Master's Thesis

A Gene-Centric Perspective of Scientific and Technological Innovations

Woochul Jung

Department of Biomedical Engineering

Ulsan National Institute of Science and Technology

2023

A Gene-Centric Perspective of Scientific and Technological Innovations

Woochul Jung

Department of Biomedical Engineering

Ulsan National Institute of Science and Technology

A Gene-Centric Perspective of Scientific and Technological Innovations

A thesis submitted to
Ulsan National Institute of Science and Technology
in partial fulfillment of the
requirements for the degree of
Master of Science

Woochul Jung

12.15.2022 of submission

Approved by

Advisor
Cheol-Min Ghim

A Gene-Centric Perspective of Scientific and Technological Innovations

Woochul Jung

This certifies that the thesis of Woochul Jung is approved.

12.15.2022 of submission

Signature

Advisor: Cheol-Min Ghim

Signature

Taejoon Kwon

Signature

Semin Lee

Abstract

Research on genes and gene products is a foundation of modern biotechnology, and recognized for its applicability in medicine, agriculture, food industry, energy supply, environmental remediation, and many others. To investigate a macroscopic and gene-centric perspective of scientific discoveries and technological innovations, we employed a large-scale curation of research papers and patents. As a raw data to represent the scientific impact of each gene, we collected the entire set of research articles available on PubMed, that have the names of genes or gene products in their title or abstract. The more dedicated literature source, the United States Patent and Trademark Office (USPTO) patent publication, was retrieved as a source data to represent the counterpart in technological innovations. In parallel with this literature curation, the gene symbols were collected from curated subset of UniProt consortium database maintained by European Bioinformatics Institute, and then clustered into the non-overlapping standardized sets to eliminate the overwhelming duplicates and possible contamination by common acronyms. Based on the annual counts of papers or patents whose titles/abstracts include a given gene, we show the overall trends of genetic research since the launch of the Human Genome Project. The gene citation fluctuated more in inventive activity compared to those in the research, while both sides were largely contributed by medicinal discipline. The volume of publications mentioning genes has been increased while the debut of new genes on titles and abstracts has been deflated. In contrast, new combinations of previously-studied genes kept actively explored, and their frequently adopted genes informed biotechnology innovations rather than sheerly top-studied genes.

Contents

List of Figures

List of Tables

I Introduction

Advent of bioinfomatics and text mining in biology

The nucleotides, building blocks of the digitized genetic information, were discovered decades ahead of the idea of 'Central Dogma' [1]. It had taken time until the molecular structure of the information carrier was elucidated, which turned out to be the stable sequence of double helix [2, 3]. By the time the sequence of tRNA was reported from the experiment of Holley, the mapping between the genetic codes and the amino acid sequence was decoded [4, 5]. The first computer program for peptide sequencing already had been introduced by Dayhoff *et al.* in the beginning of the 1960s [6].

Since these priming works, the volume of digitized information of biological molecules has been inflated. The electronic repositories for such biological data was established in respond to growing demands, as well as the online libraries for the academic articles [7–10]. Further development of the lab instruments for reading biological information [11–13] and internet accessibility accelerated the deposit of generated data by researchers. The Human Genome Project (HGP) can be seen as a profound achievement accompanied by expansion of the bioinfomatics [14].

The text mining approach in the biological context has been suggested on this background since early 1990s [15, 16]. The main interest of related works has been mining of meaningful relations between genes and proteins [17], and even predicting interactions potentially justifiable by experiments. These co-occurrence mining of genes in literature has been encouraged by the functional genomics and high-throughput analysis experiments, such as DNA microarray [18, 19]. One of the well-known online platform is 'Textpresso', which provides mining result of selected species based on full text search of the literature set [20–22]. More recent example is the mining of pathological and disease-related concepts mentioned in the literature as well as genes or proteins [23–25].

Bibliometric research and network approach

Another point of view on the text mining of genes was bibliometric approach. This approach for the scientific literature was started as the name of 'Citation Index' by Eugene Garfield, of which purpose was to providing fast and accurate system for retrieving preceding works of academic papers [26]. From this well-established custom, citation dynamics in academia has become a subject of investigation by researchers. More specifically, the citing relations between scientific literatures had been visualized in the graph map, even before the network scientists delve into the bibliometric investigation [27, 28]. With utilization of bulk data through internet and rising of network science, numerous investigation on the scientific activity has been published [29–32]. Not only as a crucial tool for assessing the impact of literature and authors, bibliometric data itself became a main focus of investigation as a part of 'science of science' researches [33].

There are a few works considering the gene or protein as an object in the citation dynamics. Hoffmann and Valencia reported that same aged genes shows power-law behavior in the probability distribution of citation popularity for the yeast data [34]. From their work, no correlation was detected between

such popularity and interaction degree of their corresponding nodes in the protein-protein network. In the work of Pfeiffer and Hoffmann, they claimed that correlation exists between the citation of genes and the impact factor of their venue journals [35].

In some of those works, the genes were considered as components of network structure. even before the sweeping discoveries regarding complex network [18, 36]. Another approach from network science can be found regarding 'protein-protein interaction' (PPI) data, which predicted tentative functional modules within proteome [37–39]. The 'sequence similarity network' (SSN) is another example connecting proteins from various species to capture homology and relatedness within network view [40, 41]

Comprehensive view on gene bibliometrics

As far as found from the literature search, no comprehensive view on literature citation has been reported on the collective set of genes from varying species. One of the reasons would be overlapping gene homologs and their symbols among wide range of species, from which ambiguity may arise. A lead-in solution for the problem was suggested in this work, which borrows the community detection method from network science [42]. By removing overlapping components between the gene communities, the landscape on thematic categorization of genes could be presented. The time series of popularity were retrieved for each gene community by counting the articles that mention one of the corresponding gene symbols. The detected gene communities were denoted simply as 'gene', and the association between article and gene as 'hit' for the convenience throughout this work. The set of analysis on scientific literature were conducted on the patent document as well, which are more closely related to utilization of associated genes. From the observation on both sides, the correlation between research and inventive activity had been explored in gene-centric view. In the last part, the discoveries of genes in literature were investigated in terms of newly spotlighted genes on the titles or abstract. The trend over 30 years of genetic research since initiation of HGP could be explained by its volume and the repertoire of combinatoric research.

II Results

2.1 The landscape of gene popularity in science and technology

The number of hit research publications overall has been increasing. In the Figure 1, the blue line shows the number of hit articles over the years between 1990 and 2018. There is a small dip in year 2017, before which monotonically rising trajectory was maintained. In the panel **a**, the 'IL-6', one of the most popular genes in terms of hit research works is shown. The time series of 'IL-6' is remarkable for its almost identical trend with that of overall research hit (panel **c**), including the dip in year 2017. These trajectories are well aligned with the trend of total volume of research publications collected regardless of whether hit exist. In the Figure 2, the dip in 2017 is not visible for research articles (solid line) but increasing trend is maintained in both cases of selected journals (panel **a**) and entire MEDLINE citation of from PubMed (panel **b**).

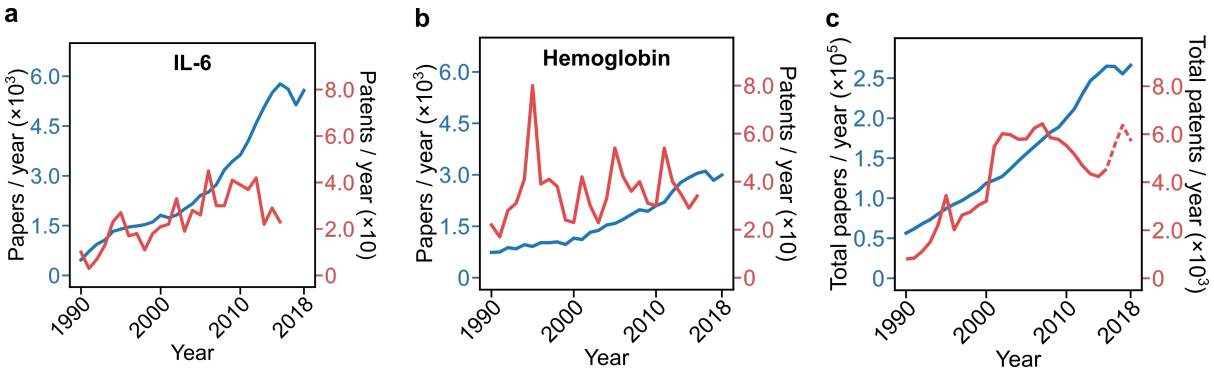

Figure 1: Annually counted hit articles from research and invention
The numbers of articles were counted if symbols of genes were mentioned in their titles or abstracts. The blue lines show the hit numbers obtained from research articles, while the red lines show those from patent articles. **a**, IL-6 was the most cited gene in research side over 1990–2018. **b**, Hemoglobin was the most cited gene in patent side over 1990—2015. **c**, The total hit articles were counted regardless of its cited gene.

The counterparts of patent publication show different patterns, more fluctuating and deviated from the ever-increasing trend. In the panel **c** of the Figure 1, the number of patent documents containing gene hits plunges after the year 2000 (red lines). This pattern can be seen in the background volume of patent most popular gene products in terms of patent in the Figure 2 (dotted line). This is mainly due to a radical change in the regulation of the patent publication procedure of 'United States Patent and Trademark Office' (USPTO). USPTO had published only granted patents by the policy, which they changed that any plausible patents be published after 1 and a half year from its filing date. In the panel **b** of the Figure 1, 'Hemoglobin', the most popular gene products in terms of patent hit in the scope of the time (1990—2015) is shown. The trend of hemoglobin shows highly fluctuating pattern, which is not aligned with that of overall hit patents. This is presumably because of the overall popularity of genes is much smaller in inventive activities compared to counterpart of research.

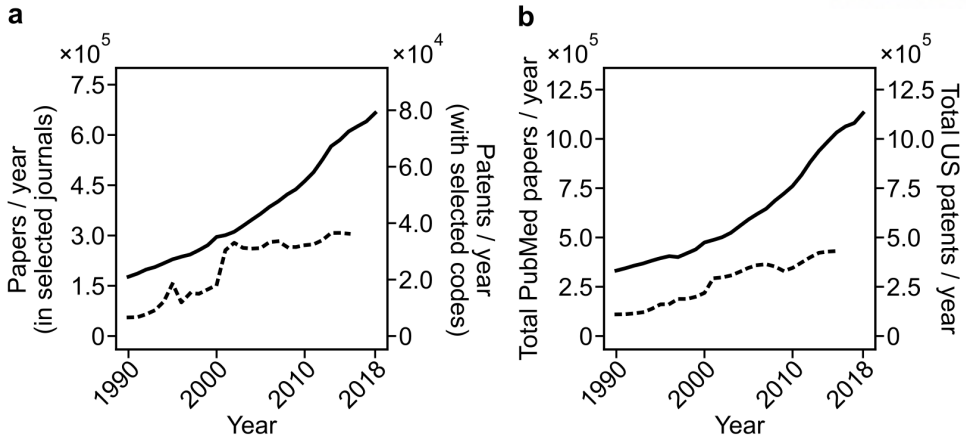

Figure 2: Annual counts of the articles from background set of literature
a, The numbers of articles after selection from the bulk resources. The solid line corresponds to research articles published in selected journals, while the dotted line corresponds to the patent publication containing selected classification codes. **b**, The number of articles obtained directly from the bulk resources. The solid line corresponds to research articles in PubMed, while the dotted line corresponds to the patent publication of USPTO.

From the list of top ranked genes, the difference between the two sides can be observed. In the panel **a** of the Figure 3, the consistency of ranked items observed in terms of their popularity between recent (2014—2018) and prolonged (1990—2018) time scope. However, in the panel **b**, much greater portion of genes in the list show changes of their ranks between the two time scope (1990—2015 for left, 2011—2015 for right side). The small volume of patent documents can be considered again, as one of the causes of this relative instability. The lists in both sides are dominated by genes with human-related homologs. Two exceptions among 40 items (22 genes) in the lists are 'Green Fluorescent Protein' (GFP) and 'Luciferase'. They are well-known genes as the genetic tools widely used in the laboratories of modern biology.

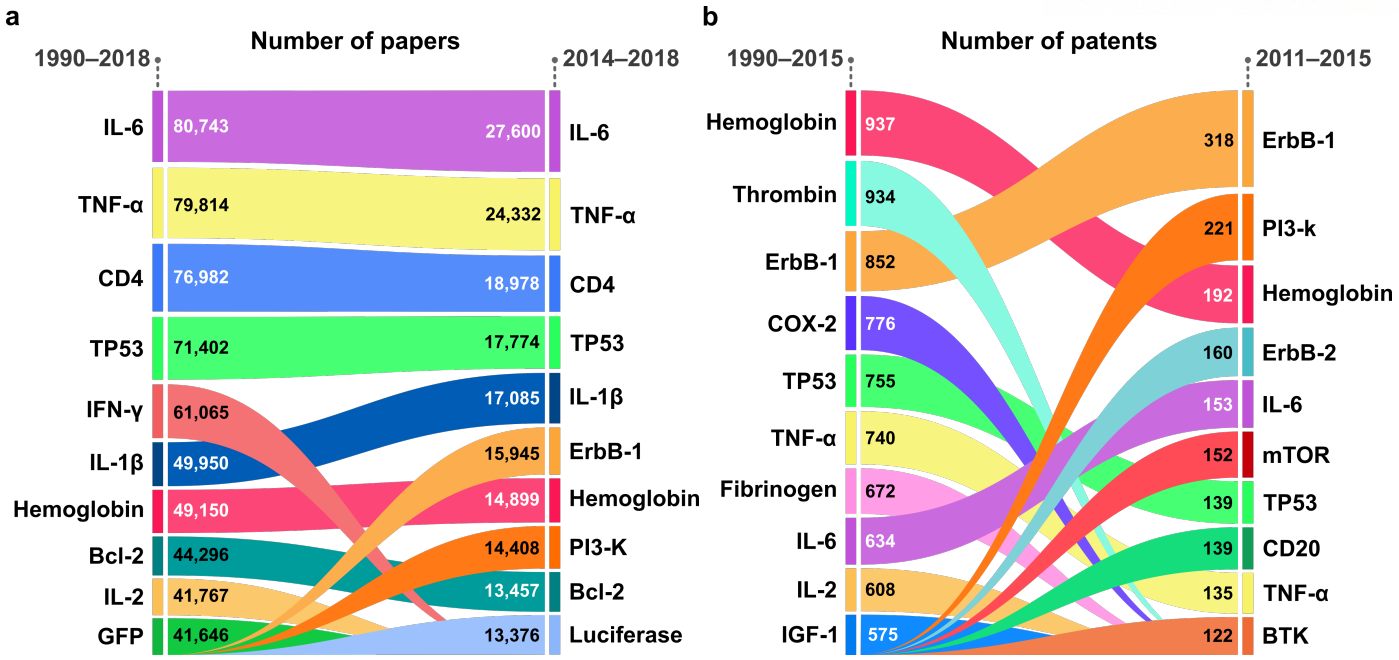

Figure 3: Rank changes in top 10 genes of prolonged and recent years
The 10 most-cited genes in the research and patent sides are listed. The thickness of lateral ribbons is proportional to respective count of the genes within each time period. **a**, The changes in the top 10 genes cited in research. The time periods of 1990–2018 and 2014–2018 were compared. **b**, The changes in the top 10 genes cited in patent publications. The time periods of 1990–2015 and 2011–2015 were compared.

The popularity of genes in research and invention sides are correlated. Total 13,065 genes are scattered on x, y plane in the Figure 4, of which hits are non-zero for both paper and patent. The x axis is for the number of research hit in log scale of each gene in the both panel **a**, and **b**. The y axis in the panel **a** is for the number of invention hit in log scale. Here, the Pearson and Spearman correlation between x and y values were calculated as 0.7866 and 0.6619 for each. The gray line is a trend guide line, which denotes that 98% of the dots are place below or on the line. In the panel **b** y values are patent-to-paper ratio of each gene. The red dots are the median values of y among the data points in the bins of logarithmically increasing size on the x axis. The number of patent hit for a genes seems to have upper limit corresponding to its number of research hits.

2.2 Thematic categorization of genes and gene products

The categories for the genes were assigned based on the information acquired from the literature items citing them and associated protein database entries. Mainly the journals that publishing the hit articles were used, of which theme and scope quantifies the character of the cited genes. The result is shown in Figure 5, which includes 61,874 genes in the pie chart. The categories are ‘Medical science’, ‘Plant science’, ‘Applied biotechnology’, ‘General microbiology’, and ‘Genetic tools’. There are subcategories within the medical science and applied biotechnology and the ‘Others’ is for genes that belong to nowhere. The category of medical science is dominating, of which number of genes are comparable with that of the protein coding genes in the human genome. There might be a claim that medical science

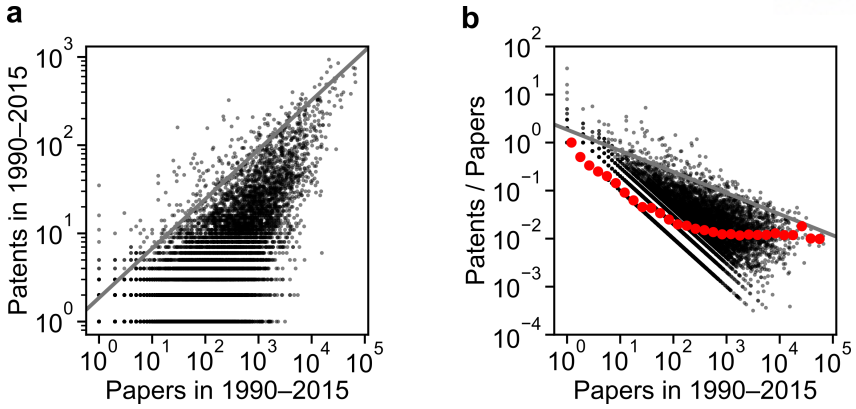

Figure 4: Scatter plot between the hit count of papers and patents
Each gene was plotted as a single dot with its x value as number of hit papers and y value as number of hit patents. Genes were included if they have non-zero values for both paper and patent citation over 1990–2015. **a**, The double-log scale view of the plot. A gray straight line indicates $y = \alpha x^{\lambda}$ with $\lambda = 0.56$, and $\alpha = 1.86$. 98% of the genes plotted are located below or on this line. **b**, The double-log scale view where the patent hits on y axis are divided by paper hits. Red dots locate median among the y values of dots in each interval of paper hit.

is over-represented due to biased selection of background literature sources. It cannot be denied that the signal related to medicine and human-related gene symbol are more prevailing than those of other themes in most of the steps of this investigation.

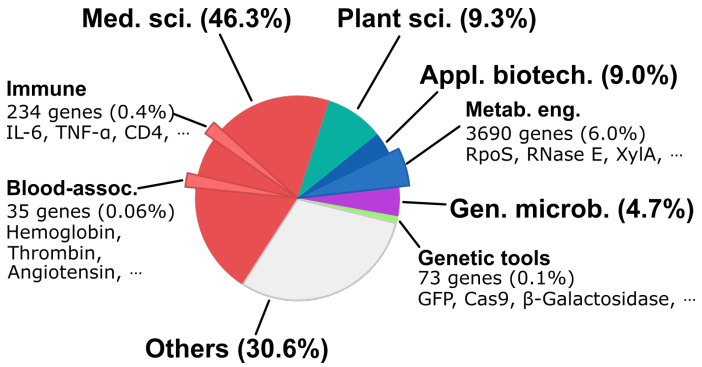

Figure 5: Fraction of thematic categories of genes
Proportion and example genes of each thematic gene category were presented. 61,874 genes were included in the chart. 'Med. sci.', is an abbreviation for medical science, 'blood-assoc.' for blood-associated, 'plant sci.' for plant science, 'appl. biotech.' for applied biotechnology, 'metab. eng.' for metabolic engineering, and 'gen. microb.' for general microbiology. The categories are based on articles over the entire time period available.

The genes were further classified by the patterns from the trajectory of annual hits. The existence of peak year for the popularity along the lifetime of genes were observed. Then behavior after the peak was considered to decide which pattern the gene should be assigned. The genes were classified into climbing (C), growth-slowed (G), sustained (S), falloff (F), or rebound (R) based on their trajectories of hits on research papers. The first two patterns were considered as growing patterns, while later three as non-growing patterns. From the basically the same approach, those for hits on patent publication were

classified into growing (G) or non-growing (N). In the panel **a** of Figure 6, distribution of these patterns are shown for the sets of genes from selected categories. The proportion of growing patterns are largest in the 'Immune' category of which genes were specifically selected from narrowed curation. Based on the patterns in the inventive activities, the category of metabolic engineering shows largest fraction of non-growing patterns (S+F+R).

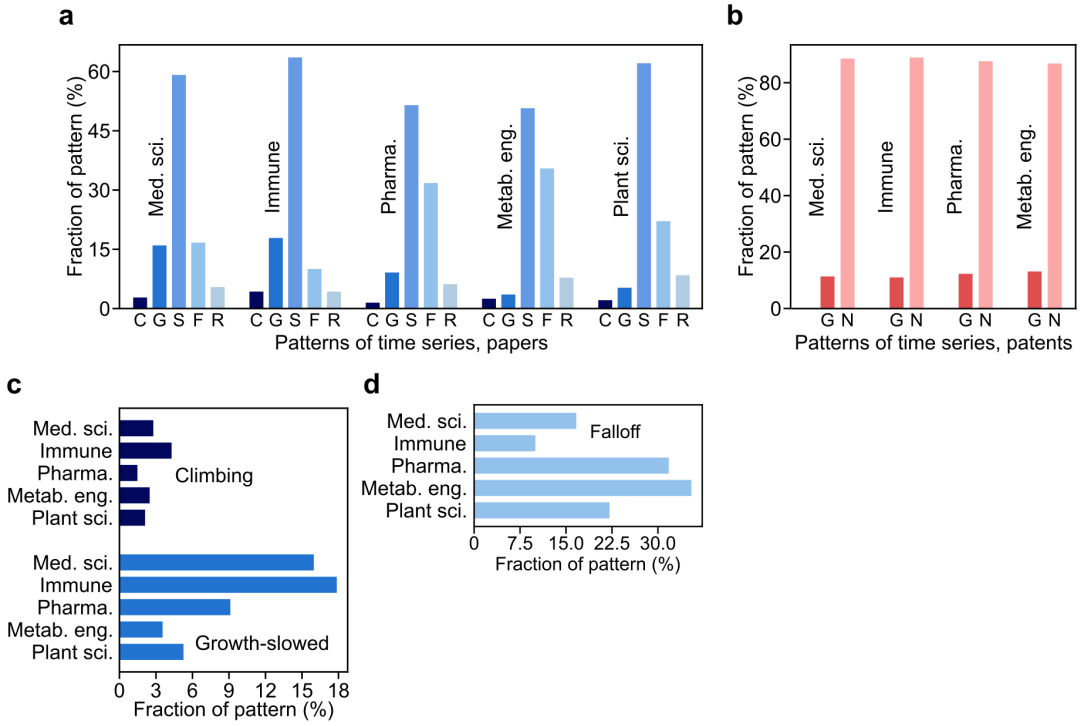

Figure 6: Fraction of time-series patterns within each gene category
Medical science, immune, pharmaceutical, metabolic engineering, and plant science were selected for comparison. **a**, Fractions of time-series patterns in research side among the genes in each category. **b**, Fractions of time-series patterns in patents among the genes in each category. **c**, Fractions of growing patterns within each category were directly compared for the research side. **d**, Fractions of 'falloff' pattern within each category were directly compared for the research side.

2.3 New combination as the driving force of research and inventive activity

Debut of genes in the title and abstract of documents is decreasing while that of unprecedented gene combinations is increasing. In the Figure 7, the trend of newly-studied genes is shown in the panel **a** when observed as the singlets. The overall trend is decreasing past the year 2000, when HGP was on its last stage. This trend is observed in both for papers (left) and patents (middle), though the trend has been more drastically plunging in the patent side. In contrast, in panel **b**, the decreasing pattern disappears from both side when focusing on the gene combinations. These trend shows remarkable similarity with that of the total hit documents in both sides of research and invention, which are shown in the panel **c** of Figure 1.

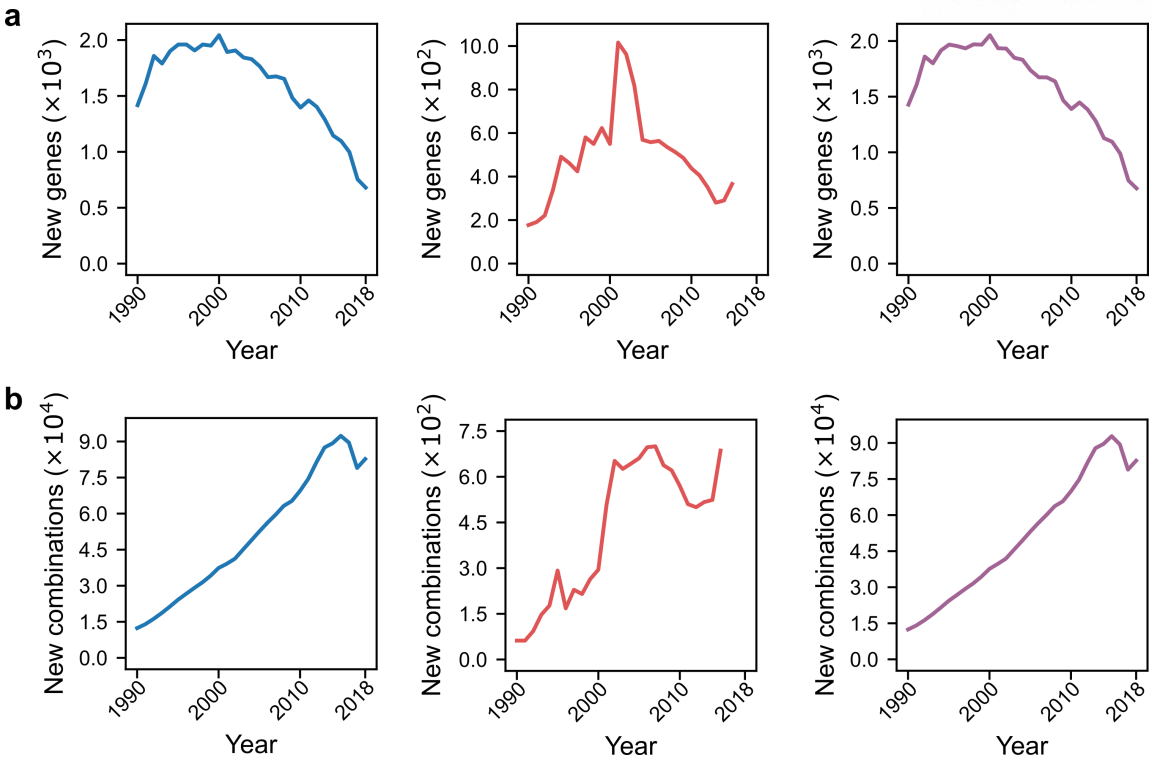

Figure 7: Annual counts of the debut of genes in the title/abstract of papers and patents
The number of genes (or gene combinations) which were mentioned in a title or abstract of article for the first time were counted for each year. Blue line corresponds to debuts on research, while red to those on patents, and purple to those on either side. **a**, The trend of debuts counted for a single gene. **b**, The trend of debuts counted for combinations of genes.

By classifying the gene debuts into theme categories, it shows differences in time trajectory of fraction between research and invention. For instance, In the panel **a** of Figure 8, the fraction of medical science among newly-focused genes has been decreasing in the paper side (blue line). This trend has been opposite for those in the patent side (red line), showing steady expansion of the discovery of newly focused genes in the category. The trend in panel **b** from both paper and patent side shows new discoveries have been expanding until recent years in the discipline of plant science. For the case of plant science, there is widely-known technological issue in the sequencing of plant genome. In case of applied biotechnology, the decrease of the portion in academic gene debuts is more dramatic (panel **c**).

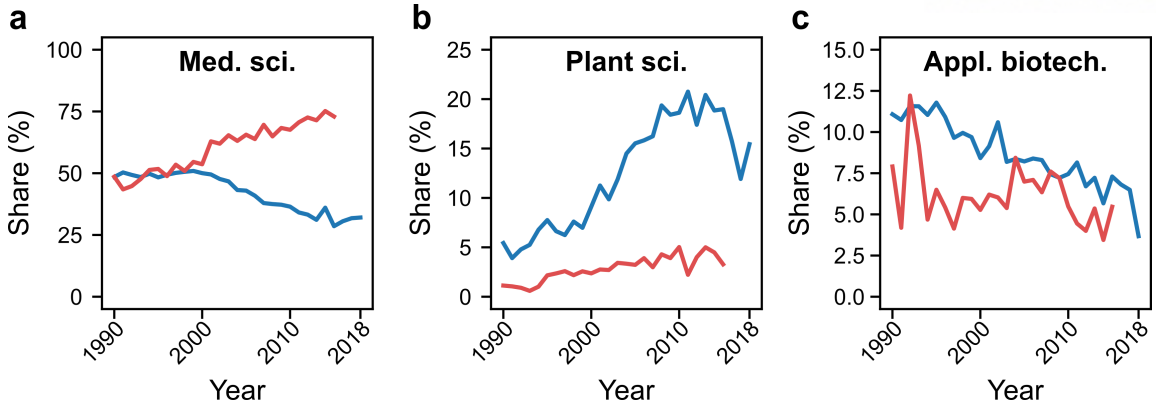

Figure 8: Fraction of gene categories among annually counted debut
Medical science, plant science, and applied biotechnology were selected for comparison. Blue lines correspond to fraction of each category among newly cited genes in research articles. Red lines correspond to those in patent documents.

More findings were derived from the new combinations of genes that were mentioned together in titles or abstracts of literature. The genes that most frequently included in these new combinations were surveyed annually in the Figure 9. The yellow band represent the period that the respective gene took the first place among others. β-galactosidase, GFP and CRISPR associated protein 9 (Cas9) were the most referred genes consecutively, except the mammalian target of rapamycin (mTOR, panel **c**) in 2015. These 3 genes are tool genes, so to speak, which were highly cited for their utility in the wet-lab experiments. They were frequently recruited by experimental protocols facilitating desired modification of model systems.

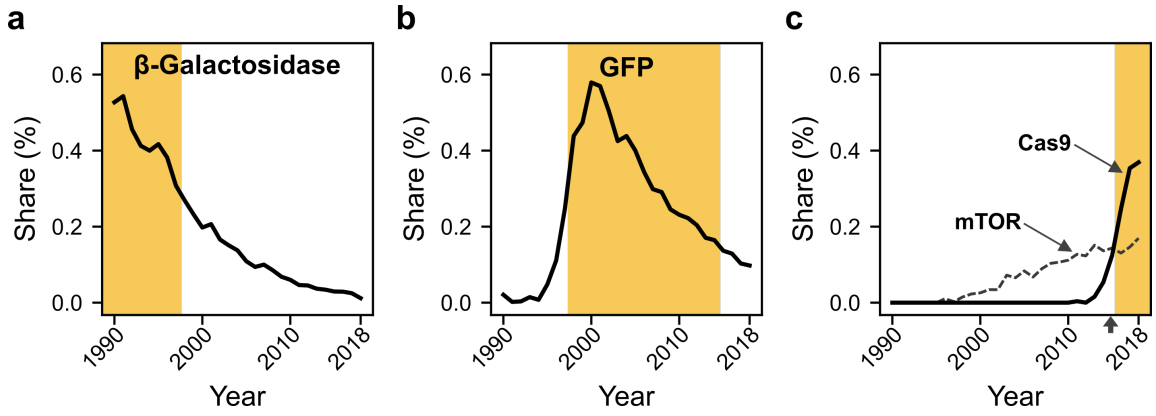

Figure 9: Fraction of gene pairs hiring selected genes among annual debuts on research
Paired debuts of genes on research were surveyed for their member genes, and the most hired genes were selected for each year. Yellow bands correspond to periods in which respective genes were most hired by paired debuts. **a**, β-galactosidase in 1990–1997. **b**, GFP in 1998–2014. **c**, Mammalian target of rapamycin (mTOR) in 2015, and CRISPR associated protein 9 (Cas9) in 2016–2018.

III Methods

3.1 Establishing connection between gene-protein and literature data

Construction of gene-protein synonym list from the UniProt data

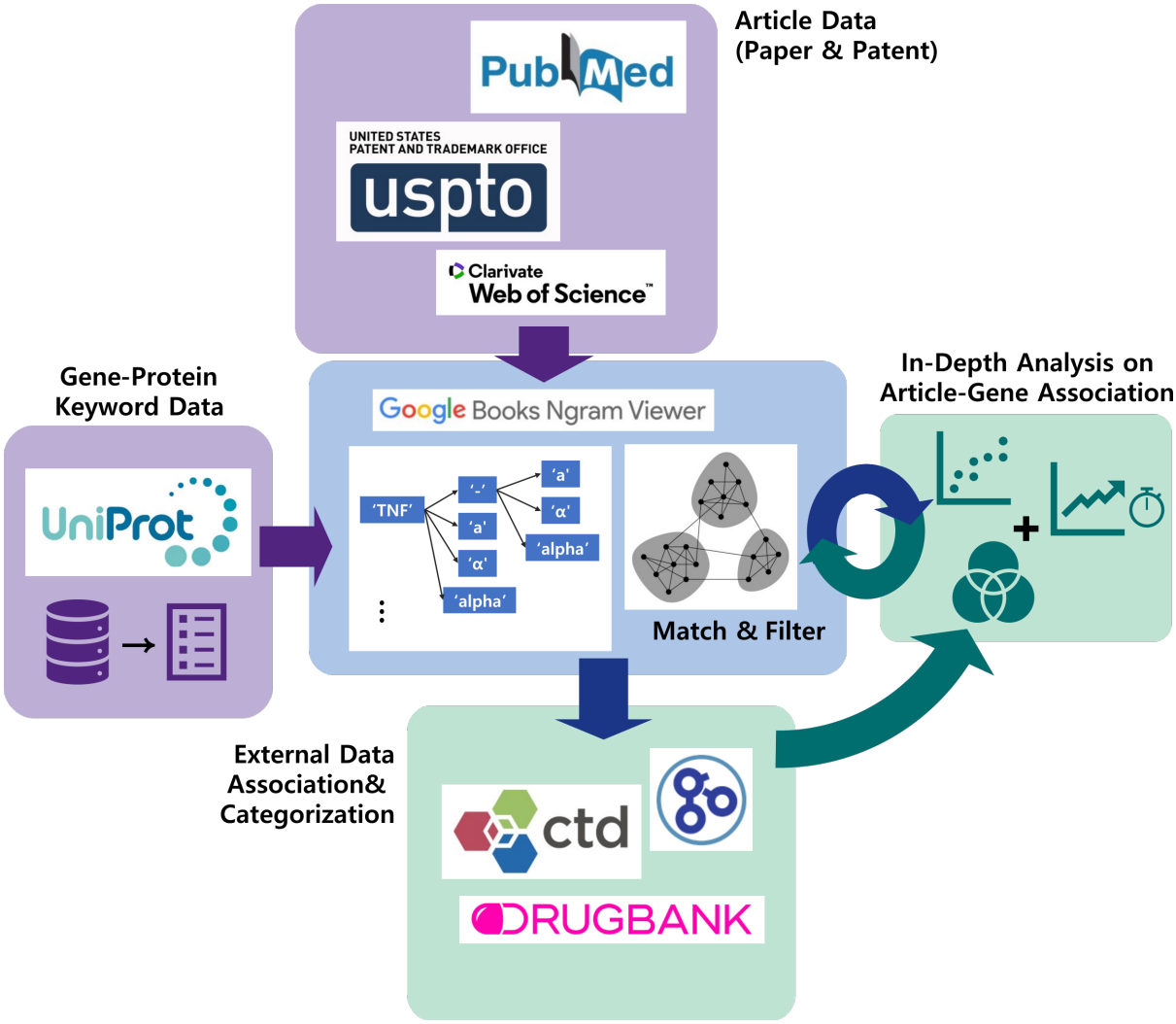

Figure 10: Schematic view on research work flows
Purple rectangles denote steps for the acquisition of raw resources. The blue rectangle represents the production step of gene citation data. Green rectangles show the analysis of data and the linkage to external data.

The protein and gene symbols were extracted from the protein entries in the Swiss-Prot database provided by UniProt Consortium [43]. The raw data was download from the source as a flat text file, of which version was '2019_01' released on 16th, January 2019. Reviewed protein entries within the Swiss-prot has protein description and gene name fields. The string information contained in both fields was considered as a source of synonymous gene symbols which were used in the text search. The synonyms improperly tagged were excluded, for example, those indicating subdomains of polypeptide or cleaved chains of main peptide body.

Preparing USPTO patent data

The scope of patent data was restricted to those published by 'United States Patent and Trademark Office' (USPTO). Patent data were downloaded from the public databases serviced by Google Cloud Platform (GCP). The US patent data provide by IFI CLAIMS Patent Services can be accessed on the 'Big Query' service in GCP. The full set of patent data was downloaded on the September 20th, 2020. The articles published by USPTO undergoes classification process and is assigned code items describing the field of technology of industry it relates.

The patents were primarily selected based their on classification code and relevance to genes or proteins. USPC was mainly used on the patents published by US office. At the same time 'International Patent Classification' (IPC) was used for patent applied to the international office of intellectual properties. In recent years, 'Cooperative Patent Classification' (CPC) is used by the European Patent Office as well as the office of United States. The gene-protein relevant patent documents were selected if they were assigned with one or more code items from the predefined classifications. The hierarchical tree and explanation of the items are serviced on the web site of CPC. The class items searched from tree were curated based on its relevance to the genes and protein entities. As the classification system itself is so vast that the scope of code items was restricted to specific levels, below the top nodes of classification sections.

Gene-relevant research articles

The academic literature were collected from PubMed based on the journals on which they are published. The relevant journals were selected based on the journal categories they belong. The categories based on the disciplines were provided in 'Web of Science' (WoS) 'Journal Citation Reports' (JCR) service. The list of journals was curated manually for each categories, and some of journals were filtered out which are seemingly irrelevant to our topic. The journal titles, abbreviations of the titles, and 8-digit ISSN numbers were collected from JCR published in 2020. Upon acquisition of the title information, the disciplines of titles related to biological science or biotechnology were selected. The baseline data file was downloaded from the PubMed FTP on May 15th, 2020, which was annual release containing the entire set of titles and abstracts of the articles serviced through the PubMed web by the end of 2019 [44]. The papers published from the journals of interest were collect by comparing the journal title, abbreviation and ISSN number obtained from WoS.

String-matching algorithms

The string-matching algorithms is based on the naïve string search which compares directly the phrase in the target literature and text string of gene-protein synonym data. The rules were added to catch prevailing variations of synonyms, including Greek letters and plural form as well as varying case-sensitivity. For the purpose of faster computation, the indexed tree and tokenizing scheme were incorporated in the matching algorithm [45,46]. On the top of this, tokenized variations were added to the synonym corpus, based on the predefined rules.

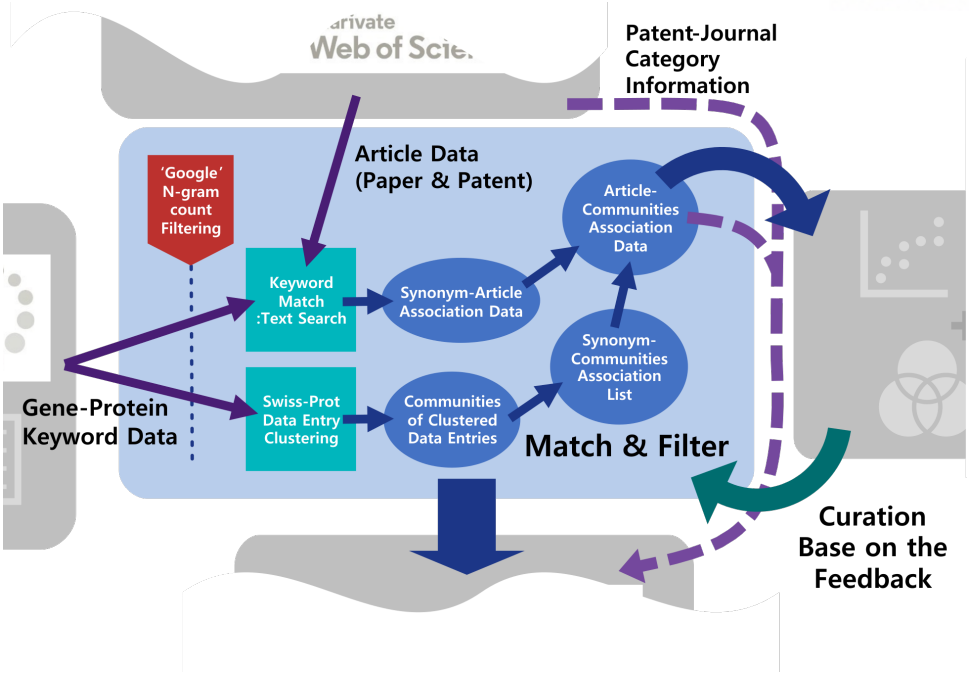

Figure 11: Schematic view on gene-text matching and filtering
The main components of this step are the gene-text matching and Swiss-Prot entry clustering. These two jobs are accomplished in a separate way to their last part respectively, then merged to produce gene-citation data.

Filtering general words from matched phrases using Google n-gram count

The significant portion of matches of most hit synonyms in test run were from the homonyms used in irrelevant contexts. The n-gram database was used to filter these irrelevant hits from the draft matching result. The usage counts of the hit phrases were extracted from Google Books n-gram database. They were considered as a proxy for its usage in the general context. The hit phrases of gene symbols were looked-up from the n-gram table for their usage count and sorted according to the value. A threshold for the usage count was determined to filter the hit phrased having counts that is equal to or more than the threshold. Then the phrases near the threshold are manually curated to determine whether to filter them or not. The Google n-gram raw data version 2 of July 2012 was used, and the download date was Mar. 4th, Oct. 28th and Sep. 7th of 2019 [47].

Clustering UniProt entries into functional entry

The schematic view on this step is presented in Figure 13. A part of synonyms of protein entries were duplicates in the Swiss-Prot. For instance, homologs from distinct tissues of single organism or those from organisms of distinct species. Most of those duplicate entries needed to be clustered into proper gene entities consistently. From this step, it becomes possible as well to obtain the collective set of genes across the varying species.

The network was constructed, of which nodes are Swiss-Prot entries and edges are similarity mea-

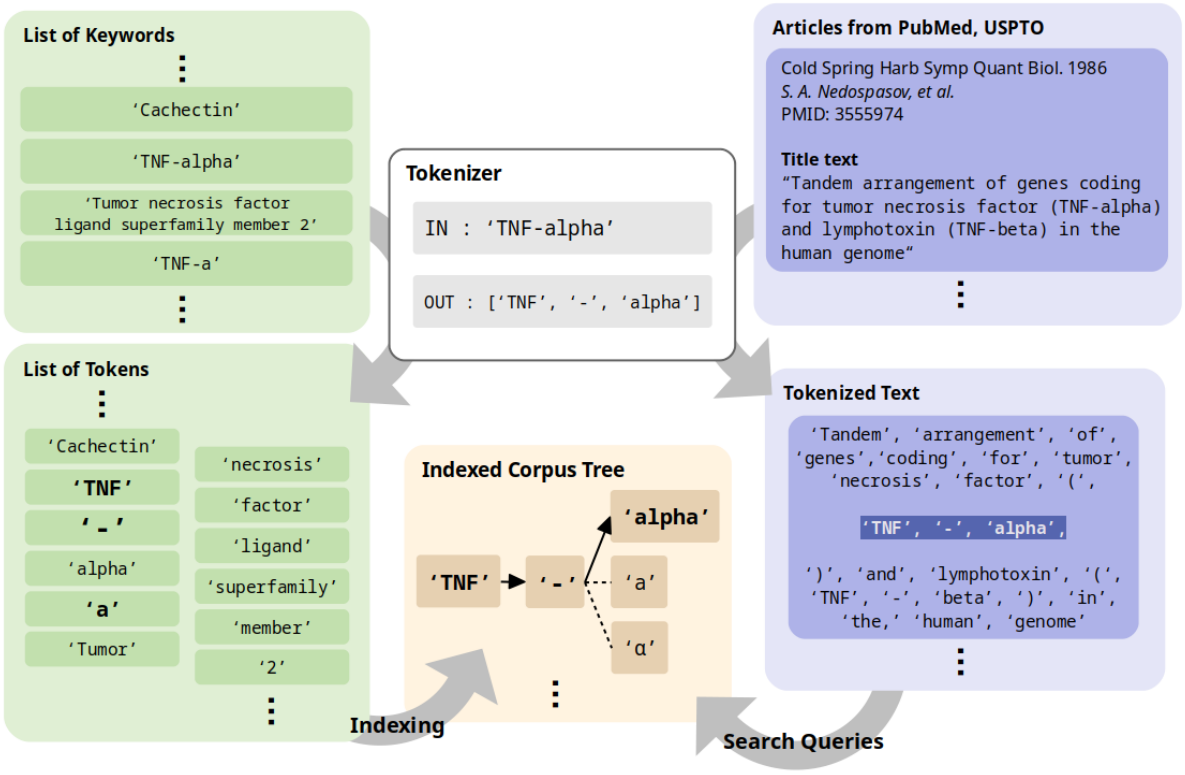

Figure 12: Schematic view on string-matching algorithm
Symbol keywords (examples on the left side) and the article texts (examples on the right side) are both tokenized in the same way. The resultant tokens from the gene symbols are indexed into a corpus tree which utilizes hash map to store the orders of tokens. The tokens obtained from article texts are queried to the corpus tree to find any correct sequential match on the indexed gene symbol.

sured between them. The 'Overlapping similarity' of set of synonyms of each entry was considered as the weight of links in-between [48]. The edges with weight lower than certain threshold were cut-off while others were uniformly converted to 1. The community detection was implemented on this unweighted version of the network [42].

Construction of annual ranking list of genes

All gene communities obtained from clustering step were assigned with non-overlapping synonym sets. When the titles and abstracts of article were searched for such synonym, the matches found were annotated as 'hits' between the gene community and articles. How many articles each gene community was cited by was counted, and the list of gene communities were sorted by their hit counts. Research article has its own publication date, which is generally same with the publication date of specific journal issue containing the article. Likewise, patent articles have their application date, and issued date and publication date in case of granted or published patent for each. The time trajectory of hit amount was obtained for each gene community, with time resolution of an annual basis. The ranked lists of win-

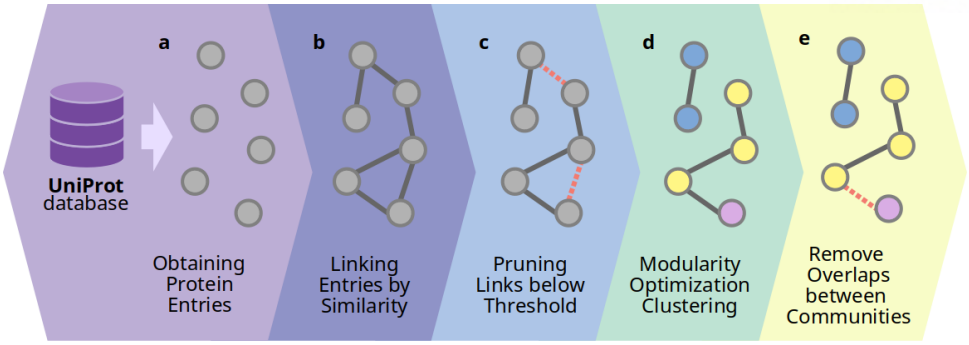

Figure 13: Flow diagram for gene community detection
a, A network structure between Swiss-Prot peptide entries was constructed. **b**, Edges are obtained from overlapping similarity between the gene symbol sets of two distinct entries. **c**, Edges with not enough weight are removed from the structure. Remaining edges are treated as unweighted. **d**, Modularity of the network was optimized from possible community allocations. **e**, The overlapping gene symbols were localized to respective communities selected by predefined rules.

dowed summation of such counts were reported with the controlled time scope, from 1990 to 2018 for paper and 1990 to 2015 for patent. The same lists from more recent time scope was reported base on the counts from the last 5 years of each time window.

3.2 Categorization and patterning from quantitative popularity of gene-protein

Patent valuation data

It is assumed that there is considerable relation between academic activities regarding a gene-protein and their industrial application. The quantifiable value of patent having matches referring to a certain gene or protein can be used as measure of potential or impact of industrial application they have. As there has been attempts to estimate such quantity in the fields of economy, the methodology was borrowed from the renowned research article on this topic [49].

The patent entries in the downloaded database could be indexed with the publication number of patent application. The valuation data shared by the Kogan *et al.*, 2017 targeted US patents published in the limited range of time, which was not enough for us to proceed analysis. The amount forward citation has been also used as a proxy for the impact or novelty of patent. The citation network has been tracked within the background Google patent data, using meta-data of each article which indicates identification number of other patents of future date referring it.

Categorization of paper, patents and associated gene communities

From the brief investigation on the popularity of genes, it was suspected that genes associated with certain group of disease tends to have specific form of trajectories on their hit time series. This gave a motivation for categorical analysis of genes based on their theme and disciplines.

The disciplines of each academic journals were already obtained from WoS which was used to set the boundary of the background literature pool. The categorical disciplines were manually modified, which

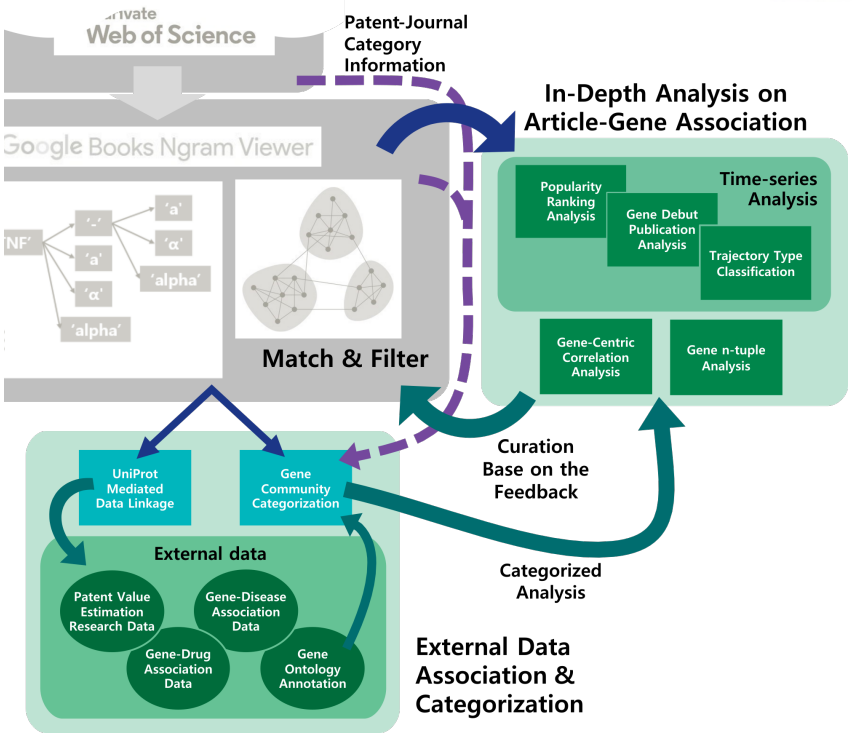

Figure 14: Schematic view on gene citation analysis and externally linked data
The gene-citation data produced were further analyzed mainly by approaches for time series data. A large portion was supported by categorization based on journal information and patent classification codes. Selected genes were further linked to external resources to specify the character of them. Cross-referencing information in UniProt mainly enabled linking and retrieving such items.

was curating them into more closely related groups, separating more theoretic or application-oriented journals, and removing overlaps between disciplines. The proportion of such category of paper hits was measured for a certain gene or protein. Those genes were classified into the designated category if the proportion exceeds threshold set for the category. Some of the classification codes (CPC) on the patent documents were used to group desired genes in the same manner with the case of academic journal articles. 'Gene Ontology' database offered proper ontology term that gene communities could be additionally curated for their additional characters, for example, those related to the 'immune response in organisms' term.

As the gene-disease information has been curated by other researchers and published on the web, detected gene communities were linked to the items the publication. The 'Comparative Toxicogenomic Database' (CTD) was considered as most suitable to our criteria after comparison with other DBs [50]. Disease items were curated in tree-like structure with hierarchical relation between levels. The top-level items of the tree in the disease classification were used to make subdivisions under the largest category of human medicine.

Large portion of medicinal purpose applications of genes and proteins results in development of drugs which affects their pathological expression inside living organs. Drug development is time-consuming and capital-intensive process, which means that popularity of a certain gene or protein as

target or carrier can be considered as one aspect of their impact on the industry. The most prominent candidate of database considered was 'DrugBank', of which data are partially serviced as free on the web [51]. Gene communities are linked to drug items through accession number of Swiss-Prot entries which are a member of each community.

3.3 Time series patterns and slowdown of gene debuts

From the time series of associations between genes and literature, a basic set of patterns were recognized regarding the trend of hit counts.

In the first step, additional time series generated from yearly summation of 3-year hit counts were prepared for each gene. Then peak years were defined as a year with the largest number of the value in the ad-hoc time series. In the cases with the peak year located on the latest year of valid time range, the patterns of corresponding genes were typed as 'Growing'. In the other cases, those were typed as 'Non-growing'. This step was applied to the time series from both sides of research and invention.

Regarding the time series from research publications, the growing patterns were subtyped into 'Climbing' and 'Growth-slowed', while the non-growing patterns into 'Sustained', 'Falloff', and 'Rebound'. Differences of summed counts between years were collected as another time series, along with the location of peak years. The 'climbing' and 'growth-slowed' patterns were distinguished based on the location of these peak years, in the same manner with the case of growing and non-growing patterns. On the other hand, the numbers of summed counts from peak, drop after peak, and the latest year were compared to distinguish 3 more subtypes within the non-growing patterns.

Journal categories	Journals listed in WoS		Journals found in PubMed		Articles found in PubMed	
	#	%	#	%	#	%
Biomedical science	1,488	38.35	1,456	40.75	6,516,614	45.31
Biomedical engineering	533	13.74	526	14.72	1,966,273	13.67
Biomed., pharmacy	262	6.75	251	7.02	959,262	6.67
Biomed., infective disease	104	2.68	104	2.91	416,000	2.89
Biomed., antibiotics	11	0.28	10	0.28	76,757	0.53
Biotechnology in general	35	0.90	34	0.95	93,162	0.65
Biotech., microbial	18	0.46	18	0.50	79,786	0.55
Biotech., food	16	0.41	14	0.39	30,644	0.21
Biotech., fuel	5	0.13	4	0.11	22,803	0.16
Biotech., nano technology	7	0.18	7	0.20	12,667	0.09
Biotech., environment	4	0.10	3	0.08	6,251	0.04
Biotech., biocontrol	19	0.49	10	0.28	5,694	0.04
Biotech., material	4	0.10	3	0.08	3,907	0.03
Biochemistry and cell biology	360	9.28	346	9.68	1,486,003	10.33
Veterinary science	107	2.76	84	2.35	272,511	1.89
Genetics	70	1.80	66	1.85	253,535	1.76
Microbial science	89	2.29	83	2.32	232,543	1.62
Proteogenomics and bioinfomatics	90	2.32	88	2.46	220,103	1.53
Plant science	193	4.97	121	3.39	196,139	1.36
Animal science	69	1.78	42	1.18	111,202	0.77
Entomology	85	2.19	48	1.34	37,339	0.26
Aquatic biology	97	2.50	60	1.68	32,742	0.23
Biomolelcuar design	6	0.15	6	0.17	12,062	0.08
Algae science	9	0.23	6	0.17	2,306	0.02
Multidisciplinary	65	1.68	56	1.57	1,013,558	7.05
N.E.C.	134	3.45	127	3.55	322,942	2.25
Total	3,880	100	3,573	100	14,382,805	100

Table 1: Summery of journals listed in JCR and those collected from PubMed
Journal list of Journal Citation Reports (JCR) were curated into intermediate list of categories which are enlisted in the table. The categories were assured to have no overlapping journals in between. The number above are obtained regardless of the publication date of articles. N.E.C. stands for 'not elsewhere classified'.

IV Conclusion and Discussion

Comparison of collective and individual gene citation

The trend of research citations referring top-ranked genes seems to align with the trend of entire gene-citing publications. This type of trend alignment is not observed for the citation of genes on the patent documents. Generally, the annual citation of a gene seems to fluctuate more on inventive activities. The population of gene-citing patents is relatively smaller than that of gene-citing research articles, which is suspected to contribute to the larger fluctuation. Once the main bodies of patent literature are included, the signal of gene-citing inventions might be amplified which can help clearer comparison of the aligned trends. Another reason can be gene-specific issues related to invention, especially for the highly patented genes. To address this, case studies on these issues are needed for the top-ranked genes on the patent side. More specifically, the contents of gene citing patents published near the time point of drastic increase of citation are the targets for this study. If a burstiness is one of the properties of highly patented technological factors, like genes in this work, the fluctuation would be more natural for the successful genes. Likewise, the gene-specific issues related to its research activity can be investigated to see whether the burstiness disappears for the highly-researched genes.

Dominance of the medical category

The dominance of medical science among categorized genes has been observed from different sides of this work. Throughout the investigation, the thematic categorization mainly depended on the journal venue where the gene-citing articles were published. The list of journal titles collected from the PubMed were seemingly not biased, as it contains most of the journals listed under the relevant JCR disciplines in WoS as shown in table 1. However, the sizes of the journal lists were much larger for the medicine-related disciplines. It means that potentially gene-citing research disciplines are highly populated by journal articles of medicinal discipline. Another explanation can be that the curated portion of UniProt database is intrinsically biased to gene products from more favored model organisms. The number of peptide entries originated from human, basically, takes up the small portion among the entire entries in Swiss-Prot ($\sim 3.6\%$). The fraction calculated from the number of gene symbols originated from human proteins is much larger ($\sim 12.5\%$). Recalculating this fraction by narrowing the numerator to the gene symbols ever cited, the human-related symbols take up much greater portion ($\sim 33.4\%$). Here, the bias exists to a certain extent toward the curation of human-related gene symbols in the UniProt. Above observations are straightforward upon considering the dominant size of research funding on the fields of biomedical and health science in US [52]. Still, it is hard to illustrate how the medical research contributed to the expansion of the knowledge related to genes and proteins.

Saturation of the gene discovery

The slowdown in the discovery of newly-cited genes was observed on the research publications. The genes cited in the patent publication first ever were negligible, largely outnumbered by those from the

research side. It is a natural consequence that the repertoire of genes available to discover would be limited, once the repertoire of the species ever discovered is limited. However, the volume of newly sequenced genome has been increased, even though not all the species ever discovered has been sequenced for their genome [53]. More importantly, it is not clear to what extent the genes ever sequenced would finally be investigated as main targets in the laboratory. The observed set of genes already cited is mainly from the human genome or related to their human homolog, which is almost saturated. In addition, the genes from other species than the human and vertebrates were investigated to a certain extent by means of the homologs or functionally parallel proteins of representative organisms in separate clades. The other genes, such as tentative coding sequences in 'TrEMBL', have no clear significance as research targets at present. The slowdown is conclusive from the perspectives of the preexisting research works and their motivations for the gene citation. At the same time, the capacity for further discovery remains in the unexplored part of sequence data, the genome obtained from untypical species.

Further investigation and improved methods

From the journal information linked to each genes, more extensive investigation is possible. For instance, the authors and their affiliation can be retrieved from database for researcher. The similar approach is possible as well in the patent document, as they contains information of inventors and assignees. It will be desirable to have access to well-curated database for researchers, institutes, or companies around the world.

Throughout this work, the text data were restricted to titles and abstract of literature, excluding main bodies such as results and methods. The insignificant mentions of genes or proteins within these excluded portion of the articles were ignored. One can argue that significance of the gene as research object is guaranteed from its mentions within abstract. This approach was taken for granted in several preceding works [24, 34, 35, 54].

However, the loss of 'citation signal' caused from the approach can be more severe regarding the genes with low popularity. The gene-to-article association data provided by 'National Library of Medicine' (NLM) were compared with the association data generated here. A draft result showed that still remarkable portion of gene-to-article associations curated from NLM were missing in the set of generated links. The loss of weak signals is suspected cause for this missing part. Another reason can be the protein entries outside the Swiss-Prot, which have associations with articles according to curation by NLM.

These false-negative issue can be further discussed regarding the text search method. The main focus of qualitiy control herein was removing false positive portion from the generated gene-to-article association data. The precision score of the result can be managed by this approach, while false negative issue from the limitation of dictionary-based search cannot be resolved. The gene symbols obtained from Swiss-Prot were the items of such dictionary, excluding those from un-curated portion of the UniProt database. There are more sophisticated approaches on this task which mainly borrow analytical routines within the scope of natural language processing (NLP) [55–57]. Most prominent example is 'named en-

tity recognition' (NER) which can recognize the nouns related to concepts in biology. The target mainly includes biological entities like genes, proteins, diseases and experimental specimens [58]. It can be implemented without predefined dictionary of such target keywords. The recent achievements in machine learning were applied for these NLP tasks and showed remarkable improvement in the biological context [59]. The major portion of genes with strong citation signals would not show much improvement in retrieved trends, even those cutting-edge techniques are implemented. Still it is recommended to survey available, easy-to-use, options of the NLP approach for more manageable quality control in these types of analysis.

The definition of the gene used through the series of analysis within this work is open to improvements. The clustered peptide entries obtained from community detection are somewhat analogous to the sets of homologous proteins. However, the identity between distinct peptides cannot easily be captured by comparison of gene-protein symbols denoting them. Moreover, there is the 'gray area' where term 'homologous' cannot be clearly defined for a pair of proteins.

One solution can be direct comparison of the sequence information between peptides, which is an already well-established method measuring relatedness of pairs of proteins [60]. Despite of its heavy computational burden, it can capture the similarity between proteins which is more directly related to their function. The 'sequence similarity network' (SSN) is an approach which can possibly substitute the network realization processed within this work [40, 41]. In addition to this, shared symbols need be assigned to multiple gene, which were not shared but rather localized to a single gene throughout this work. Many synonymous symbols capture ontological meaning from different level of the hierarchy, such as protein families or strain specific variants of protein. Once the hierarchy is reflected in the gene citation analysis, more comprehensive view on the landscape can be obtained.

References

[1] P. Levene, "The structure of yeast nucleic acid," *Journal of Biological Chemistry*, vol. 40, pp. 415–424, 12 1919.

[2] J. D. Watson and F. H. C. Crick, "Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid," *Nature*, vol. 171, pp. 737–738, 4 1953.

[3] R. E. Franklin and R. G. Gosling, "Molecular configuration in sodium thymonucleate," *Nature*, vol. 171, no. 4356, pp. 740–741, 1953.

[4] M. W. Nirenberg, O. W. Jones, P. Leder, B. F. C. Clark, W. S. Sly, and S. Pestka, "On the coding of genetic information," *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 28, pp. 549–557, 1 1963.

[5] R. W. Holley, G. A. Everett, J. T. Madison, and A. Zamir, "Nucleotide sequences in the yeast alanine transfer ribonucleic acid," *J Biol Chem*, vol. 240, no. 5, p. 2122, 1965.

[6] M. O. Dayhoff and R. S. Ledley, "Comprotein: a computer program to aid primary protein structure determination," in *Proceedings of the December 4-6, 1962, American Federation of Information Processing (AFIPS) '62 (Fall)*, 1962, pp. 262–274.

[7] D. Benson, D. J. Lipman, and J. Ostell, "Genbank," *Nucleic Acids Research*, vol. 21, pp. 2963–2965, 1993.

[8] G. N. Cameron, "The embl data library." *Nucleic acids research*, vol. 16, no. 5 Pt A, p. 1865, 1988.

[9] Protein Data Bank, "Protein data bank," *Nature New Biol*, vol. 233, p. 223, 1971.

[10] U. Poland, "MEDLARS online history," *Bulletin of the Medical Library Association*, vol. 78, no. 1, p. 72, 1990.

[11] K. B. Mullis, "The unusual origin of the polymerase chain reaction," *Scientific American*, vol. 262, no. 4, pp. 56–65, 1990.

[12] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, 2003.

[13] E. R. Mardis, "The impact of next-generation sequencing technology on genetics," *Trends in genetics*, vol. 24, no. 3, pp. 133–141, 2008.

[14] M. Olson, L. Hood, C. Cantor, and D. Botstein, "A common language for physical mapping of the human genome," *Science*, vol. 245, no. 4925, pp. 1434–1435, 1989.

[15] K. Baclawski, R. P. Futrelle, N. F. Noy, and M. J. Pescitelli, "Database techniques for biological materials and methods." in *ISMB*, 1993, pp. 21–28.

[16] K. Baclawski *et al.*, "Data/knowledge bases for biological papers and techniques," *Proc. Sympos. Adv. Data Management for the Scientist and Engineer*, pp. 23–28, 1993.

[17] T. Sekimizu, H. S. Park, and J. . I. Tsujii, "Identifying the interaction between genes and gene products based on frequently seen verbs in medline," 1998. [Online]. Available: https://www.researchgate.net/publication/12255953

[18] B. J. Stapley and G. Benoit, "Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in medline abstracts." WORLD SCIENTIFIC, 12 1999, pp. 529–540.

[19] T. K. Jenssen, A. Lægreid, J. Komorowski, and E. Hovig, "A literature network of human genes for high-throughput analysis of gene expression," *Nature Genetics*, vol. 28, 2001.

[20] H.-M. Müller, E. E. Kenny, and P. W. Sternberg, "Textpresso: An ontology-based information retrieval and extraction system for biological literature," *PLoS Biology*, vol. 2, p. e309, 9 2004.

[21] H.-M. Müller, A. Rangarajan, T. K. Teal, and P. W. Sternberg, "Textpresso for neuroscience: Searching the full text of thousands of neuroscience research papers," *Neuroinformatics*, vol. 6, pp. 195–204, 9 2008.

[22] H.-M. Müller, K. M. V. Auken, Y. Li, and P. W. Sternberg, "Textpresso central: a customizable platform for searching, text mining, viewing, and curating biomedical literature," *BMC Bioinformatics*, vol. 19, p. 94, 12 2018.

[23] À. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, and L. I. Furlong, "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research," *BMC bioinformatics*, vol. 16, no. 1, pp. 1–17, 2015.

[24] S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, and L. J. Jensen, "DISEASES: Text mining and data integration of disease–gene associations," *Methods*, vol. 74, pp. 83–89, 2015.

[25] M. Krallinger, F. Leitner, and A. Valencia, "Analysis of biological processes and diseases using text mining approaches," *Bioinformatics Methods in Clinical Research*, pp. 341–382, 2010.

[26] E. Garfield, "Citation indexes for science," *Science*, vol. 122, pp. 108–111, 7 1955.

[27] X. Lin, "Map displays for information retrieval," *Journal of the American Society for information Science*, vol. 48, no. 1, pp. 40–54, 1997.

[28] H. Small, "Visualizing science by citation mapping," *Journal of the American society for Information Science*, vol. 50, no. 9, pp. 799–813, 1999.

[29] F. Radicchi, S. Fortunato, and C. Castellano, "Universality of citation distributions: Toward an objective measure of scientific impact," *Proceedings of the National Academy of Sciences*, vol. 105, no. 45, pp. 17 268–17 272, 2008.

[30] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the national academy of sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[31] L. M. Bettencourt, D. I. Kaiser, and J. Kaur, "Scientific discovery and topological transitions in collaboration networks," *Journal of Informetrics*, vol. 3, no. 3, pp. 210–221, 2009.

[32] T. Kuhn, M. Perc, and D. Helbing, "Inheritance patterns in citation networks reveal scientific memes," *Physical Review X*, vol. 4, no. 4, p. 041036, 2014.

[33] S. Fortunato *et al.*, "Science of science," *Science*, vol. 359, no. 6379, p. eaao0185, 2018.

[34] R. Hoffmann and A. Valencia, "Life cycles of successful genes," *Trends in Genetics*, vol. 19, pp. 79–81, 2 2003.

[35] T. Pfeiffer and R. Hoffmann, "Temporal patterns of genes in scientific publications," *Proceedings of the National Academy of Sciences*, vol. 104, pp. 12 052–12 056, 7 2007.

[36] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[37] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[38] S.-H. Yook, Z. N. Oltvai, and A.-L. Barabási, "Functional and topological characterization of protein interaction networks," *Proteomics*, vol. 4, no. 4, pp. 928–942, 2004.

[39] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular systems biology*, vol. 3, no. 1, p. 88, 2007.

[40] H. J. Atkinson, J. H. Morris, T. E. Ferrin, and P. C. Babbitt, "Using sequence similarity networks for visualization of relationships across diverse protein superfamilies," *PloS one*, vol. 4, no. 2, p. e4345, 2009.

[41] A. J. Enright and C. A. Ouzounis, "BioLayout—an automatic graph layout algorithm for similarity visualization," *Bioinformatics*, vol. 17, no. 9, pp. 853–854, 2001.

[42] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, 10 2008.

[43] A. Bateman *et al.*, "UniProt: the universal protein knowledgebase in 2021," *Nucleic Acids Research*, vol. 49, 1 2021.

[44] R. Agarwala *et al.*, "Database resources of the national center for biotechnology information," *Nucleic Acids Research*, vol. 46, 1 2018.

[45] E. Fredkin, "Trie memory," *Communications of the ACM*, vol. 3, 9 1960.

[46] G. Barth, "An analytical comparison of two string searching algorithms," *Information Processing Letters*, vol. 18, 6 1984.

[47] Y. Lin, J.-B. Michel, E. A. Lieberman, J. Orwant, W. Brockman, and S. Petrov, "Syntactic annotations for the google books ngram corpus." Association for Computational Linguistics, 7 2012, pp. 169–174. [Online]. Available: https://www.aclweb.org/anthology/P12-3029

[48] V. M.K and K. K, "A survey on similarity measures in text mining," *Machine Learning and Applications: An International Journal*, vol. 3, 3 2016.

[49] L. Kogan, D. Papanikolaou, A. Seru, and N. Stoffman, "Technological innovation, resource allocation, and growth," *The Quarterly Journal of Economics*, vol. 132, 5 2017.

[50] A. P. Davis *et al.*, "Comparative Toxicogenomics Database (CTD): update 2021," *Nucleic Acids Research*, vol. 49, 1 2021.

[51] D. S. Wishart *et al.*, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Research*, vol. 46, 1 2018.

[52] National Center for Science and Engineering Statistics (NCSES). (accessed Dec. 12, 2022) "Higher Education Research and Development: fiscal year 2021". ncses.nsf.gov. [Online]. Available: https://ncses.nsf.gov/pubs/nsf23304/

[53] National Library of Medicine. (accessed Dec. 12, 2022) "GenBank and WGS statistics". ncbi.nlm.nih.gov. [Online]. Available: https://www.ncbi.nlm.nih.gov/genbank/statistics/

[54] A. M. Cohen, W. R. Hersh, C. Dubay, and K. Spackman, "Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts," *BMC bioinformatics*, vol. 6, no. 1, pp. 1–15, 2005.

[55] A. A. Morgan *et al.*, "Overview of BioCreative II gene normalization," *Genome Biology*, vol. 9, p. S3, 2008.

[56] L. Smith *et al.*, "Overview of BioCreative II gene mention recognition," *Genome Biology*, vol. 9, p. S2, 9 2008.

[57] Z. Lu *et al.*, "The gene normalization task in BioCreative III," *BMC Bioinformatics*, vol. 12, p. S2, 12 2011.

[58] S. Zhang and N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts," *Journal of biomedical informatics*, vol. 46, no. 6, pp. 1088–1098, 2013.

[59] J. Lee *et al.*, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, 9 2019.

[60] W. R. Pearson, "An introduction to sequence similarity ("homology") searching," *Current protocols in bioinformatics*, vol. 42, no. 1, pp. 3–1, 2013.

Acknowledgements

The results presented in this thesis were the part of research work suggested by Dr. Cheol-Min Ghim and Dr. Pan-Jun Kim. This project, including my M.S. course in UNIST, was thoroughly supported and advised on by Prof. Ghim. I consider myself much fortunate having got a chance to join Ghim Lab, where I could get a glimpse of joy in academia. Besides, there had been numerous remote conferences where Dr. Pan-Jun Kim gave detailed supervisions on the methodology. I hope he earn worthwhile return for his straight devotion. Before I participated full time in this project, Giju Jung conducted a preparatory investigation. He had left all the data and comments from his work before he moved. WooJoong Kim led many parts of the research and helped me to assure integrity of the program codes and result. He produced remarkable results on modeling as well which are out of scope of this thesis. Junghun Chae, Juneil Jang and Dr. Yong-Su Jin with his students helped to check and curate on the quality of produced data. Moreover, Junghun is now working on the update of overall story after me and WooJoong. I'd like to give thanks to all above and many others who supported this work and kindly shared their opinions.