriented	293	1563.251343
204	acid	biology-oriented	676	1529.149048
205	quantification	neutral	263	1523.612793
206	pathogenesis	biology-oriented	276	1503.769897
207	topology	systems-oriented	223	1503.326172
208	variability	neutral	302	1503.035034
209	atp	biology-oriented	256	1492.965088
210	optimal	systems-oriented	347	1486.030151
211	therapy	biology-oriented	459	1476.247803
212	genetics	biology-oriented	279	1476.083862
213	membrane	biology-oriented	363	1474.13623
214	vaccine	biology-oriented	291	1469.239502
215	extracellular	biology-oriented	239	1458.199341
216	infection	biology-oriented	515	1454.011475
217	clustering	neutral	258	1445.763794

218	thaliana	biology-oriented	178	1426.447144
219	platform	neutral	492	1420.926758
220	perturbation	systems-oriented	204	1416.167725
221	plasma	biology-oriented	360	1407.12146
222	linear	systems-oriented	395	1402.744629
223	mice	biology-oriented	360	1380.913208
224	sequence	biology-oriented	592	1375.648315
225	multivariate	neutral	213	1367.607666
226	dataset	neutral	199	1364.912354
227	correlation	neutral	387	1339.389038
228	egfr	biology-oriented	162	1318.164429
229	screening	neutral	371	1315.606445
230	estimation	neutral	260	1302.946533
231	paradigm	neutral	290	1300.897217
232	ligand	biology-oriented	208	1281.961914
233	control	systems-oriented	1502	1266.481934
234	redox	biology-oriented	169	1261.966553
235	mapping	systems-oriented	297	1258.210938
236	metabonomics	biology-oriented	154	1253.069214
237	vegf	biology-oriented	153	1244.932251
238	reconstruction	neutral	342	1244.490967
239	structure	neutral	977	1239.567017

240	topological	systems-oriented	170	1215.211182
241	hypothesis	neutral	383	1212.452026
242	substrate	biology-oriented	240	1195.788208
243	boolean	systems-oriented	190	1188.632813
244	multiscale	neutral	146	1187.973999
245	adaptive	biology-oriented	238	1182.934448
246	viral	biology-oriented	247	1180.207642
247	activity	neutral	867	1179.803833
248	efficacy	neutral	265	1175.087158
249	subcellular	biology-oriented	152	1167.703369
250	inhibition	biology-oriented	298	1167.310425
251	treatment	biology-oriented	888	1162.755981
252	component	neutral	441	1153.129395
253	qualitative	neutral	268	1132.602783
254	genotype	biology-oriented	178	1129.328613
255	kegg	neutral	140	1127.296753
256	diabetes	biology-oriented	269	1099.407593
257	fluorescence	biology-oriented	189	1085.655273
258	proliferation	biology-oriented	276	1085.204468
259	eukaryotic	biology-oriented	169	1074.305786
260	heterogeneity	neutral	189	1071.25354
261	biomolecular	biology-oriented	133	1070.440796

262	tf	biology-oriented	141	1068.713379
263	exposure	neutral	398	1063.145264
264	innate	biology-oriented	232	1057.610596
265	production	neutral	971	1056.266479
266	cardiovascular	biology-oriented	203	1053.821045
267	heterogeneous	neutral	210	1051.323853
268	drosophila	biology-oriented	172	1047.172852
269	molecule	biology-oriented	271	1046.451172
270	mouse	biology-oriented	365	1046.100342
271	variation	neutral	419	1043.874146
272	mapk	biology-oriented	127	1033.373291
273	omic	biology-oriented	126	1025.236328
274	inhibitor	biology-oriented	186	1019.371582
275	toxicology	biology-oriented	149	1018.670776
276	connectivity	neutral	175	1012.273621
277	therapeutics	biology-oriented	142	1011.547546
278	bioinformatic	biology-oriented	124	1008.962646
279	automated	systems-oriented	221	1002.647034
280	lc	biology-oriented	188	1001.705505
281	specificity	biology-oriented	233	998.7128906
282	challenge	neutral	610	993.2953491
283	adaptation	biology-oriented	263	990.6032104

284	transcript	biology-oriented	204	986.9748535
285	assay	biology-oriented	229	978.2520142
286	holistic	neutral	195	976.3835449
287	neuronal	biology-oriented	158	967.8212891
288	pharmacology	biology-oriented	156	967.7330933
289	peptide	biology-oriented	205	962.6178589
290	chromatin	biology-oriented	145	962.25
291	microbiome	biology-oriented	118	960.1414795
292	erk	biology-oriented	129	958.4816895
293	replication	biology-oriented	195	958.2965698
294	bacteria	biology-oriented	299	949.5690918
295	mutant	biology-oriented	227	945.4503784
296	endogenous	biology-oriented	176	940.8521729
297	glycolysis	biology-oriented	123	935.4240112
298	angiogenesis	biology-oriented	128	933.0911865
299	comparative	neutral	300	906.3901367
300	microscopy	biology-oriented	170	900.6907959

APPENDIX F

THE STANDARDIZED DEGREE CENTRALITY FOR ALL 300 KEYWORDS

This appendix shows the standardized degree centrality of 300 key words in co-word networks between 2003 and 2013.

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
biology	0.4412	0.4000	0.4863	0.4047	0.4219	0.4078	0.4163	0.4719	0.3931	0.3727	0.4133
systems	0.5588	0.4960	0.5216	0.5175	0.4297	0.4980	0.4436	0.4869	0.4618	0.4760	0.4834
gene	0.3782	0.2600	0.3059	0.2179	0.3438	0.2745	0.2840	0.3446	0.1718	0.2952	0.2546
data	0.3361	0.4160	0.3216	0.4241	0.4297	0.4196	0.4125	0.3483	0.3664	0.3764	0.4576
Keywords	0.4160	0.3240	0.3412	0.2802	0.3242	0.2510	0.3152	0.2921	0.3473	0.2399	0.3063
protein	0.2479	0.3040	0.3059	0.3268	0.3438	0.3725	0.3113	0.3521	0.3626	0.3137	0.2768
cell	0.3403	0.3200	0.2549	0.2802	0.3203	0.4314	0.2918	0.3670	0.2786	0.3432	0.3579
metabolic	0.2899	0.2920	0.2275	0.2646	0.2109	0.2000	0.3658	0.3146	0.3053	0.2362	0.3727
network	0.2983	0.2960	0.2706	0.2296	0.3867	0.2627	0.3074	0.3783	0.2519	0.2546	0.3210
molecular	0.1597	0.3200	0.3255	0.2412	0.2617	0.2510	0.2101	0.2996	0.2748	0.2804	0.2657
model	0.2857	0.3120	0.3216	0.2607	0.2617	0.2863	0.2451	0.4120	0.3588	0.3284	0.2878
expression	0.3235	0.3000	0.2549	0.1595	0.3398	0.2549	0.2490	0.2622	0.1908	0.2546	0.2841
signaling	0.1807	0.1440	0.2000	0.1595	0.1758	0.1608	0.2451	0.2285	0.1756	0.1513	0.2214

cellular	0.2017	0.2040	0.2078	0.0973	0.1953	0.2314	0.2568	0.2022	0.2366	0.1882	0.2325
genome	0.2353	0.2400	0.2000	0.1712	0.2266	0.0863	0.1323	0.2060	0.2023	0.2251	0.1993
pathway	0.1555	0.0920	0.2627	0.0973	0.1445	0.0431	0.1673	0.1348	0.0916	0.1882	0.1808
computational	0.1849	0.2120	0.1647	0.1634	0.1406	0.1137	0.1634	0.1423	0.1756	0.2362	0.1993
regulatory	0.2395	0.1840	0.1843	0.2374	0.1406	0.1176	0.1323	0.1610	0.1565	0.1808	0.1107
modeling	0.1849	0.1880	0.2353	0.1245	0.0977	0.1451	0.1790	0.1423	0.0802	0.1218	0.1328
experimental	0.1807	0.2440	0.1412	0.1556	0.2617	0.1490	0.1868	0.2060	0.1031	0.2066	0.1181
genetic	0.2437	0.1600	0.1647	0.3230	0.1641	0.1725	0.2101	0.1873	0.1908	0.2509	0.1218
metabolism	0.1471	0.1120	0.0863	0.0350	0.1523	0.1098	0.1401	0.2135	0.1832	0.0480	0.1993
functional	0.2941	0.2000	0.2745	0.1012	0.2109	0.2196	0.2101	0.1948	0.1832	0.1292	0.1550
complex	0.2143	0.2720	0.2706	0.2802	0.2930	0.2392	0.2335	0.2060	0.2328	0.2214	0.2620
disease	0.2059	0.2160	0.1922	0.1790	0.1641	0.2157	0.1634	0.2509	0.2634	0.2066	0.2030
proteomics	0.1639	0.1640	0.0941	0.1128	0.1875	0.1412	0.1284	0.0524	0.1183	0.0775	0.0701
drug	0.1134	0.2080	0.1255	0.1479	0.2617	0.1020	0.1479	0.2509	0.2061	0.0959	0.2066
throughput	0.1008	0.1000	0.1216	0.1128	0.1484	0.1412	0.0739	0.1648	0.1336	0.0738	0.1070

cancer	0.0126	0.1200	0.0863	0.1245	0.1875	0.0392	0.1673	0.1461	0.0840	0.1292	0.1808
quantitative	0.1933	0.1480	0.1373	0.1323	0.1875	0.2314	0.2257	0.0861	0.1069	0.1476	0.1439
metabolomics	0.0756	0.0800	0.0471	0.1284	0.1211	0.0863	0.0467	0.0861	0.0992	0.0701	0.1476
regulation	0.1176	0.1600	0.1216	0.0973	0.1484	0.1333	0.1595	0.1985	0.1985	0.1144	0.1587
biochemical	0.1345	0.1640	0.1725	0.1012	0.1172	0.1098	0.0700	0.0787	0.1412	0.1181	0.1550
interaction	0.0294	0.1800	0.1020	0.0973	0.1484	0.0667	0.0895	0.1273	0.1221	0.0701	0.1328
genomics	0.1765	0.1240	0.1176	0.1012	0.1719	0.1176	0.1012	0.0749	0.0763	0.0517	0.0886
dynamics	0.1345	0.1400	0.0902	0.0817	0.1133	0.0980	0.1634	0.0524	0.0878	0.0886	0.1107
genomic	0.1849	0.2160	0.1294	0.1518	0.0820	0.0353	0.1206	0.1386	0.0305	0.0812	0.1033
multiple	0.1597	0.1120	0.1647	0.0428	0.0742	0.0902	0.1128	0.1161	0.0916	0.1550	0.1771
transcriptional	0.0882	0.0680	0.0902	0.0350	0.0703	0.0863	0.0700	0.1461	0.0649	0.0738	0.1365
mathematical	0.1387	0.1360	0.1216	0.1012	0.1016	0.1216	0.1595	0.1124	0.1031	0.0996	0.1218
transcription	0.0924	0.0600	0.0431	0.1167	0.0391	0.0941	0.0973	0.1273	0.0992	0.0886	0.1181
microarray	0.1176	0.1320	0.1059	0.0661	0.0859	0.0667	0.0661	0.1049	0.0573	0.0664	0.0258
response	0.2605	0.3040	0.1608	0.2140	0.2344	0.1529	0.1401	0.2584	0.1336	0.1734	0.1771

omics	0.0000	0.0640	0.0118	0.0506	0.0664	0.0588	0.0311	0.1049	0.0611	0.0554	0.1328
behavior	0.1555	0.1200	0.0667	0.0817	0.0352	0.0235	0.1167	0.0637	0.0763	0.0812	0.0406
metabolite	0.0798	0.0560	0.0275	0.0584	0.1055	0.1059	0.0506	0.0899	0.1718	0.0738	0.1107
dynamic	0.0840	0.1640	0.1333	0.1479	0.1445	0.1569	0.0895	0.1124	0.1794	0.0996	0.0406
profiling	0.0882	0.1280	0.0510	0.0778	0.1172	0.1294	0.1167	0.2285	0.1221	0.1070	0.0590
human	0.2395	0.2240	0.1843	0.1245	0.2148	0.1882	0.1751	0.1723	0.2519	0.2509	0.1734
yeast	0.0798	0.0440	0.0353	0.0389	0.0898	0.0902	0.0700	0.1124	0.0191	0.0849	0.0738
bioinformatics	0.1050	0.0880	0.0588	0.0506	0.0859	0.0314	0.0817	0.0674	0.0420	0.0738	0.0590
simulation	0.0966	0.0760	0.0392	0.0817	0.1406	0.0000	0.0934	0.0412	0.0115	0.0996	0.0996
clinical	0.0252	0.1320	0.0549	0.1284	0.0859	0.0863	0.0895	0.1199	0.0725	0.1882	0.1292
function	0.2227	0.1280	0.1490	0.0973	0.1445	0.1373	0.1634	0.1461	0.1221	0.0886	0.1218
identification	0.0714	0.0680	0.0824	0.1323	0.1953	0.0941	0.1245	0.1348	0.0916	0.1476	0.1328
proteomic	0.1218	0.0880	0.1255	0.1051	0.0977	0.0980	0.1401	0.0300	0.1031	0.0148	0.0480
spectrometry	0.0378	0.0560	0.0353	0.0661	0.1367	0.1529	0.0895	0.0449	0.0649	0.0664	0.0554
tumor	0.0210	0.0200	0.0510	0.0856	0.0508	0.0196	0.0934	0.0599	0.0000	0.0627	0.1144

discovery	0.1387	0.1040	0.0667	0.1128	0.1523	0.1020	0.1362	0.1086	0.0573	0.0295	0.0959
proteome	0.0504	0.0200	0.1059	0.1440	0.1445	0.0588	0.0817	0.1124	0.0611	0.0738	0.0332
immune	0.0126	0.0880	0.0510	0.0545	0.0664	0.0510	0.0856	0.0000	0.0344	0.1181	0.0627
phenotype	0.0336	0.0520	0.0549	0.0661	0.0234	0.0353	0.0623	0.1199	0.0534	0.0664	0.0886
activation	0.0294	0.0760	0.0314	0.0428	0.1445	0.0902	0.1245	0.0936	0.0153	0.0849	0.0295
tissue	0.0882	0.0360	0.0510	0.0545	0.0586	0.0824	0.1206	0.1199	0.0763	0.0664	0.0554
stochastic	0.0504	0.0120	0.0667	0.0739	0.0117	0.0706	0.0700	0.0262	0.0229	0.0480	0.0000
parameter	0.0546	0.0000	0.0745	0.0545	0.0078	0.0667	0.0700	0.0150	0.0687	0.0406	0.0480
integrated	0.1639	0.0960	0.2157	0.0973	0.0820	0.1333	0.1479	0.0899	0.0992	0.1144	0.1993
physiological	0.1218	0.1600	0.0941	0.0389	0.0313	0.0902	0.0195	0.0861	0.0725	0.0701	0.0369
complexity	0.0966	0.0960	0.0902	0.0973	0.0820	0.1569	0.0623	0.1536	0.0611	0.0701	0.1144
ms	0.0252	0.0200	0.0275	0.0778	0.0508	0.0745	0.0506	0.0337	0.0763	0.0849	0.1033
development	0.2185	0.2400	0.1922	0.1829	0.1484	0.1804	0.2257	0.2247	0.2786	0.2362	0.2583
silico	0.0924	0.0560	0.0353	0.0817	0.0938	0.0431	0.0389	0.0637	0.0763	0.0554	0.0738
therapeutic	0.0588	0.0640	0.0078	0.0623	0.0703	0.0392	0.0545	0.0075	0.0344	0.0554	0.0517

integration	0.1050	0.1440	0.1098	0.1051	0.1289	0.0784	0.1051	0.1423	0.1183	0.0369	0.1328
intracellular	0.0462	0.0600	0.0549	0.0078	0.0586	0.0667	0.0817	0.0187	0.0382	0.0406	0.0590
kinetic	0.0798	0.0040	0.0353	0.0661	0.0508	0.0431	0.0156	0.0449	0.0344	0.0923	0.0554
algorithm	0.0672	0.0280	0.0275	0.0506	0.0273	0.0196	0.0000	0.0974	0.0229	0.0443	0.0258
vivo	0.0630	0.0000	0.0431	0.0817	0.0117	0.0510	0.0156	0.0037	0.1489	0.0480	0.0627
receptor	0.0630	0.0400	0.0627	0.0661	0.0703	0.0902	0.1206	0.1124	0.0802	0.0590	0.0738
transduction	0.1555	0.0280	0.1333	0.0895	0.0781	0.1176	0.0117	0.0225	0.0000	0.0221	0.0000
dna	0.1176	0.0520	0.1059	0.1051	0.0938	0.1412	0.1440	0.1386	0.0534	0.1292	0.0443
modelling	0.0966	0.0840	0.0588	0.1362	0.1094	0.1373	0.0739	0.0899	0.0344	0.0664	0.1070
microbial	0.0504	0.0800	0.0118	0.0389	0.0000	0.0353	0.0856	0.0412	0.0267	0.0738	0.0480
system	0.2227	0.2200	0.3373	0.2451	0.2422	0.3020	0.2646	0.2622	0.2557	0.3137	0.1882
flux	0.0084	0.0360	0.0078	0.0661	0.0313	0.0118	0.0311	0.0225	0.0305	0.0111	0.0443
kinase	0.0252	0.0560	0.0392	0.1051	0.0742	0.0314	0.0195	0.0637	0.0038	0.0332	0.0923
global	0.1597	0.1400	0.0824	0.1051	0.0938	0.1020	0.1284	0.1536	0.0344	0.1107	0.1476
signalling	0.0000	0.0120	0.1020	0.0311	0.0234	0.0549	0.0156	0.0487	0.0649	0.0627	0.0590

binding	0.0126	0.0360	0.0431	0.0934	0.0352	0.0314	0.0156	0.0749	0.1603	0.0406	0.0221
differential	0.0672	0.0480	0.0588	0.0778	0.0664	0.0902	0.0195	0.0449	0.0000	0.0664	0.1144
regulated	0.0462	0.0840	0.1137	0.1051	0.0664	0.0784	0.0467	0.0524	0.0954	0.0590	0.0369
robustness	0.0420	0.0600	0.0627	0.0039	0.0000	0.0314	0.1012	0.0412	0.0611	0.0258	0.0443
multi	0.0462	0.0360	0.0627	0.0272	0.0664	0.0588	0.0661	0.0300	0.0115	0.0738	0.1697
integrative	0.0546	0.0680	0.0902	0.0272	0.0313	0.0431	0.0428	0.0487	0.0916	0.0221	0.0738
information	0.2815	0.2200	0.2157	0.1829	0.2734	0.1843	0.1440	0.1723	0.1908	0.1661	0.1587
transcriptome	0.0714	0.0480	0.0235	0.0233	0.0586	0.0275	0.0156	0.0375	0.0458	0.0554	0.0775
medicine	0.0294	0.0640	0.0706	0.0156	0.1445	0.0157	0.0817	0.0637	0.1412	0.1181	0.0554
predictive	0.0294	0.0880	0.0471	0.0350	0.1328	0.0431	0.0584	0.0337	0.0878	0.0332	0.0443
cerevisiae	0.0462	0.0640	0.0118	0.0428	0.0352	0.0235	0.0778	0.0300	0.0000	0.0664	0.0701
prediction	0.0336	0.0440	0.0078	0.0272	0.0469	0.0588	0.0389	0.0749	0.0382	0.0000	0.0443
signal	0.1681	0.0560	0.1686	0.1245	0.1133	0.1137	0.0389	0.0674	0.0496	0.0480	0.0000
apoptosis	0.0294	0.0000	0.0510	0.0000	0.0820	0.0275	0.0156	0.0337	0.0000	0.0406	0.1070
mechanistic	0.0840	0.0000	0.0667	0.0545	0.0859	0.0471	0.0389	0.0936	0.0763	0.0369	0.0886

framework	0.0924	0.0920	0.1451	0.0856	0.0781	0.0588	0.0467	0.0824	0.1183	0.0812	0.0701
coli	0.1218	0.1400	0.1294	0.0545	0.0547	0.0314	0.0817	0.0225	0.0000	0.0000	0.0000
mrna	0.0546	0.0400	0.0510	0.0156	0.0898	0.0275	0.0973	0.0412	0.0382	0.0590	0.0332
synthetic	0.0168	0.0080	0.0275	0.0272	0.0781	0.0706	0.0895	0.0749	0.0611	0.0480	0.0775
growth	0.1513	0.1240	0.1020	0.0739	0.0625	0.0980	0.2101	0.2060	0.1107	0.1292	0.1033
phosphorylation	0.0210	0.0120	0.0902	0.0117	0.0313	0.1294	0.0078	0.0674	0.0344	0.0221	0.1218
msb	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
chemical	0.1429	0.0560	0.1529	0.1440	0.1055	0.0706	0.0506	0.1199	0.0878	0.0627	0.1033
tcm	0.0000	0.0000	0.0235	0.0000	0.0000	0.0000	0.0895	0.0000	0.0305	0.0369	0.0074
vitro	0.0000	0.0240	0.0392	0.0389	0.0469	0.0431	0.0350	0.0899	0.0687	0.0185	0.0295
optimization	0.0210	0.0200	0.0039	0.0389	0.0625	0.0471	0.0233	0.0262	0.0534	0.0406	0.0590
rna	0.0168	0.1000	0.0667	0.0661	0.0273	0.0353	0.0195	0.0599	0.1069	0.0886	0.0627
robust	0.0126	0.0000	0.0157	0.0506	0.0664	0.0392	0.0428	0.0449	0.0611	0.0849	0.0148
host	0.0294	0.0440	0.0275	0.0661	0.0469	0.0118	0.1051	0.0899	0.0840	0.0517	0.0369
arabidopsis	0.0294	0.0720	0.0078	0.0156	0.0234	0.0549	0.0272	0.0487	0.0573	0.0590	0.0148

engineering	0.0756	0.0240	0.0235	0.0350	0.1094	0.1098	0.0661	0.1348	0.0305	0.1328	0.1218
sequencing	0.0252	0.0640	0.0902	0.0467	0.0000	0.0471	0.0000	0.0599	0.0420	0.0369	0.0517
target	0.0798	0.1040	0.0863	0.0700	0.1406	0.0353	0.0311	0.1423	0.1756	0.0812	0.1292
metabolomic	0.0168	0.0680	0.0196	0.0195	0.0000	0.0235	0.1167	0.0262	0.0153	0.0258	0.0111
biomarker	0.0000	0.0000	0.0000	0.0428	0.0781	0.0118	0.0545	0.0337	0.0229	0.0258	0.0517
enzyme	0.0336	0.0640	0.1020	0.0000	0.0469	0.0196	0.0078	0.0037	0.0458	0.0258	0.0664
comprehensive	0.0756	0.1240	0.1294	0.1401	0.1797	0.0980	0.0545	0.1423	0.1031	0.1328	0.1070
mass	0.0462	0.1000	0.0549	0.1128	0.1445	0.1490	0.1401	0.0749	0.0802	0.0664	0.0812
translational	0.0210	0.0000	0.0784	0.0467	0.0195	0.0392	0.0661	0.0899	0.0458	0.0258	0.0480
organism	0.0588	0.1680	0.1216	0.0778	0.0391	0.0902	0.0311	0.0599	0.0992	0.0627	0.0221
annotation	0.0378	0.0000	0.0667	0.0000	0.0391	0.0039	0.0117	0.0487	0.0382	0.0000	0.0000
toxicity	0.1008	0.0560	0.0353	0.0778	0.0313	0.0000	0.1012	0.0337	0.0000	0.0000	0.0295
stress	0.0336	0.0360	0.0745	0.0039	0.0547	0.0667	0.1595	0.0749	0.0573	0.0849	0.0554
personalized	0.0210	0.0200	0.0000	0.0117	0.0313	0.0118	0.0389	0.0225	0.0458	0.0443	0.0332
phenotypic	0.0714	0.0240	0.0549	0.0623	0.0469	0.0000	0.0078	0.0150	0.0649	0.0148	0.0295

cycle	0.0924	0.1280	0.0745	0.0156	0.0938	0.0353	0.1089	0.1086	0.0153	0.0664	0.0000
mitochondrial	0.0000	0.0200	0.0118	0.0000	0.1016	0.0471	0.0467	0.0974	0.0458	0.0000	0.0886
web	0.0210	0.0120	0.0549	0.0233	0.1016	0.0431	0.0156	0.0000	0.0458	0.0221	0.0221
visualization	0.0210	0.0200	0.0471	0.0272	0.0234	0.0157	0.0311	0.0449	0.0305	0.0000	0.0000
sbml	0.0378	0.0120	0.0157	0.0506	0.0000	0.0118	0.0156	0.0000	0.0000	0.0221	0.0000
database	0.0546	0.0920	0.0196	0.1089	0.1602	0.0392	0.0584	0.0375	0.0076	0.0185	0.0221
transcriptomics	0.0462	0.0600	0.0314	0.0389	0.0508	0.0392	0.0000	0.0375	0.0458	0.0664	0.0000
sensitivity	0.0630	0.0040	0.0157	0.0195	0.0625	0.0157	0.0739	0.0262	0.0267	0.0443	0.0332
analytical	0.0588	0.0840	0.0706	0.1712	0.0195	0.0471	0.0623	0.0674	0.0267	0.0590	0.0369
saccharomyces	0.0336	0.0720	0.0118	0.0389	0.0391	0.0235	0.0778	0.0375	0.0000	0.0590	0.0664
validation	0.0546	0.1000	0.0471	0.0389	0.1055	0.0549	0.0156	0.0637	0.0458	0.0775	0.0111
nmr	0.0504	0.0000	0.0235	0.0895	0.0273	0.0275	0.0856	0.0412	0.1031	0.0443	0.0332
metabolome	0.0126	0.0560	0.0392	0.0661	0.0313	0.0118	0.0039	0.0749	0.0153	0.0111	0.0886
mechanism	0.0504	0.0200	0.0510	0.0233	0.0313	0.0196	0.0778	0.0412	0.0267	0.0701	0.0849
kinetics	0.0336	0.0560	0.0667	0.0195	0.0234	0.0157	0.0000	0.0000	0.0267	0.0221	0.0148

evolution	0.0798	0.0560	0.0471	0.1051	0.0938	0.0471	0.0895	0.0412	0.0344	0.0406	0.0923
biosynthesis	0.0084	0.0120	0.0314	0.0000	0.0195	0.0627	0.0000	0.0150	0.0382	0.0000	0.0590
physiology	0.0420	0.0600	0.0784	0.0233	0.0469	0.1020	0.0078	0.0674	0.0153	0.0258	0.0590
glucose	0.0000	0.0240	0.0431	0.0545	0.0352	0.0039	0.0545	0.0375	0.0687	0.0000	0.0664
homeostasis	0.0294	0.0120	0.0157	0.0156	0.0078	0.0431	0.0233	0.0262	0.0344	0.0406	0.0000
biomedical	0.0000	0.0200	0.0196	0.0233	0.0000	0.0157	0.0078	0.0187	0.0153	0.0037	0.0000
escherichia	0.0840	0.1040	0.1059	0.0389	0.0469	0.0353	0.0623	0.0262	0.0000	0.0000	0.0074
plant	0.0336	0.0600	0.0549	0.0817	0.0938	0.0431	0.0623	0.0449	0.0344	0.0775	0.0332
structural	0.0840	0.0000	0.0510	0.0584	0.0430	0.0627	0.0856	0.1199	0.0420	0.0295	0.0221
aging	0.0000	0.0000	0.0235	0.0000	0.0078	0.0118	0.0623	0.0375	0.1374	0.0000	0.0111
dimensional	0.0420	0.0760	0.0431	0.0389	0.0703	0.0471	0.0272	0.0375	0.0649	0.0554	0.0554
lipid	0.0000	0.0240	0.0549	0.0000	0.0195	0.0000	0.0739	0.0674	0.0153	0.0148	0.0886
species	0.1387	0.0360	0.0588	0.0817	0.0898	0.1294	0.0895	0.1199	0.0687	0.1402	0.0295
statistical	0.0630	0.0800	0.0980	0.0934	0.0547	0.0353	0.0584	0.0674	0.0344	0.0185	0.1439
reaction	0.0588	0.1000	0.0588	0.1206	0.0977	0.0510	0.1206	0.0861	0.0496	0.0849	0.1144

feedback	0.0546	0.1120	0.0588	0.0272	0.0234	0.0157	0.0350	0.0225	0.0763	0.0185	0.0221
inference	0.0042	0.0000	0.0000	0.0156	0.0195	0.0314	0.0000	0.0412	0.0153	0.0295	0.0295
pathogen	0.0000	0.0280	0.0000	0.0350	0.0469	0.0392	0.0195	0.0262	0.0382	0.0258	0.0185
transcriptomic	0.0336	0.0200	0.0784	0.0778	0.0000	0.0000	0.0428	0.0262	0.0687	0.0627	0.0295
evolutionary	0.0294	0.0280	0.0118	0.0000	0.0664	0.0196	0.0545	0.0112	0.0763	0.0664	0.0701
progression	0.0252	0.0200	0.0314	0.0272	0.0000	0.0235	0.1206	0.0562	0.0802	0.1033	0.0590
factor	0.0756	0.0560	0.0980	0.1206	0.0703	0.1020	0.0584	0.1236	0.0496	0.0480	0.1328
emerging	0.1387	0.0760	0.1373	0.0895	0.0938	0.0588	0.0350	0.0899	0.0840	0.0000	0.0443
dynamical	0.0378	0.0440	0.0392	0.0545	0.0352	0.0392	0.0039	0.0300	0.0420	0.0074	0.0701
mammalian	0.0924	0.0320	0.0353	0.0467	0.0391	0.0196	0.0428	0.0787	0.0229	0.0738	0.0111
ontology	0.0000	0.0000	0.0157	0.0078	0.0430	0.0314	0.0195	0.0337	0.0000	0.0074	0.0000
inflammatory	0.0546	0.0680	0.0431	0.0156	0.0195	0.0157	0.0428	0.0112	0.0076	0.0111	0.0664
application	0.0840	0.0600	0.1059	0.1518	0.1328	0.1176	0.1206	0.0712	0.1183	0.0185	0.1292
temporal	0.0714	0.0000	0.0471	0.0311	0.0547	0.0392	0.0000	0.0000	0.0878	0.0295	0.0221
design	0.1134	0.0840	0.0745	0.0739	0.0664	0.1686	0.1012	0.1124	0.1565	0.1697	0.1476

differentiation	0.0546	0.0160	0.0510	0.0778	0.0000	0.0667	0.0000	0.0936	0.1031	0.0369	0.0775
role	0.1092	0.0760	0.1608	0.0895	0.1055	0.1412	0.0467	0.1236	0.1221	0.1550	0.1476
nonlinear	0.0588	0.0320	0.0196	0.0428	0.0000	0.0314	0.0000	0.0412	0.0267	0.0111	0.0185
bacterial	0.0672	0.0240	0.0118	0.0117	0.0234	0.0784	0.0000	0.0300	0.0611	0.1144	0.0111
oxidative	0.0000	0.0160	0.0000	0.0156	0.0508	0.0118	0.0545	0.0112	0.0687	0.0517	0.1218
environmental	0.1681	0.1040	0.0314	0.1089	0.0234	0.1373	0.0311	0.0824	0.1107	0.0590	0.0849
bayesian	0.0126	0.0000	0.0000	0.0156	0.0078	0.0118	0.0117	0.0187	0.0000	0.0221	0.0369
liver	0.0000	0.0360	0.0157	0.0272	0.0234	0.0000	0.0817	0.0225	0.0038	0.1255	0.0480
dependent	0.0294	0.0200	0.0353	0.0117	0.0000	0.0706	0.1089	0.0674	0.0687	0.0701	0.0443
developmental	0.0924	0.0240	0.0902	0.0350	0.0078	0.0784	0.0117	0.0449	0.0840	0.0664	0.0221
mirna	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0406	0.0000
imaging	0.0000	0.0000	0.0235	0.0117	0.0273	0.0039	0.0233	0.0000	0.0115	0.0258	0.0000
diverse	0.1092	0.0720	0.0510	0.0428	0.0547	0.0392	0.0545	0.0150	0.0344	0.0480	0.0480
circadian	0.0378	0.0000	0.0196	0.0311	0.0000	0.0000	0.0000	0.0487	0.0382	0.0332	0.0000
systematic	0.0714	0.0920	0.0824	0.0584	0.0586	0.0627	0.0934	0.0562	0.0954	0.0738	0.0480

ppi	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0262	0.0191	0.0000	0.0074
stem	0.0000	0.0520	0.0000	0.0156	0.0234	0.1098	0.0000	0.0712	0.0191	0.0554	0.0554
epigenetic	0.0000	0.0000	0.0000	0.0000	0.0078	0.0235	0.0000	0.0187	0.0229	0.0590	0.0111
interactome	0.0000	0.0240	0.0392	0.0000	0.0000	0.0235	0.0233	0.0150	0.0191	0.0148	0.0332
chromatography	0.0000	0.0520	0.0157	0.0350	0.0508	0.0627	0.0545	0.0300	0.0573	0.0517	0.0258
inflammation	0.0546	0.0120	0.0000	0.0000	0.0195	0.0549	0.0000	0.0300	0.0344	0.0258	0.0369
synthesis	0.0504	0.0440	0.0549	0.0233	0.1133	0.0627	0.1128	0.0637	0.0344	0.0443	0.0369
amino	0.0168	0.0200	0.0549	0.0195	0.0234	0.0392	0.0545	0.0187	0.0153	0.0148	0.0406
detection	0.0546	0.0480	0.0000	0.0739	0.0781	0.0549	0.0311	0.0225	0.0076	0.0111	0.0111
systemic	0.0588	0.0520	0.0392	0.0117	0.0156	0.0549	0.0311	0.0787	0.0573	0.0295	0.0258
acid	0.0084	0.0320	0.0431	0.0233	0.0234	0.0314	0.0700	0.0562	0.0305	0.0185	0.0517
quantification	0.0000	0.0200	0.0000	0.0467	0.0664	0.0157	0.0311	0.0000	0.0458	0.0148	0.0000
pathogenesis	0.0000	0.0360	0.0000	0.0000	0.0000	0.0000	0.0311	0.0337	0.0305	0.0185	0.0221
topology	0.0168	0.0120	0.0510	0.0000	0.0156	0.0000	0.0000	0.0524	0.0229	0.0037	0.0185
variability	0.0000	0.0160	0.0118	0.0195	0.0195	0.0000	0.0467	0.0112	0.0305	0.0295	0.0185

atp	0.0294	0.0000	0.0000	0.0000	0.0313	0.0196	0.0039	0.0000	0.0344	0.0037	0.0037
optimal	0.0336	0.0280	0.0431	0.0700	0.0000	0.0471	0.0350	0.0337	0.0076	0.0923	0.0074
therapy	0.0000	0.0360	0.0157	0.0389	0.0508	0.0667	0.0467	0.0861	0.0344	0.0590	0.0996
genetics	0.0000	0.0400	0.0353	0.0506	0.0469	0.0275	0.0506	0.0300	0.0649	0.0258	0.0406
membrane	0.0000	0.0160	0.0275	0.0467	0.0156	0.0588	0.0428	0.0112	0.0305	0.0000	0.0443
vaccine	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0078	0.0075	0.0000	0.0148	0.0332
extracellular	0.0000	0.0440	0.0196	0.0389	0.0469	0.0314	0.0428	0.0637	0.0382	0.0185	0.0221
infection	0.0000	0.0000	0.0118	0.0000	0.0547	0.0000	0.0233	0.0375	0.0611	0.0996	0.0443
clustering	0.0336	0.0280	0.0118	0.0233	0.0000	0.0118	0.0467	0.0000	0.0076	0.0221	0.0000
thaliana	0.0252	0.0440	0.0078	0.0117	0.0000	0.0353	0.0000	0.0375	0.0420	0.0295	0.0148
platform	0.0294	0.0640	0.0000	0.0467	0.0586	0.0627	0.0700	0.0449	0.0763	0.0627	0.0923
perturbation	0.0378	0.0640	0.0118	0.0000	0.0352	0.0000	0.0117	0.0112	0.0153	0.0221	0.0185
plasma	0.0000	0.0360	0.0275	0.0545	0.0313	0.0627	0.0389	0.0599	0.0382	0.0074	0.0480
linear	0.0210	0.0160	0.0275	0.0700	0.0352	0.0235	0.0078	0.0449	0.0611	0.0664	0.0664
mice	0.0546	0.0240	0.0000	0.0506	0.0117	0.0314	0.0623	0.0075	0.0191	0.0258	0.0074

sequence	0.0672	0.0560	0.0392	0.0428	0.0898	0.0353	0.0428	0.0150	0.0611	0.0443	0.0443
multivariate	0.0168	0.0480	0.0235	0.0156	0.0547	0.0196	0.0623	0.0787	0.0115	0.0295	0.0185
dataset	0.0000	0.0080	0.0000	0.0195	0.0273	0.0000	0.0156	0.0000	0.0000	0.0000	0.0406
correlation	0.0084	0.0280	0.0196	0.0117	0.0273	0.0157	0.0039	0.0337	0.0382	0.0443	0.0664
egfr	0.0588	0.0320	0.0353	0.0584	0.0000	0.0314	0.0117	0.0337	0.0191	0.0258	0.0627
screening	0.0168	0.0000	0.0431	0.0856	0.0859	0.0392	0.0272	0.0637	0.1221	0.0221	0.0148
estimation	0.0252	0.0000	0.0353	0.0661	0.0156	0.0000	0.0000	0.0000	0.0763	0.0111	0.0517
paradigm	0.0546	0.0920	0.0392	0.0545	0.0000	0.0353	0.0545	0.0562	0.0458	0.0258	0.0000
ligand	0.0000	0.0240	0.0275	0.0039	0.0508	0.0000	0.0195	0.0787	0.0496	0.0000	0.0258
control	0.1555	0.1840	0.0863	0.1829	0.0859	0.1137	0.2218	0.1573	0.1718	0.1181	0.1365
redox	0.0168	0.0000	0.0000	0.0000	0.0195	0.0431	0.0195	0.0000	0.0000	0.0000	0.0812
mapping	0.0336	0.0320	0.0667	0.0856	0.0469	0.0863	0.0233	0.0000	0.0305	0.0148	0.0812
metabonomics	0.0588	0.0360	0.0431	0.0739	0.0273	0.0078	0.0506	0.0000	0.0000	0.0258	0.0000
vegf	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0117	0.0712	0.0000	0.0000	0.0590
reconstruction	0.0336	0.0480	0.0314	0.0117	0.1055	0.0275	0.0272	0.0225	0.0000	0.0295	0.0111

structure	0.1765	0.0840	0.1098	0.0506	0.1133	0.0392	0.0428	0.1011	0.1221	0.1107	0.0554
topological	0.0000	0.0000	0.0000	0.0117	0.0078	0.0039	0.0000	0.0225	0.0000	0.0369	0.0074
hypothesis	0.0504	0.0400	0.0392	0.0272	0.0820	0.0275	0.0272	0.0000	0.0229	0.0627	0.0295
substrate	0.0336	0.0440	0.0588	0.0000	0.0352	0.0118	0.0117	0.0075	0.0000	0.0111	0.0295
boolean	0.0084	0.0000	0.0000	0.0233	0.0039	0.0000	0.0000	0.0075	0.0191	0.0258	0.0369
multiscale	0.0000	0.0000	0.0157	0.0000	0.0000	0.0824	0.0000	0.0075	0.0000	0.0185	0.0000
adaptive	0.0336	0.0000	0.0392	0.0000	0.0078	0.0000	0.0000	0.0075	0.0115	0.1328	0.0701
viral	0.0000	0.0360	0.0863	0.0233	0.0195	0.0000	0.0195	0.0000	0.0000	0.0664	0.0664
activity	0.0924	0.0800	0.0627	0.0389	0.0898	0.0196	0.1167	0.0974	0.0649	0.0111	0.0812
efficacy	0.0168	0.0480	0.0000	0.0350	0.0000	0.0039	0.0584	0.0150	0.0573	0.0738	0.0443
subcellular	0.0168	0.0240	0.0078	0.0117	0.0000	0.0000	0.0195	0.0150	0.0191	0.0074	0.0000
inhibition	0.0126	0.0680	0.0549	0.0117	0.0078	0.0196	0.0039	0.0449	0.0763	0.0332	0.0369
treatment	0.0378	0.0640	0.0196	0.0195	0.0508	0.1255	0.0856	0.1423	0.0802	0.0627	0.1292
component	0.0420	0.0200	0.1020	0.0623	0.0547	0.0157	0.0311	0.0861	0.0763	0.0517	0.0775
qualitative	0.0756	0.0480	0.0000	0.0233	0.0078	0.0314	0.0623	0.0187	0.0344	0.0221	0.0221

genotype	0.0042	0.0240	0.0275	0.0272	0.0430	0.0000	0.0311	0.0487	0.0000	0.0185	0.0369
kegg	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0389	0.0000	0.0000	0.0111	0.0037
diabetes	0.0126	0.0200	0.0314	0.0000	0.0000	0.0157	0.0000	0.0075	0.0267	0.0000	0.0406
fluorescence	0.0000	0.0000	0.0000	0.0233	0.0000	0.0000	0.0078	0.0000	0.0153	0.0295	0.0221
proliferation	0.0294	0.0080	0.0392	0.0156	0.0000	0.0392	0.0272	0.0112	0.0267	0.0332	0.0701
eukaryotic	0.0294	0.0120	0.0000	0.0000	0.0273	0.0000	0.0156	0.0262	0.0000	0.0221	0.0185
heterogeneity	0.0000	0.0000	0.0000	0.0156	0.0156	0.0118	0.0233	0.0375	0.0191	0.0627	0.0000
biomolecular	0.0000	0.0000	0.0745	0.0039	0.0117	0.0000	0.0000	0.0487	0.0153	0.0148	0.0000
tf	0.0000	0.0000	0.0549	0.0000	0.0000	0.0000	0.0000	0.0787	0.0267	0.0000	0.0332
exposure	0.0714	0.0520	0.0078	0.0545	0.0234	0.0078	0.0117	0.0712	0.0344	0.0554	0.0111
innate	0.0000	0.0920	0.0431	0.0117	0.0000	0.0000	0.0195	0.0000	0.0115	0.0517	0.0590
production	0.0000	0.0240	0.0275	0.0195	0.0273	0.0275	0.0428	0.0861	0.0305	0.0923	0.1255
cardiovascular	0.0630	0.0120	0.0000	0.0000	0.0000	0.0157	0.0467	0.0861	0.0305	0.0074	0.0037
heterogeneous	0.0294	0.0320	0.0353	0.0545	0.0195	0.0157	0.0000	0.0375	0.0000	0.0000	0.0554
drosophila	0.0252	0.0160	0.0157	0.0272	0.0313	0.0000	0.0311	0.0000	0.0038	0.0000	0.0185

molecule	0.0672	0.0280	0.0157	0.0117	0.0352	0.0588	0.0000	0.0375	0.0687	0.0000	0.0406
mouse	0.0378	0.0240	0.0627	0.0506	0.0430	0.0118	0.0272	0.0974	0.0305	0.0000	0.0775
variation	0.0672	0.0240	0.0275	0.0311	0.0586	0.0235	0.0506	0.0300	0.0649	0.0443	0.0332
mapk	0.0000	0.0480	0.0353	0.0000	0.0195	0.0000	0.0117	0.0262	0.0115	0.0221	0.0000
omic	0.0504	0.0000	0.0549	0.0156	0.0000	0.0000	0.0389	0.0000	0.0573	0.0111	0.0000
inhibitor	0.0252	0.0000	0.0196	0.0389	0.0000	0.0000	0.0311	0.0637	0.0191	0.0221	0.0590
toxicology	0.1176	0.0160	0.0431	0.0117	0.0391	0.0000	0.0000	0.0637	0.0076	0.0148	0.0000
connectivity	0.0588	0.0000	0.0863	0.0195	0.0195	0.0235	0.0156	0.0487	0.0115	0.0185	0.0000
therapeutics	0.0000	0.0160	0.0000	0.0000	0.0117	0.0039	0.0000	0.0225	0.0000	0.0258	0.0000
bioinformatic	0.0000	0.0320	0.0235	0.0156	0.0078	0.0471	0.0000	0.0300	0.0382	0.0221	0.0074
automated	0.0000	0.0480	0.0000	0.0195	0.0156	0.0275	0.0117	0.0300	0.0916	0.0258	0.0148
lc	0.0000	0.0000	0.0000	0.0389	0.0625	0.0235	0.0117	0.0262	0.0382	0.0258	0.0517
specificity	0.0000	0.0200	0.0314	0.0156	0.0234	0.0000	0.0545	0.0037	0.0191	0.0221	0.0000
challenge	0.0252	0.0400	0.0588	0.0311	0.0938	0.0431	0.0000	0.0562	0.0725	0.0554	0.1033
adaptation	0.0420	0.0120	0.0235	0.0000	0.0156	0.0235	0.0000	0.0187	0.0000	0.0775	0.0221

transcript	0.0210	0.0400	0.0000	0.0311	0.0000	0.0471	0.0000	0.0112	0.0153	0.0369	0.0258
assay	0.0168	0.0680	0.0000	0.0117	0.0078	0.0000	0.0000	0.0037	0.0000	0.0000	0.0369
holistic	0.0000	0.0640	0.0000	0.0428	0.0078	0.0157	0.0233	0.0112	0.0305	0.0886	0.0000
neuronal	0.0000	0.0000	0.0000	0.0000	0.0039	0.0235	0.0000	0.0000	0.0038	0.0664	0.0000
pharmacology	0.0126	0.0000	0.0275	0.0350	0.0156	0.0000	0.0000	0.0187	0.0000	0.0627	0.0000
peptide	0.0000	0.0000	0.0000	0.0156	0.0664	0.0118	0.0117	0.0000	0.0000	0.0369	0.0369
chromatin	0.0000	0.0000	0.0000	0.0272	0.0000	0.0000	0.0000	0.0000	0.0076	0.0074	0.0000
microbiome	0.0000	0.0000	0.0000	0.0000	0.0039	0.0157	0.0545	0.0000	0.0954	0.0148	0.0000
erk	0.0000	0.0400	0.0157	0.0311	0.0000	0.0000	0.0233	0.0000	0.0000	0.0000	0.0000
replication	0.0000	0.0680	0.0392	0.0000	0.0078	0.0471	0.0661	0.0112	0.0000	0.0369	0.0000
bacteria	0.0252	0.0520	0.0588	0.0000	0.0195	0.0471	0.0233	0.0000	0.0191	0.0221	0.0369
mutant	0.0210	0.0320	0.0000	0.0000	0.0000	0.0118	0.0117	0.0337	0.0115	0.0000	0.0664
endogenous	0.0756	0.0080	0.0196	0.0117	0.0195	0.0235	0.0000	0.0262	0.0000	0.0258	0.0664
glycolysis	0.0000	0.0000	0.0196	0.0000	0.0156	0.0000	0.0117	0.0000	0.0267	0.0000	0.0443
angiogenesis	0.0000	0.0200	0.0471	0.0000	0.0000	0.0000	0.0000	0.0375	0.0000	0.0185	0.0369

comparative	0.0714	0.0080	0.0000	0.0584	0.0430	0.0431	0.0117	0.0300	0.0076	0.0258	0.0443
microscopy	0.0000	0.0000	0.0000	0.0000	0.0000	0.0157	0.0000	0.0000	0.0344	0.0148	0.0369

APPENDIX G

THE COUNTRY ORIGIN OF CORRESPONDING AUTHORS FROM 2005 TO 2013

This appendix shows the number of articles that have corresponding authors in five major countries and other countries.

	US	China	Germany	England	Japan	Other countries	Total
2000	1	0	0	0	1	0	2
2001	5	0	0	0	0	0	5
2002	8	0	3	1	4	9	25
2003	63	0	12	10	1	22	108
2004	116	0	30	25	6	64	241
2005	189	7	42	31	15	117	401
2006	249	16	59	50	15	183	572
2007	258	18	88	60	15	216	655
2008	305	35	91	83	18	311	843
2009	404	32	89	99	29	387	1040
2010	483	69	119	105	38	455	1269
2011	458	80	146	121	31	513	1349
2012	464	112	147	120	34	541	1418
2013	445	129	119	109	32	529	1363
Growth rate	0.1129	0.4394	0.1390	0.1701	0.0993	0.2076	0.1652