Trend Prediction of Biomedical Technology by Semantic Analysis

Xiaomeng Sun[1,3], Kexu Zhang[2], Peng Nan[3], Lei Liu[1]

[1]Fudan University
No.138, Medical School Road, Shanghai, China
liulei@fudan.edu.cn
[2]Shanghai Ebuinfo Co., Ltd.
No. 170, Lane 277, Chen Hui Road, Shanghai, China
kexu.zhang@ebuinfo.com
[3]Fudan University
No. 2005, Songhu Road, Shanghai, China
nanpeng@fudan.edu.cn

Abstract: This paper proposes a solution to establish a biomedical technology analysis platform to laying foundation for an expert knowledge-based biomedical system, which aims to give intelligent medical decision in the end. It curates 23.44 million biomedical articles since 1970 for serving as repository to supporting as knowledge base. Based on the platform, trend prediction in biomedical technology has been made by semantic analysis.

1 Introduction

Scientific researches lead great progress on human race in all ages, especially the science of health and medicine. With the explosive growth of digitized data in the past few decades, remarkable development has been achieved in the realm of biomedicine study, computational technology and their crossing domain. Transforming experience based biomedical knowledge into computer sensitive language, developing logical algorithm to learn the pattern and give optimized diagnosis, building intelligent autonomic learning knowledge-based systems are of great interest today [1,2,3,4].

An expert knowledge-based system (KBS) stores, organizes, manages and utilizes complex structured and unstructured domain-specific information to give valuable knowledge from mass data, and therefore solves complex problems that we can't deal with before the era of information explosion. Medical library is the most widely used KBS which contains full text or abstract of medical research articles [5,6]. It normally provides multiple search interfaces, grouping and sorting feedback, citation indexes and knowledge network graphs for user retrieval request [7,8]. It represents the simplest model of KBS and settles the basis for more advanced knowledge integrated and decision making models.

Below, we elaborate on the challenges facing effective knowledge discovery from mass data and describe a biomedical technology analysis system that contains 23.44 million articles in the field of biomedical research. At the core of this platform is technical classification and hotspot keywords analysis so far. Both results are based on word frequency statistics.

2 Biomedical technology analysis platform

This platform has collected large amount of biomedical articles from different open source databases on the Internet from February 1970 to February 2015. Considering the quality of the data, we eventually selected the biomedical articles of recent ten years, which is from February 2005 to February 2015, as our analysis object. Data of these ten years includes 0.5 million Chinese articles and 22.94 million English articles. All of the Chinese information is gathered from the WANFANG MED ONLINE (http://www.wanfangdata.com.cn) website with only abstracts, headings, keywords, authors and some other related data. English information is downloaded mainly from PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) and BioMed Central (http://www.biomedcentral.com) in the same form, along with 0.8 million full text among them.

Again, building on the biomedical library, we employed word frequency statistics to yield accurate technical classification and keyword analysis. An important component of this platform is the ability to automatically capture technical keywords and roughly cluster articles in different technical categories. The platform illustrates each technical classification and its article amount in percentage. Also important for this platform use is the ability to acquire research hotspots through keyword frequency statistics. This result is especially useful for industry development analysis, which helps us better understanding the ongoing biomedical research territory.

3 Trend prediction of biomedical technology by semantic analysis

3.1 Method

Gaining and reediting biomedical research spheres information from UMLS system (http://www.nlm.nih.gov/research/umls/), we summarized a list of widely used biomedical technologies in the recent decade. This list of words then served as search keys by scanning the library to cluster articles into technical classifications and give their corresponding distribution probabilities. This process requires repeatedly integration and optimization according to the distribution result to approach the most reasonable and balanced classification result.

To verify and support this biomedical technology taxonomy, we simultaneously conducted content analysis based on keyword. Word frequency statistics is the analytic method of giving list of keywords grouped by frequency of occurrence within the

giving text corpus. By extracting and sorting keywords from Chinese and English library, we generated 3512 Chinese keywords and 440,650 English keywords, with total frequency amount of 87932 and 1,693,163 times respectively.

3.2 Result

At the beginning, we chose the top 500 keywords in both languages as raw data for further discussion. For Chinese library, in contrast to the fact that 500 only account for 14.24% of total keyword amount, they contain 67% of total frequency, which means that they represent 67% of all Chinese articles. Both figures for English library are 0.11% and 18.16% respectively. These indexes strongly indicate these 1000 keywords could reflect the overall biomedical study in both languages.

Combining evidences from both technical classification and keyword analysis, screening obscure content, connecting categories with detailed high frequency keywords, we exhibit top 3 technical classification results and their high frequency subcategory keywords in English language (Table 1).

Table. 1. Top three technical classifications and their keywords with frequency

Category	Keyword	Frequency
Genomics Technology	Genome wide association study (GWAS)	38534
	Next generation sequencing (NGS)	20473
	Single cell transcriptome sequencing	19695
	Chromatin structural genomics	13527
	Haplotype analysis	11289
Proteomics Technology	Western blot	60275
	Immunofluorescence	23688
	Protein microarray	14470
	Isoelectric focusing (IEF)	12681
	Bioinformatics analysis	6185
Glycomics Technology	Glyco-catch method	5357
	Fluorophore-assisted carbohydrate electrophoresis (FACE)	4440
	Two-step proteolytic digestion combined with sequential microcolumns technology	1726
	Chemoselective glycoblotting	1500
	Mass spectrum	1444

Since the announcement of the essentially complete human genome a decade ago, various omics subjects welcome a booming generation by benefitting from the rapid progress of biotechnology and computer science. GWAS, NGS, different kinds of electrophoresis and GC-MS technology are shared a comparative short history and rapid development speed in general. Among every top three well-studied omics areas, these technologies are all widely used, that we can tell from the investigated frequency showed in table 1. In this statistical result, articles related to genomics, proteomics and glycomics study amount to 21.30% of the English library, accounted for a large proportion of biomedical research in recent ten years. This article will give extra analysis of genomics study and its relevant crucial technologies.

4 Discussion

Genomics study dedicates on solving problems of genome in two directions: structure and function. Therefore the field includes efforts to determine the entire DNA sequence of organisms and fine-scale genetic mapping, which could be inducted as structural genomics, and to understand the intragenomic phenomena and other interactions between loci and alleles within the genome, which can be deemed to functional genomics or postgenome study.

Table. 2. Genomics technology major sub-classifications with frequency

Category	Keyword	Frequency
Genomics Technology	Genome wide association study (GWAS)	38534
	Next generation sequencing (NGS)	20473
	Single cell transcriptome sequencing	19695
	Chromatin structural genomics	13527
	Haplotype analysis	11289
	Pan genomics technology	10827
	Immune genomics technology	9438
	Epigenetics technology	5974
	Nutrigenomics technology	5161
	Pharmacogenomics related technology	3950
	Metagenomics technology	2564
	Druggable genome discovery technology	1107

Next generation sequencing (NGS), chromatin structural genomics and single cell transcriptome sequencing focus on the first question. NGS technologies enable us to

parallelize the sequencing process, producing millions of sequences concurrently, lower the cost of DNA sequencing to a great extent and make sequencing much more convenient and more accurate than ever before. NGS sets a lower base for conducting genomics study today and helps to accelerate the knowledge discovery routine. One important purpose of getting genome sequences is to find out the connection between human genome and drug susceptibility rules of gene expression and interpreting transcriptional bursting phenomenon. Beyond DNA studies, single cell transcriptome sequencing gives information about different kinds of RNA molecules of a specified cell population. With that, the processes of cellular differentiation and carcinogenesis can be learned in depth. Articles clustered under these keywords verify our research hotspot analysis.

Genome wide association study (GWAS) and haplotype analysis, on another direction, belong to the fundamental functional genomics research. If structural genomics is about gathering information, functional genomics works to make sure of these vast produced data turning into valuable knowledge wealth. GWAS exams common genetic variants in different individuals to test if any variant is associated with a trait, mainly focused on associations between single nucleotide polymorphisms (SNPs) and traits like major diseases. Haplotype analysis also proceeds with SNPs and polymorphic sites to provide valuable evidence for investigating the genetics of common diseases. Relevancy to medical care and drug delivery explains the high frequency of both technologies.

Besides, there are some other areas of expertise in our statistical list: pan genomics technology, immune genomics technology, epigenetics technology, nutrigenomics technology, metagenomics technology and druggable genome discovery technology. Each keyword mirrors an aspect of genomics study and numerous neoteric technologies, combined to form a tip of the iceberg in biomedical research today. To better understand the trend in biomedical industry, we furthermore looked into the top five most studied sub-classifications in genomics technology.

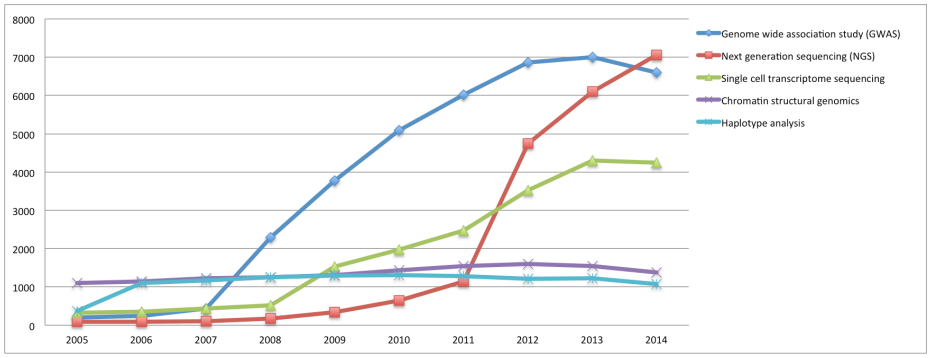

Fig. 1. Article amount of top five genomics technology in last years

Figure 1 concludes the article amount of top five genomics technologies of past ten years, while genomics study has faced huge improvement by means of technology advancement. GWAS' application experiences rapidly growth since 2007, peaks in

2013, and still holds leading position of genomics study this year. NGS also sees a quick growth stage around 2013 with even higher rates of increase and surpasses GWAS last year to becoming the most widely used genomics technology today. Single cell transcriptome sequencing starts to becoming a research hotspot in 2008, gently increases and remains steady last year. Chromatin structural genomics and haplotype analysis, on the other hand, shares much smoother curves in last ten years, indicated fundamental status of them in genomics industry. All these figures indicates that structural genomics embraces a significant milestone after the accomplishment of Human Genome Project in May 2006, and succeeding genomics study starts to focus on intragenomic phenomena afterwards: heterosis, epistasis, pleiotropy, other interactions within genome and their roles in etiopathology. Functional genomics is welcoming great development opportunity today to give more clue to personal medicine and intelligent medical treatment, which would revolutionize our healthy care industry in the long run, and corresponding technologies will similarly be extensively applied over the coming decades.

This article only discusses the most widely studied genetics technology in English language. Thus far, we can summarize and demonstrate industry hotspot analysis report of all biomedical fields in both languages on the biomedical technology analysis platform website. Integrality of Chinese library and more accurate technical classification for English library still require further efforts. This work would lay a foundation for establishing a reasonable structured knowledge-based system in biomedical realm.

References

1. Kuru, K., M. Niranjan, Y. Tunca, E. Osvank, and T. Azim. 'Biomedical Visual Data Analysis to Build an Intelligent Diagnostic Decision Support System in Medical Genetics', Artificial Intelligence in Medicine Vol. 62, No. 2, 105-118, 2014.
2. Peek, N., and S. Swift. 'Intelligent Data Analysis for Knowledge Discovery, Patient Monitoring and Quality Assessment', Methods of Information in Medicine Vol. 51, No. 4, 318-322, 2012.
3. Torralba-Rodrigueza, F. J., V. Bixquert-Montagud, J. T. Fernandez-Breis, and R. Martinez-Bejar. 'An Incremental Knowledge Acquisition-Based System for Supporting Decisions in Biomedical Domains', Computer Methods and Programs in Biomedicine Vol. 98, No. 2, 161-171, 2010.
4. Gonzalez-Velez, H., M. Mier, M. Julia-Sape, T. N. Arvanitis, J. M. Garcia-Gomez, M. Robles, P. H. Lewis, S. Dasmahapatra, D. Dupplaw, A. Peet, C. Arus, B. Celda, S. Van Huffel, and M. Lluch-Ariet. 'Healthagents: Distributed Multi-Agent Brain Tumor Diagnosis and Prognosis', Applied Intelligence Vol. 30, No. 3, 191-202, 2009.
5. Cain, T. J., R. L. Rodman, F. Sanfilippo, and S. M. Kroll. 'Managing Knowledge and Technology to Foster Innovation at the Ohio State University Medical Center', Academic Medicine Vol. 80, No. 11, 1026-1031, 2005.
6. Chen, P., and R. Verma. 'A Query-Based Medical Information Summarization System Using Ontology Knowledge', 19th Ieee International Symposium on Computer-Based Medical Systems, Proceedings, 37-42, 2006.

7. Montani, S. 'How to Use Contextual Knowledge in Medical Case-Based Reasoning Systems: A Survey on Very Recent Trends', Artificial Intelligence in Medicine Vol. 51, No. 2, 125-131, 2011.
8. Ranky, P. G. 'A 3d Web-Enabled, Case-Based Learning Architecture and Knowledge Documentation Method for Engineering, Information Technology, Management, and Medical Science/Biomedical Engineering', International Journal of Computer Integrated Manufacturing Vol. 16, No. 4-5, 346-356, 2003.