Journal of Informetrics 9 (2015) 1–15

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Tracing database usage: Detecting main paths in database link networks

Qi Yu [a,*], Ying Ding [b], Min Song [c], Sungjeon Song [c], Jianhua Liu [d], Bin Zhang [e]

[a] School of Management, Shanxi Medical University, 56 Xinjian South Road, Taiyuan 030001, China
[b] Department of Information and Library Science, Indiana University, Bloomington, IN, USA
[c] Department of Library and Information Science, Yonsei University, Seoul, South Korea
[d] National Science Library, Chinese Academy of Sciences, Beijing, China
[e] Center for the Studies of Information Resources, Wuhan University, Wuhan, China

ARTICLE INFO

Article history:
Received 4 July 2014
Received in revised form
28 September 2014
Accepted 29 October 2014
Available online 20 November 2014

Keywords:
Database link network
Main path
Bibliometrics
Bioinformatrics

ABSTRACT

This paper presents a database link network to measure the impact of databases on biological research. To this end, we used the 20,861 full-text articles from PubMed Central in the field of Bioinformatics. We then extracted databases from the methodology sections of these articles and their references. The list of databases was built with *The 2013 Nucleic Acids Research Molecular Biology Database Collection* (available online), which includes 1512 databases. The database link network was constructed from sets of pairs of databases mentioned in the methodology sections of full-text PubMed Central articles. The edges of the database link network represent the link relationships between two databases. The weight of each edge is determined either by the link frequency of the two databases (i.e., in the link-weighted database link network) or the topic similarity between two databases (i.e., in the similarity-weighted database link network). With the database link network, we analyzed the topological structure and main paths of the database link network to trace the usage, connection, and evolution of databases. We also conducted content analysis by comparing content similarities among the papers citing databases.

1. Introduction

Biomedical data are being produced at an extraordinary rate (Luscombe, Greenbaum, & Gerstein, 2001; Reichhardt, 1999). Valuable aggregations of data (e.g., GO, SwissPro) have been shared online (Mons et al., 2011). These biological databases are an important tool in assisting biologists to hypothesize or understand biological phenomena, from biomolecule structure and interactions, to the metabolism of entire organisms, and to the evolution of species. Compared to fields like physics, astronomy, and computer science, which have been dealing with the challenges of massive databases for decades, the big-data revolution in biology has been sudden, allowing little time for researchers to adapt to it. Databases are difficult to find, and annotation and curation by the scientific community are increasing at a rate that is painfully slow (Mons et al., 2011). Biologists now find themselves unable to extract all they need from the large amount of available data. Therefore, it is worth considering how biologists might make more effective use of these databases.

* Corresponding author. Tel.: +86 0351 4135652.
E-mail address: yuqi351@gmail.com (Q. Yu).

http://dx.doi.org/10.1016/j.joi.2014.10.002
1751-1577/

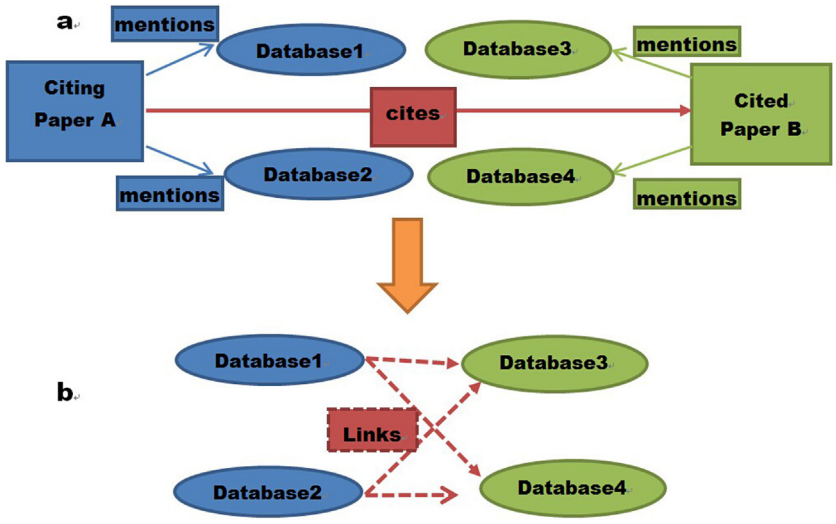

Fig. 1. Database link network.

Much work has been done to organize, categorize, and rate these databases, so that the information they contain can be most effectively exploited. An important resource for finding biological databases is a special yearly issue of the journal *Nucleic Acids Research* (NAR). The Database Issue of NAR is freely available, and categorizes many of the publicly available online databases related to biology and bioinformatics. A companion database to the issue – the Online Molecular Biology Database Collection – lists 1512 online databases (Fernandez-Suarez & Galperin, 2013). Other collections of databases include MetaBase (Bolser et al., 2012) and the Bioinformatics Links Collection (Brazas, Yim, Yamada, & Ouellette, 2011). The biological databases are well organized and described in these resources, and they can be searched, listed, browsed, or queried. However, it takes time for biologists to choose the exact database they want simply by browsing the descriptions and comparing aspects of the databases.

Bibliometric methods can be applied to evaluate databases so as to make the evaluation easier and more objective. A few studies can be found to measure the impact of databases using bibliometric methods. Urquhart and Dunn (2013) applied bibliometrics to assess the usage of the National Minimum Dataset for Social Care data in scholarly publications and grey literature (Urquhart & Dunn, 2013). Eccles, Thelwall, and Meyer (2012) conducted a webometric analysis of digital resources and found that the comparative link analysis approach was both practical and useful (Eccles et al., 2012). However, one may want to know not only the impact of databases, but also how they connect with each other and evolve with time.

Bibliometric analyses mainly focus on "entities" that can be extracted from the text of publications. Entities are either evaluative entities or knowledge entities (Ding et al., 2013). Evaluative entities have been widely used to evaluate scholarly impact, including papers (de la Pena, 2011), authors (Ding, Yan, Frazho, & Caverlee, 2009; Sun & Han, 2013; Tan, Li, Zhang, & Guo), journals (Medina & van Leeuwen, 2012), institutions (Vieira & Gomes, 2010), and countries (Bornmann & Leydesdorff, 2013). Knowledge entities act as carriers of knowledge units in scientific articles. The most-often-used knowledge entity in bibliometric studies is the keyword, which can represent a research topic or the subject of a field (Hu, Hu, Deng, & Liu, 2013). Knowledge entities can also be topics, key methods, key theories, domain entities (e.g., biological entities: genes, drugs, and diseases), and databases. Ding (2011) combined evaluative entities (i.e., authors and papers) and knowledge entities (i.e., topics) to explain whether productive authors tended to collaborate with and/or cite researchers with the same or different topical interests (Ding, 2011). Ding et al. (2013) proposed the "Entitymetric," to measure the impact of biological entities, such as genes, drugs, and diseases (Ding et al., 2013). Theories are treated as knowledge entities to explore authors' use of theory in information science research (Pettigrew & McKechnie, 2001) and family therapy research (Hawley & Geske, 2000). However, few bibliometric analyses have extended the knowledge entity to "database" and trace the usage of databases.

To this end, we propose a "database link network" (shown in Fig. 1). The connections among databases show the citing/cited relationships that can be exploited by analyzing the topological structure of this network. We further develop the main-path algorithm to trace the evolution of the databases. Main-path analysis identifies those entities that make significant contributions to the knowledge diffusion process. It was first introduced by Hummon and Doreian (1989), in which they used citation information from academic papers to trace the main flow of ideas in DNA development (Hummon & Doreian, 1989). Since then, exploring the development trajectory of a scientific field has commonly been done through main-path analysis (Carley, Hummon, & Harty, 1993; Lu & Liu, 2013; Lucio-Arias & Leydesdorff, 2008). However, these studies confined the application of main-path algorithms to paper citation networks. These algorithms assume the networks are: (1) binary—that is, all citations are treated equally; and (2) acyclic—that is, there are no loops in the network. In fact, there exist networks that are weighted and cyclic: their links have different strengths, and they have at least one directed path that starts and ends at the same node. Examples of this kind of network include author citation networks, journal citation

networks, and the database link network described in this paper. We have modified the original main-path algorithms so that they can trace important knowledge flow in our database link network.

This paper extends current content-based citation analysis to databases by taking the database as one kind of knowledge entity, to form database link networks (Ding et al., 2013). It proposes an easy way to evaluate database usage through citing and cited relationships between databases documented in scholarly publications. Through database link networks, not only can successful databases be promptly identified via degree centrality, but also their usage can be traced through network paths. The main path algorithm (MPA) has been developed by optimization of previous main-path related research, by considering edge-weight differences and cyclic features of the network. This paper uses the bioinformatics literature as a data source for the generation of a database link network, then illustrates the usage, connection, and evolution of databases by analyzing its topological structure and main paths. In addition, we conduct content analysis by comparing content similarities among the papers citing databases.

The present paper is organized as follows: Section 2 outlines a literature review; Section 3 provides details about the methods we developed and applied; Section 4 discusses and evaluates the research results; and Section 5 gives our conclusion and identifies possible future work.

2. Material and methods

2.1. Data

The targeted domain is bioinformatics, and all databases used in this domain are analyzed. PubMed Central (PMC) was chosen as a source for bioinformatics articles. First, key journals in bioinformatics were identified, based on criteria provided by Huang et al. (2012). An additional set of journal-selection criteria was applied, resulting in the inclusion of: (1) The International Society of Computational Biology (http://www.iscb.org/iscb-publications-journals), (2) the bioinformatics journal list on Wikipedia (http://en.wikipedia.org/wiki/List_of_bioinformatics_journals), and (3) the Mathematical and Computational Biology section in the Web of Science's Science Journal Citation Reports (SJCR). From these sources, we drew a comprehensive list of 48 bioinformatics journals. Second, all 20,861 articles published in these 48 journals between 2004 and 2010 were collected from PMC; they include 804,067 references.

2.2. Database extraction

Databases were extracted from the methodology sections of the collected articles and their references. A dictionary containing the list of the available databases was built up based on the online version of *The 2013 Nucleic Acids Research Molecular Biology Database Collection*, which now includes 1512 databases, sorted into 14 categories and 41 subcategories. "Exact-string match" was used to extract databases from the methodology sections. The whole databases in this dictionary were divided into two groups by their names: case-sensitive ones and case-insensitive ones. For example, databases such as "ACTIVITY" and "FLIGHT," which are common words, were extracted with the help of case-sensitive exact-match search; databases such as "CCDB" and "2D-Page," whose names are not common words, were extracted by applying exact match with case ignored.

To identify the methodology sections of collected bioinformatics articles is challenging, as section headers differ in different publications. Database extraction is conducted on those sections relevant to "Methodology": Intro|Methods, Material, Materials, Materials-Methods, Materials|Methods, Methods, Methods|Conclusions, Methods|Discussion, Methods|Materials, Methods|Results, and Methods|Subjects.

2.3. Database link network

Fig. 1 shows how the database link network was generated. For example, if paper A cites paper B (i.e., A → B) (Fig. 1a), and database 1 and database 2 are methoned in the methodology section of paper A, while database 3 and database 4 are mentioned in the methodology section of the cited paper B, then we assume that database 1 cites both database 3 and database 4 (database 1 → database 3, database 1 → database 4), and that database 2 also cites both database 3 and database 4 (database 2 → database 3, database 2 → database 4) (Fig. 1b). For all the references, only those that appear in PMC were used to create the database link network. Because these are full-text references provided by PMC, databases can then be extracted from their methodology sections. In the end, 32,718 references were identified in PMC, which account for 4.5% of all the journal references.

Two database link networks were generated: a link-weighted network and a similarity-weighted network. For the first, nodes represent databases, links represent "cites," and link weight represents link count. This network has 591 nodes and 15,449 links. The density of the network is 0.044. The largest link weight (link count) is 1281, with database "GO" being both the start and end nodes. For the second, the nodes and links are the same as the first one, while link weights represent topical similarity between two databases.

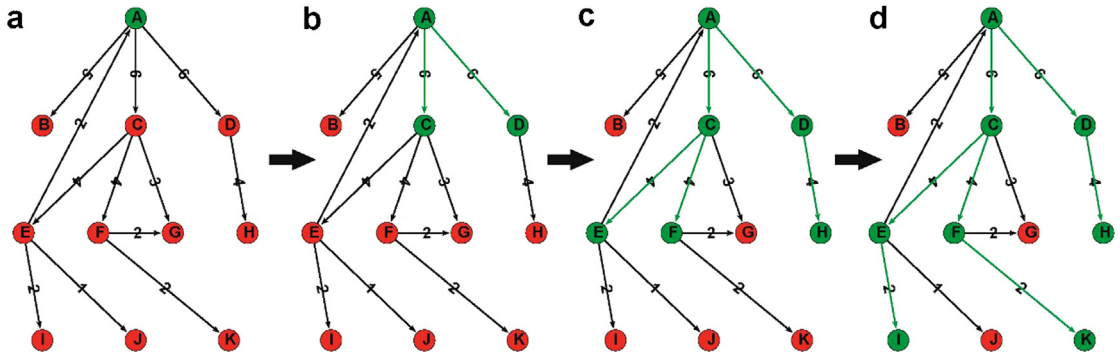

Fig. 2. Main path procedures.

2.4. Database topical similarity

Bio-LDA was used to calculate the topical distribution for a given database (Jie, Ruoming, & Jing, 2008). Bio-LDA is an extended simultaneous Latent Dirichlet allocation of modeling papers, topics, and bio entities (e.g., database, gene, drug, disease). It calculates the probability of a topic for a given bio entity (such as database), the probability of a bio entity for a given topic, the probability of a topic for a given document, the probability of a document for a given topic, the probability of a topic for a given word, and the probability of a word for a given topic. Base on the calculated topical probability distribution for each database, the dissimilarity between any two databases can be measured by Kullback–Leibler divergence (K–L divergence), which is a non-symmetric measure of the difference between two probability distributions P and Q, denoted as $D_{KL}(P||Q)$ (Kullback & Leibler, 1951). For discrete probability distributions P and Q, the K–L divergence of Q from P is defined to be:

$$D_{KL}(P \parallel Q) = \sum_i \ln\left(\frac{P_i}{Q_i}\right) p(i)$$

It is the expectation of the logarithmic difference between the probabilities P and Q, where the expectation is taken using the probabilities P. Then the database topical similarity S_{PQ} is computed as follow:

$$S_{PQ} = 1 - D_{KL}(P \parallel Q).$$

2.5. Weighted main path algorithm

The original main path algorithms, such as search path link count (SPLC), search path node pair (SPNP), node pair projection count (NPPC), and search path count (SPC), simplify binary and acyclic citation networks. However, many networks are weighted and cyclic—for example, author citation networks, journal citation networks, or other entity citation networks. A high link-weight always indicates a strong connection. Obviously, original main path algorithms are inapplicable to these kinds of networks, as they cannot make full use of the relationships among the databases (such as edge weight and database topical similarity). In view of this, a new algorithm for finding a main path in weighted cyclic networks (called the "weighted MP" algorithm) is proposed here. The weighted MP algorithm is as follows:

(1) Create an empty network N and an empty node-set S.
(2) Choose a node to start with. This can be done either by selecting a node with high centrality value (e.g., degree, closeness, betweenness, or pagerank), or one in which specific users have an interest. Add the node to network N and node-set S.
(3) Create a new empty node set, S_current. Find all the outgoing links for the current start point(s). Select the link(s) with the highest weight. For each of these links, check whether its end node is in S. If not, add the end node to N, S, and S_current, and add the link to N. Take all the node(s) in S_current as the start point(s) for the next step.
(4) Repeat step 3 until there are no outgoing links for all the current start point(s); i.e., all paths hit sinks.
(5) Find the longest path(s) in N, and take these paths as the main paths.

A simple database link network in Fig. 2 is used to demonstrate how the weighted MP is calculated.

(1) Step 1: Create empty network N. Choose node A as a start point, and add node A to network N (Fig. 2a).
(2) Step 2: Find all the outgoing links from node A. Select those with the highest weight: A–C and A–D. Add edges A–C and A–D and nodes C and D to network N (Fig. 2b).

Table 1
Top five main paths (citation-weighted dataset citation network).

Path no.	Main path
Path 1 (45 steps)	GenBank ≫ GO ≫ UniGene ≫ RefSeq ≫ UCSC Genome Browser ≫ FlyBase ≫ GEO ≫ COMe ≫ Pfam ≫ PDB ≫ SCOP ≫ DIP ≫ SGD ≫ Inparanoid ≫ InterPro ≫ SMART ≫ COG ≫ CDD ≫ IntAct ≫ KEGG ≫ EcoCyc ≫ RegulonDB ≫ Stanford Microarray Database ≫ CC+ ≫ UniProt ≫ PubMed ≫ OMIM ≫ HPRD ≫ MINT ≫ BioGRID ≫ IPI ≫ PeptideAtlas ≫ PROSITE ≫ PseudoGene ≫ dbSNP ≫ HapMap Project ≫ PANTHER ≫ Entrez Gene ≫ HomoloGene ≫ TRANSFAC ≫ SCPD ≫ TAIR ≫ AGRIS ≫ PLACE ≫ PlantCARE
Path 2 (37 Steps)	GO ≫ GenBank ≫ Pfam ≫ PDB ≫ SCOP ≫ DIP ≫ SGD ≫ RefSeq ≫ UniGene ≫ GEO ≫ COMe ≫ FlyBase ≫ UCSC Genome Browser ≫ TRANSFAC ≫ KEGG ≫ COG ≫ SMART ≫ CDD ≫ IntAct ≫ Inparanoid ≫ OMIM ≫ HPRD ≫ UniProt ≫ PubMed ≫ WormBase ≫ CE ≫ CC+ ≫ HomoloGene ≫ Entrez Gene ≫ Stanford Microarray Database ≫ RegulonDB ≫ Rfam ≫ miRBase ≫ TAIR ≫ AGRIS ≫ PLACE ≫ PlantCARE
Path 3 (37 steps)	RefSeq ≫ GenBank ≫ GO ≫ UniGene ≫ GEO ≫ COMe ≫ Pfam ≫ PDB ≫ SCOP ≫ DIP ≫ SGD ≫ Inparanoid ≫ InterPro ≫ SMART ≫ COG ≫ CDD ≫ FlyBase ≫ UCSC Genome Browser ≫ RANSFAC ≫ KEGG ≫ EcoCyc ≫ RegulonDB ≫ Stanford Microarray Database ≫ CC + ≫ UniProt ≫ PubMed ≫ IntAct ≫ HPRD ≫ OMIM ≫ Entrez Gene ≫ HomoloGene ≫ Rfam ≫ miRBase ≫ TAIR ≫ AGRIS ≫ PLACE ≫ PlantCARE
Path 4 (45 steps)	Pfam ≫ GenBank ≫ GO ≫ UniGene ≫ RefSeq ≫ UCSC Genome Browser ≫ FlyBase ≫ GEO ≫ COMe ≫ PDB ≫ SCOP ≫ DIP ≫ SGD ≫ Inparanoid ≫ InterPro ≫ SMART ≫ COG ≫ CDD ≫ IntAct ≫ KEGG ≫ EcoCyc ≫ RegulonDB ≫ Stanford Microarray Database ≫ CC+ ≫ UniProt ≫ PubMed ≫ OMIM ≫ HPRD ≫ MINT ≫ BioGRID ≫ IPI ≫ PeptideAtlas ≫ PROSITE ≫ PseudoGene ≫ dbSNP ≫ HapMap Project ≫ PANTHER ≫ Entrez Gene ≫ HomoloGene ≫ TRANSFAC ≫ SCPD ≫ TAIR ≫ AGRIS ≫ PLACE ≫ PlantCARE
Path 5 (38 steps)	UniProt ≫ GO ≫ GenBank ≫ Pfam ≫ PDB ≫ SCOP ≫ DIP ≫ SGD ≫ RefSeq ≫ UniGene ≫ GEO ≫ COMe ≫ FlyBase ≫ UCSC Genome Browser ≫ TRANSFAC ≫ KEGG ≫ COG ≫ SMART ≫ CDD ≫ IntAct ≫ Inparanoid ≫ OMIM ≫ HPRD ≫ MINT ≫ PubMed ≫ WormBase ≫ CE ≫ CC+ ≫ HomoloGene ≫ Entrez Gene ≫ Stanford Microarray Database ≫ RegulonDB ≫ Rfam ≫ miRBase ≫ TAIR ≫ AGRIS ≫ PLACE ≫ PlantCARE

(3) Step 3: Find all the outgoing links from nodes C and D. Select those with the highest weight: C–E, C–F, and D–H. Add edges C–E, C–F, and D–H, and nodes E, F, and H, to network N (Fig. 2c).
(4) Step 4: Find all the outgoing links from nodes E, F, and H. Select those with the highest weight: E–A, E–I, and F–K. For link E–A, end-node A has been visited before, so this link should be ignored; only add edges E–I and F–K, and nodes I and K, to network N (Fig. 2d).
(5) Step 5: Nodes K and N are sinks, so the calculation stops here. In network N, the longest paths starting from node A are A–C–E–I and A–C–F–K; these two paths are therefore the main paths.

Main paths for both link-weighted networks and similarity-weighted networks can be calculated using the weighted MP algorithm. The database evolution based on both database link count and database topical similarity can be identified and analyzed.

3. Results

Our database link networks are weighted and cyclic. There are two ways to calculate the weight of two databases: one is based on the number of times one database cites another database, and the other is based on the topic similarity of the two databases. A database link network whose weight is link frequency is called a link-weighted database link network. One whose weight is topic similarity is called a similarity-weighted database link network. The weighted MP algorithm was applied to both link-weighted and similarity-weighted database link networks.

3.1. Main path analysis: Link-weighted database link network

To examine the database diffusion pattern, we select the top five databases by degree in the database link network—GenBank, GO, RefSeq, Pfam, and UniPort. For each of these databases, we generated its main path by applying weighted MP to the link-weighted database link network (see Table 1). There are 49 unique databases shown in the five main paths, and all of them belong to 12 categories. For example, nine databases are in Nucleotide Sequence (18.4%), eight

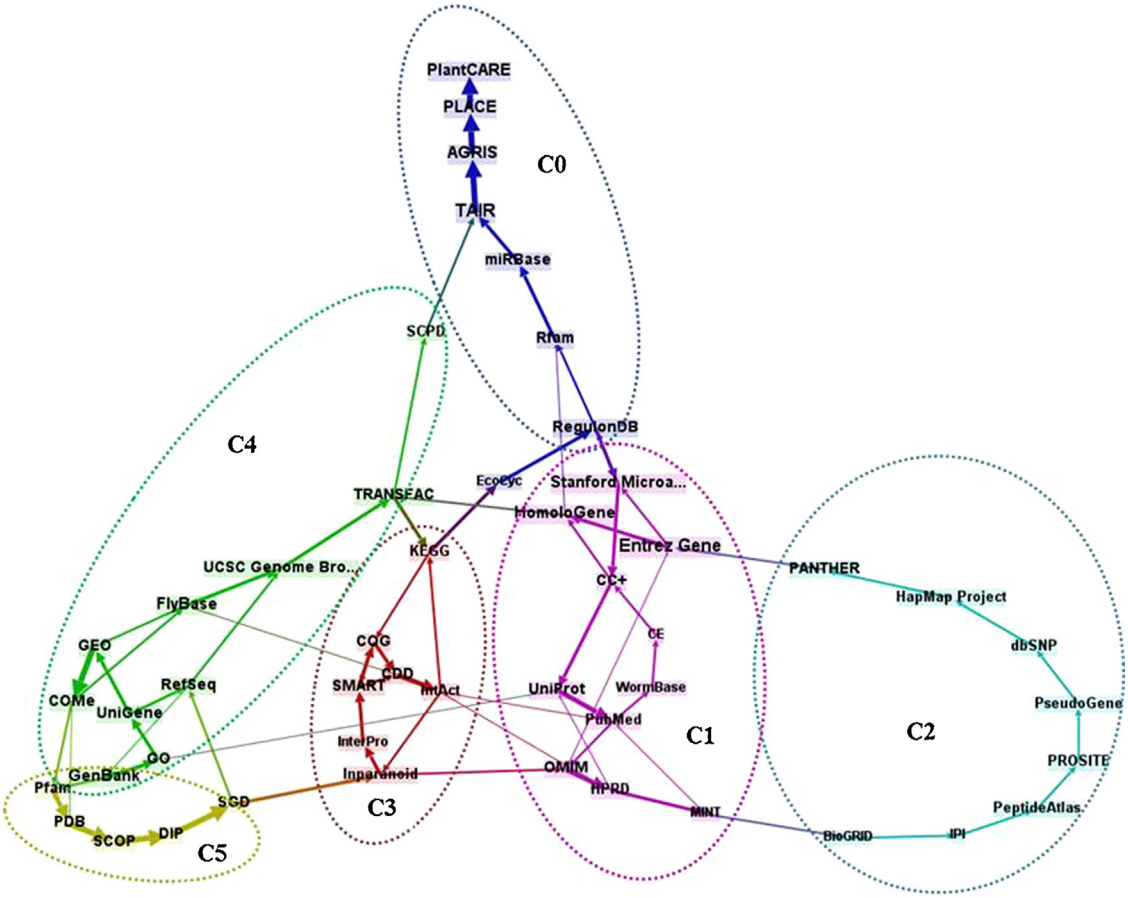

Fig. 3. Main path network (link-weighted database link network).

databases in Genomics and in Protein Sequence (16.3%), respectively, and six databases in Human and other Vertebrate Genomes (12.2%). The common path appearing in the five main paths is TAIR → AGRIS → PLACE → PlantCARE.

A directed network containing these five main paths was built; it contains 49 nodes and 80 edges (see Fig. 3). Each node represents an individual database, each edge indicates a link relationship, and the weight of an edge shows the strength of the two databases in these five main paths. For example, if the link from GenBank to GO appears in paths 1, 3, and 4, the weight of the link is three. The Louvain method was applied to detect the major components of this directed network, and six components (modularity value = 0.637) were identified. The Louvain method is a modularity algorithm to identify communities in large networks by optimizing the modularity of a partition of the network (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). We used the Louvain method provided in Gephi. The optimization consists of two steps. First, it searches for small communities by optimizing modularity locally. Second, it builds a new network by aggregating nodes in the same community. Separation of sequential components into sub-components results mainly from the characteristics of community detection by the Louvain method. In the Louvain method, a portion of a network is separated into different components if its network property is clearly different from the random network's one. Component 1 has the highest degree where it is linked to the other three components by both in-degree and out-degree links. Component 2 has a direct link to component 1, and is indirectly linked to other components via component 1. The components that span the longest distance are components 2 and 5. Two components are in between components 2 and 5, which indicates that there is not much link flow between them.

Since a category of a database can be treated as a subfield of the biomedical domain, information flow between subject areas can be analyzed by incorporating the category of the database into the main path analysis. For example, for the link-weighted database link network, main path analysis shows that GenBank and GO are connected by links. If the category of these two databases was included in the main path, it shows that Nucleotide Sequence and Genomics (non-vertebrate) are connected by link, which indicates that information flows from Genomics to Nucleotide Sequence, if GenBank cites GO. Therefore, by adding category information to the main path, it is possible to identify the diffusion of information among different subject areas. Database category information is available at NAR (http://www.oxfordjournals.org/nar/database/c/). By replacing databases with their categories, Fig. 3 can be converted into Fig. 4.

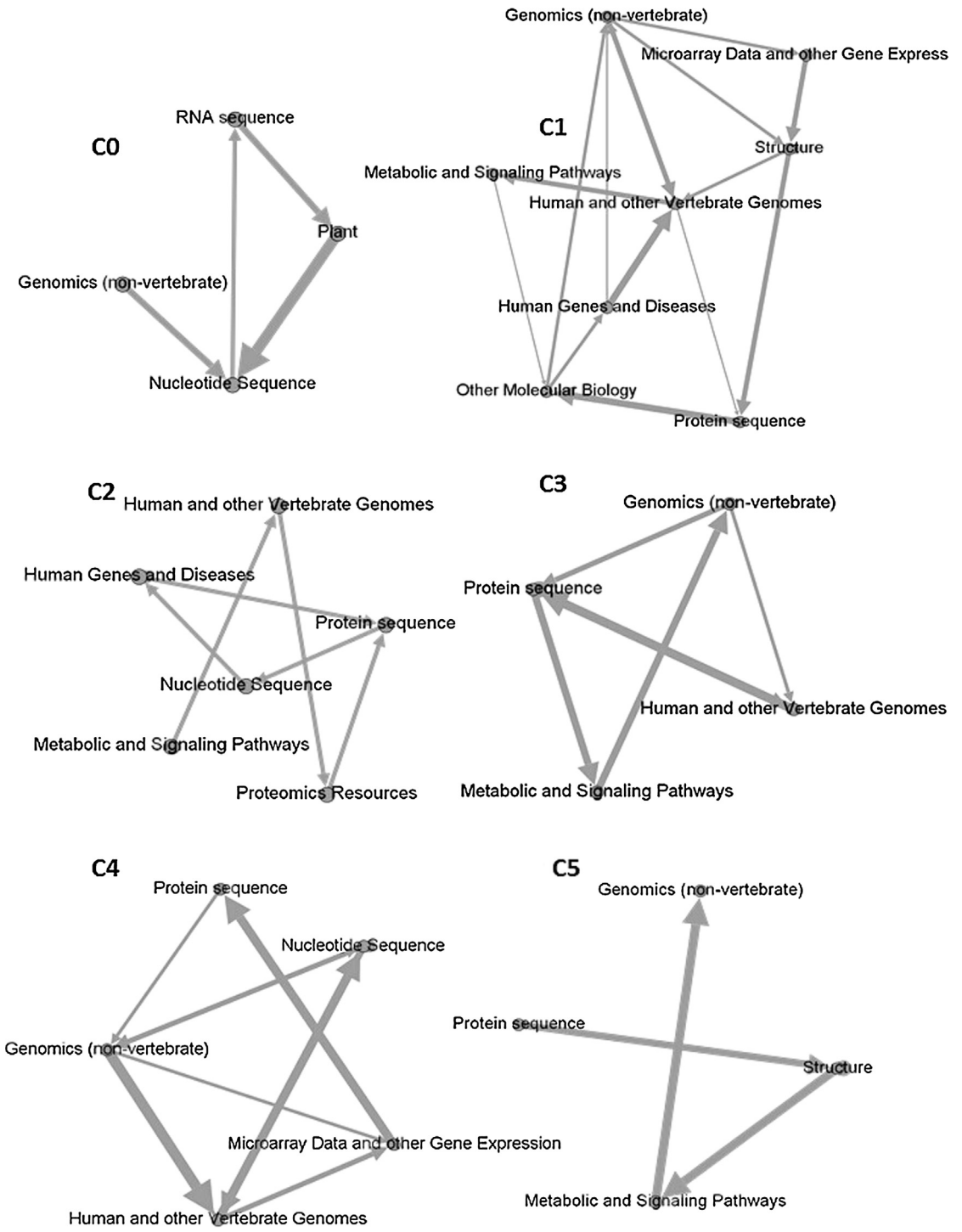

Fig. 4. Patterns of information transfer within subject categories (link-weighted database link network).

The component that has the most subject categories is component 1, which maps 11 databases into eight subject categories. In components 0, 2, 3, and 5, only a single path appears. In component 0, the research expansion shows the following flow: Genomics (non-vertebrate) → Nucleotide Sequence → RNA sequence → Plant → Nucleotide Sequence. In component 2, the following path is shown: Metabolic and Signaling Pathways → Human and other Vertebrate Genomes → Proteomics Resources → Protein Sequence → Nucleotide Sequence → Human Genes and Diseases → Protein Sequence. In component 3, the research expansion among subjects flows as follows: Genomics

Table 2
Top 20 keywords for each component (citation-weighted dataset citation network).

Rank	Component 0	Component 1	Component 2	Component 3	Component 4	Component 5
1	miRNA	Gene	Gene	Genome	Gene	Datum
2	Window	Protein	Genotyping	Gene	Genome	Gene
3	Homolog	Interaction	Process	Sequence	Sequence	Set
4	Agilent	Network	Expression	Protein	Region	Term
5	Subset	Set	Datum	Family	Set	Sequence
6	Mix	Datum	Microarray	Group	Site	Annotation
7	Level	Sequence	System	Set	Datum	Process
8	Equation	Pair	Sequence	Datum	Alignment	Protein
9	Ortholog	Annotation	List	Annotation	Position	Model
10	Adaptor	List	Set	Alignment	Pair	Algorithm
11	Homology	Model	Test	Search	Annotation	Case
12	Mutation	Case	Category	Cluster	DNA	Distribution
13	Specificity	Test	Genome	Tree	Test	Information
14	Pool	Term	Protein	Model	Model	Parameter
15	Density	Information	Query	Pair	Exon	Probability
16	Precursor	Source	Cell	Comparison	Level	Test
17	Format	Node	Model	Domain	Distribution	Level
18	CT	Group	Site	Acid	Primer	Pair
19	Strain	Resource	Rate	Region	Cell	Example
20	Stem	Measure	Kit	Strain	Parameter	Size

(non-vertebrate) → Human and other Vertebrate Genomes ↔ Protein Sequence → Metabolic and Signaling Pathways → Genomics (non-vertebrate) → Protein Sequence ↔ Human and other Vertebrate Genomes. Component 5 shows the path of Protein Sequence → Structure → Metabolic and Signaling Pathways → Genomics (non-vertebrate). In components 1 and 4, the paths among subject categories are complex, which indicates that research among subject fields is cited in several different paths.

Furthermore, a content analysis was conducted by extracting keywords from the methodology sections of articles that mentioned at least one database from the NAR list. Keywords were extracted by:

Step 1: For each database from the top five main paths, extract keywords from the full-text methodology sections of articles that mention this database.
Step 2: Select one pair (A → B) of databases from one of the top five main paths that belongs to a component. For each pair, select the top twenty keywords that appear in the methodology sections of both database A and database B.
Step 3: Select representative keywords in the keyword list that combines the top twenty keywords from each pair of the main path from one component.

These keywords show the major concepts or themes of a component. Table 2 shows the top 20 keywords for each component.

In component 0, the top-ranking keyword is miRNA, and terms such as homolog and ortholog uniquely appear in component 0. Component 1, whose top 4 terms are gene, protein, interaction, and network, has the widest range of subject categories linking to other components. Component 2 has genotyping, expression, and microarray as its major keywords, which indicates the component is pertinent to microarray analysis. Component 3 has family, group, cluster, and tree as its top keywords, showing that it is related to analysis of similar genes and gene sequences. Component 4 has unique keywords DNA and exon, which do not appear in other components. Component 5 has datum, process, distribution, information, and probability as keywords, demonstrating that its major theme is related to data analysis.

3.2. Main path analysis: Similarity-weighted database link network

Similarly, the weighted MP algorithm was applied to the similarity-weighted database link network on the directed network formed by the top five database link paths from a similarity-weighted database link network (see Table 3). There are 48 unique databases that appeared in these top five main paths, belonging to nine categories. For example, 14 databases are in Genomics (29.2%), 11 databases are in Nucleotide Sequence (22.9%), and nine databases are in Protein Sequence (18.8%). The most frequently occurring databases in the five main paths are ABA, EPD, GenePaint, HomoloGene, and SAGEmap, which all belong to component 0.

The directed network built from the top five main paths consists of 48 nodes and 48 edges (see Fig. 5). Each node represents a database, edges show link flow, and the weight of an edge is determined by the number of occurrences of two given linked nodes in the top five paths. For example, if the link from EPD to SAGEmap appears in paths 1, 3, 4, and 5, the weight of the link is four. The Louvain method was applied to detect the major components of the directed network, and seven components (modularity value = 0. 719) were identified. Component 0 acts as a hub to connect other components, and the databases in this component connect different main paths. The relationship between components 1 and 4 and the relationship between

Table 3
Top five main paths (similarity-weighted dataset citation network).

No	Main path
Path 1 (19 steps)	GenBank ≫ UCSC Genome Browser ≫ BPS ≫ WormBase ≫ HomoloGene ≫ ABA ≫ GenePaint ≫ EPD ≫ SAGEmap ≫ FlyBase ≫ IntAct ≫ PRO ≫ CC+ ≫ ABS ≫ ORegAnno ≫ CORG ≫ TRANSPATH ≫ MINT ≫ Yeast Resource Center
Path 2 (14 steps)	GO ≫ RegulonDB ≫ STRING ≫ EcoCyc ≫ ENZYME ≫ BRENDA ≫ ASAP ≫ TransportDB ≫ MetaCyc ≫ The Comprehensive Microbial Resource ≫ ProDom ≫ CAMERA ≫ Enzyme Nomenclature ≫ ApiDB
Path 3 (19 steps)	RefSeq ≫ UCSC Genome Browser ≫ BPS ≫ WormBase ≫ HomoloGene ≫ ABA ≫ GenePaint ≫ EPD ≫ SAGEmap ≫ FlyBase ≫ IntAct ≫ PRO ≫ CC+ ≫ ABS ≫ ORegAnno ≫ CORG ≫ TRANSPATH ≫ MINT ≫ Yeast Resource Center
Path 4 (32 steps)	Pfam ≫ UniProt ≫ InterPro ≫ NCBI Taxonomy ≫ PROSITE ≫ BAliBASE ≫ SCPD ≫ PlantProm ≫ TAIR ≫ CAGE ≫ DBTSS ≫ ASC ≫ EPD ≫ SAGEmap ≫ HomoloGene ≫ ABA ≫ GenePaint ≫ COMe ≫ SGD ≫ STRING ≫ EcoCyc ≫ ENZYME ≫ BRENDA ≫ ASAP ≫ TransportDB ≫ MetaCyc ≫ The Comprehensive Microbial Resource ≫ ProDom ≫ CAMERA ≫ Enzyme Nomenclature ≫ ApiDB
Path 5 (30 steps)	UniProt ≫ InterPro ≫ NCBI Taxonomy ≫ PROSITE ≫ BAliBASE ≫ SCPD ≫ PlantProm ≫ TAIR ≫ CAGE ≫ DBTSS ≫ ASC ≫ EPD ≫ SAGEmap ≫ HomoloGene ≫ ABA ≫ GenePaint ≫ COMe ≫ SGD ≫ STRING ≫ EcoCyc ≫ ENZYME ≫ BRENDA ≫ ASAP ≫ TransportDB ≫ MetaCyc ≫ The Comprehensive Microbial Resource ≫ ProDom ≫ CAMERA ≫ Enzyme Nomenclature ≫ ApiDB

components 3 and 5 are sequential, unlike those of the other components in Fig. 3. Component 2, 4, 5 and 6 are directly connected to component 0. On the other hand, component 1 and 3 are indirectly connected to component 0 via component 4 and 5. For example, in the main path going from Pfam in component 3 to ASC in component 5, the first half – from Pfam to BAliVASE – is in component 3, and the second half – from SCPD to SAC – is in component 5. These paths show the databases' usage diffusion; for instance, component 2 shows that a study employing the PRO database is expanded to a study employing the CC+, and subsequently to a study employing the Yeast Resource Center database.

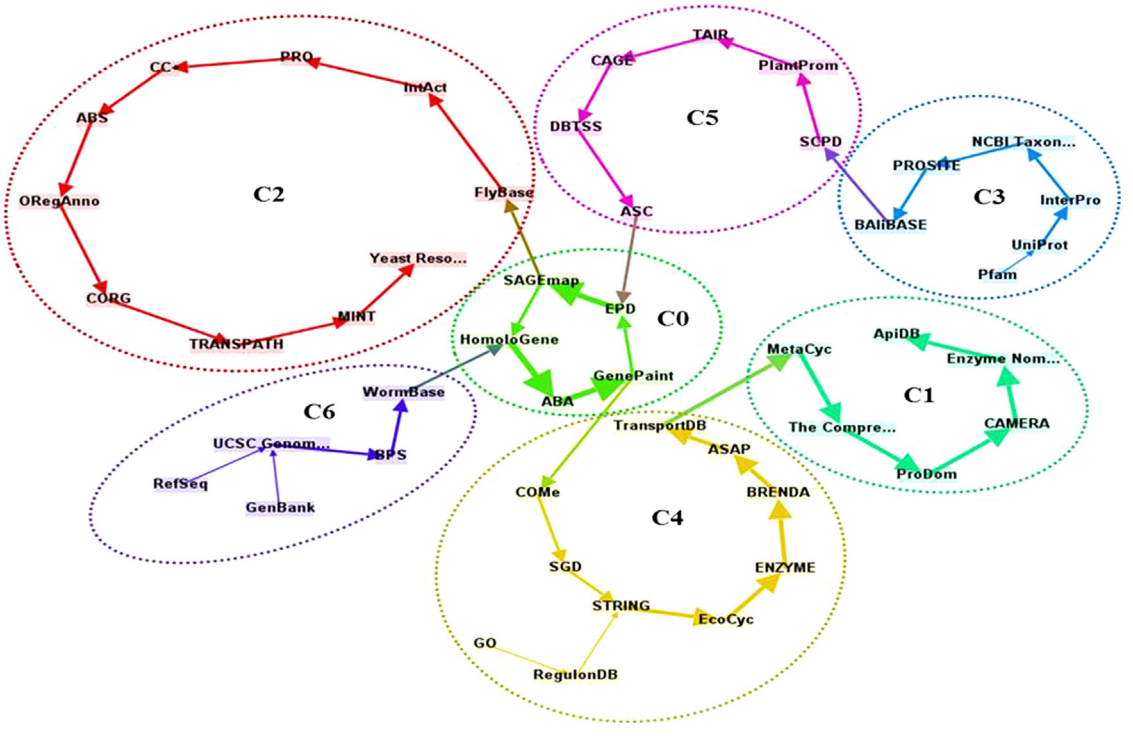

Fig. 5. Main path network (similarity-weighted database link network).

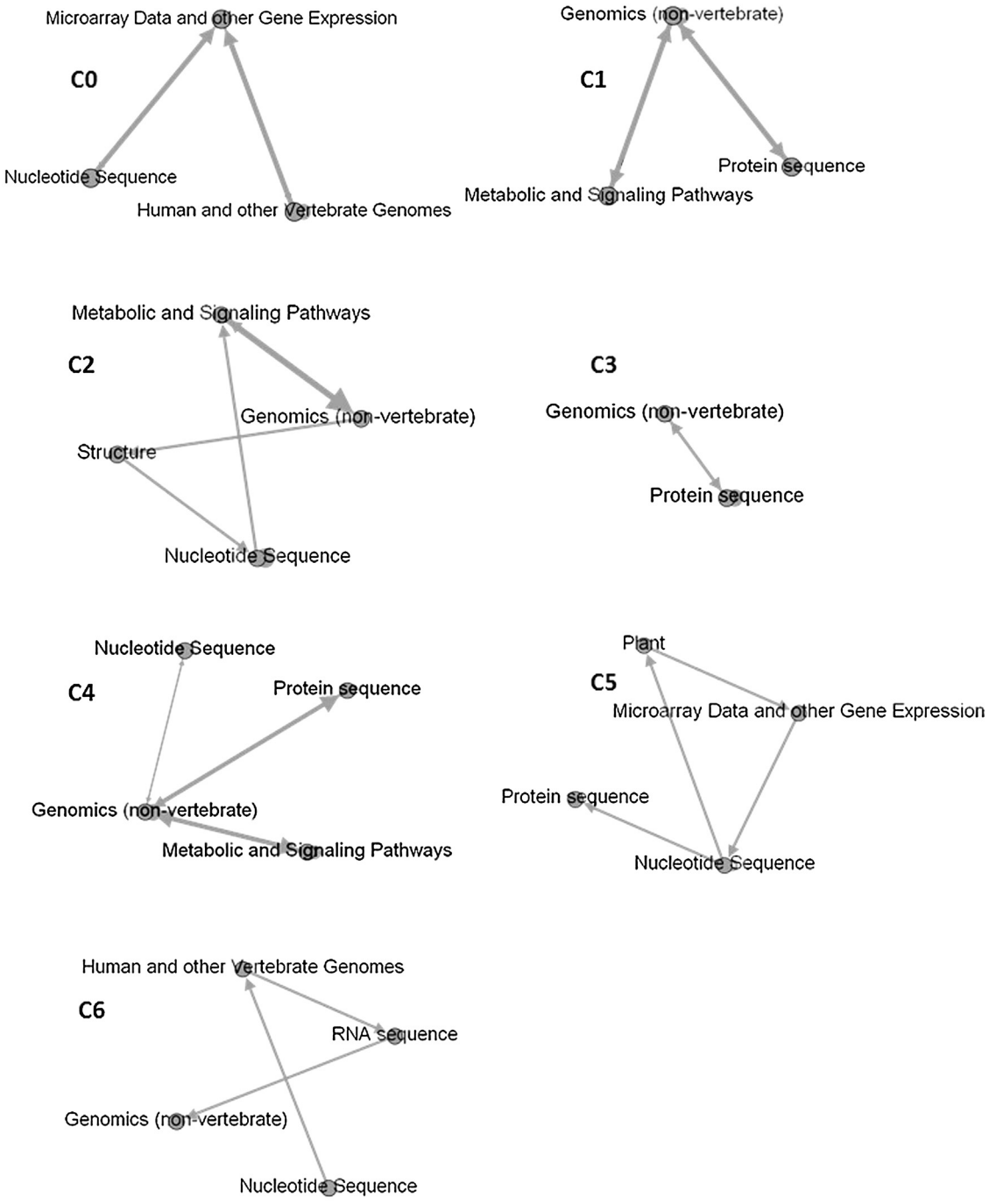

Fig. 6. Patterns of information transfer within subject categories (similarity-weighted database link network).

By replacing each database with its category, Fig. 5 can be converted into Fig. 6. Solid lines denote database link relations within components, and dotted lines show link relations between components. Component 3 and 5 have the most subject categories (i.e., 5), while component 0 has the fewest (i.e., 3).

In components 2, (3, 5), and 6, only a single path appears. Thus, the information diffusion path of researches by subject is relatively clear. In component 2, the information diffusion shows the following path: Metabolic and Signaling Pathways ↔ Genomics (non-vertebrate) → Structure → Nucleotide Sequence. In component (3, 5), the following path is shown: Genomics (non-vertebrate) ↔ Protein Sequence → Nucleotide Sequence → Plant → Microarray Data and other Gene Expression → Nucleotide Sequence ↔ Protein Sequence ↔ Genomics (non-vertebrate). In component 6, information flows as follows: Nucleotide Sequence → Human and other Vertebrate Genomes → RNA sequence → Genomics (non-vertebrate).

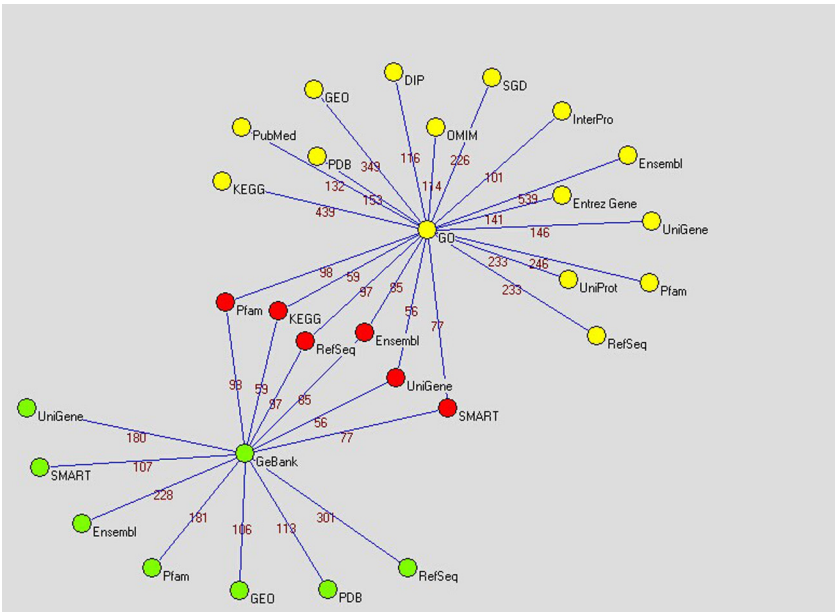

Fig. 7. The connections between GO and GeneBank with other major databases.

Components 0, 1, and 4 show similar patterns, which connect to other categories from one central category. In component 0, Nucleotide Sequence and Human and other Vertebrate Genomes are connected by Microarray Data and other Gene Expression categories. In components 1 and 4, three categories – Metabolic and Signaling Pathways, Nucleotide Sequence, and Protein sequence – are connected by Genomics (non-vertebrate). These two cases show that the central categories play a pivotal role in connecting categories in a given component.

3.3. Use cases

3.3.1. GO and GeneBank

To examine what databases are closely related to GO and GeneBank, respectively, we calculated the co-occurrence frequency between these two databases and other databases that are co-mentioned in the methodology section of the full-text papers. In calculation of co-occurrence frequency, we separately counted frequency when both GO and GenBank appear or when one of them appears in the methodology section of the full-text papers. Fig. 7 shows how databases are connected to each other by having GO or GenBank a hub.

In case of GO only, the total 252 databases were co-mentioned with GO and 14 of them (highlighted in yellow) were retained (frequency > 100). In case of GenBank only, 241 databases were co-mentioned with GenBank and 7 of them (highlighted in green) were retained (frequency > 100). In case of GO and GenBank both, 113 databases were co-mentioned with both GO and GenBank and 6 of them (highlighted in red) were retained (frequency > 50). The edge weight represents how often these databases are co-mentioned with GO only, GenBank only or GO and GenBank both in the methodology section of the full-text articles.

As shown in Fig. 7, GO is co-mentioned with databases such as Entrez Gene, KEGG, and GEO whereas GenBank is co-mentioned with RefSeq, Ensembl, and Pfam in the methodology section. Although there is a difference between GO connection and GenBank connection in ranking by frequency, there is a high overlap of databases. 6 out of 7 databases connected to GenBank are also connected to GO. Except for SMART and KEGG, the rest of databases that appear in case of GO and GenBank both are also shown in GO only and GenBank only case.

One interesting observation is that GO is connecting to various different databases since it functions as general gene identification. On the contrary, GenBank is limited to gene and protein databases only. It is attributed to the fact that GO and GenBank are two distinct databases preserving different data characteristics. GenBank is the database for Nucleotide Sequence and Protein Sequence of organisms whereas GO is an ontology for genes.

3.3.2. PlantCARE and PLACE

We created the semantic graphs (see Fig. 8) about PlantCARE and PLACE to see how they were semantically mentioned in one whole sentence of the methodology sections of our whole set of publications.

Both PlantCARE and PLACE are pertinent to *cis*-acting regulatory elements of plants. Because both share common themes, they are often cited together. These two databases are primarily used to understand *cis*-acting previously discovered. Since datasets stored in the databases are specialized in a certain domain and there are only a handful size of databases related

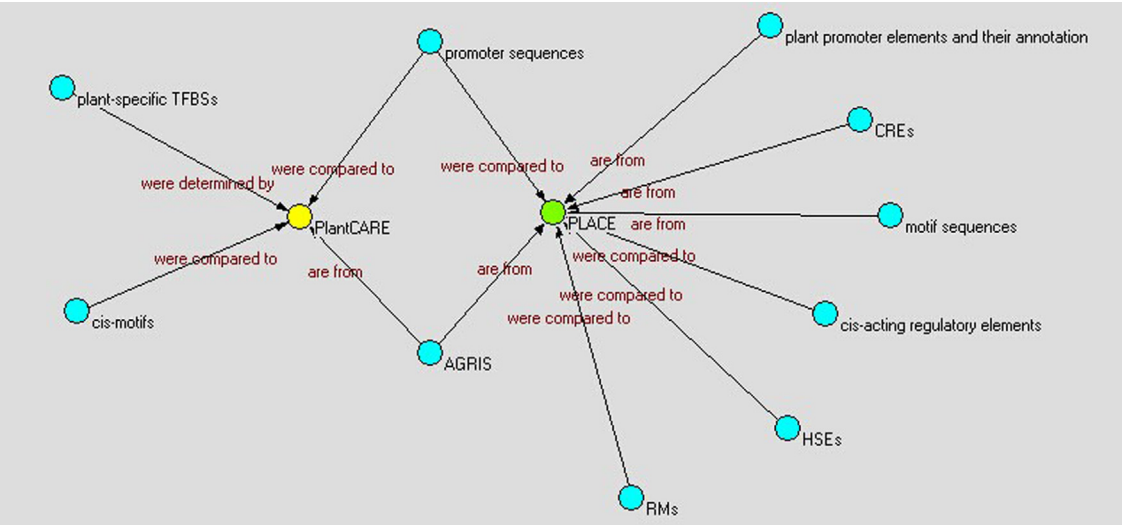

Fig. 8. Semantic graph of PlantCARE and PLACE.

to plants, it is rare for PlantCARE and PLACE to be used with other databases. If we examine the list of verbs used in the sentence where a database name is mentioned in the methodology section, we can see the verbs such as "compare", "search", "retrieve", etc. This indicates that the database is used to retrieve data and compare the newly discovered facts with the database entry. As mentioned earlier, these two databases are not co-mentioned with other databases except for TAIR, TRANSFAC, and PubMed.

4. Discussion

To examine topical resemblance between databases in one main path, the topic similarity of a main path was calculated by ALC (Average-Linkage Clustering) based on articles that mention these databases in their methodology:

$$S(\mathbf{X}, \mathbf{Y}) = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} D\left(x_i, y_j\right),$$

where **X** and **Y** denote two databases connected in a main path, and $D(x_i, y_j)$ denotes the cosine similarity between documents x_i, y_j, whose methodology sections mention X or Yr. The average similarity per main math was calculated using the mean of S(X,Y). The top five main paths from the link-weighted database link network (see Table 1) and the similarity-weighted database link network (see Table 3) were compared to determine whether there is a difference in topic similarity (see Table 4). If the difference in the average similarity per main path is significant, that may imply that there are main paths that have high or low topic similarity among databases in the path.

As shown in Table 4, the difference of the average similarity is minor. The main paths in the link-weighted database link network show a smaller difference than those main paths in the similarity-weighted database link network. One reason is that there is a high level of overlap among databases in main paths in link-weighted database link network. Among these ten main paths, path 1 of the similarity-weighted database link network shows the lowest similarity (0.1803), and path 2 of the same network shows the highest similarity (0.2033). The average topic similarity of paths does not show a big difference, but there is a meaningful difference in topic similarity between databases within a path. For example, even if path 1 of the similarity-weighted database link network is lowest-ranked by similarity, the pair CORG and TRANSPATH has a similarity of 0.2396, which is higher than that of the top-ranked pair in path 2, Enzyme Nomenclature and ApiDB, whose similarity is 0.2306.

Table 4
Average topic similarity.

Citation-weighted dataset citation network					
Path 1	Path 2	Path 3	Path 4	Path 5	Average
0.1995	0.1923	0.1957	0.1996	0.1948	0.1964
Similarity-weighted dataset citation network					
0.1803	0.2033	0.1817	0.1984	0.1988	0.1925

Table 5
Topic similarity (Path 1 in the citation-weighted dataset citation network).

Source	Target	Similarity	Source	Target	Similarity
GenBank	GO	0.1743	SMD**	CC+	0.1656
GO	UniGene	0.2022	CC+	UniProt	0.1514
UniGene	RefSeq	0.2023	UniProt	PubMed	0.1679
RefSeq	UCSC GB*	0.2084	PubMed	OMIM	0.1857
UCSC GB*	FlyBase	0.1858	OMIM	HPRD	0.2146
FlyBase	GEO	0.1825	HPRD	MINT	0.2573
GEO	COMe	0.1662	MINT	BioGRID	0.2488
COMe	Pfam	0.1602	BioGRID	IPI	0.1890
Pfam	PDB	0.1683	IPI	PeptideAtlas	0.2030
PDB	SCOP	0.2074	PeptideAtlas	PROSITE	0.1581
SCOP	DIP	0.2107	PROSITE	PseudoGene	0.1848
DIP	SGD	0.2122	PseudoGene	dbSNP	0.1894
SGD	Inparanoid	0.2084	dbSNP	HapMap Project	0.2202
Inparanoid	InterPro	0.2038	HapMap Project	PANTHER	0.1867
InterPro	SMART	0.1875	PANTHER	Entrez Gene	0.2013
SMART	COG	0.1810	Entrez Gene	HomoloGene	0.2147
COG	CDD	0.1981	HomoloGene	TRANSFAC	0.2189
CDD	IntAct	0.1526	TRANSFAC	SCPD	0.2481
IntAct	KEGG	0.1657	SCPD	TAIR	0.1951
KEGG	EcoCyc	0.2010	TAIR	AGRIS	0.2381
EcoCyc	RegulonDB	0.2164	AGRIS	PLACE	0.2750
RegulonDB	SMD**	0.2053	PLACE	PlantCARE	0.2635

* UCSC Genome Browser.
** Stanford Microarray Database.

To examine the distribution pattern of topic similarity between databases in a main path, the longest main path from both networks were selected, and their database topic similarities were compared (see Tables 5 and 6). In Table 5, the database pair showing the highest topic similarity is AGRIS and PLACE (0.275), and the pair showing the lowest topic similarity is CC+ and UniProt (0.1514). In path 4 of the similarity-weighted database link network, the highest topic similarity belongs to the pair SCPD and PlantProm (0.2856), and the lowest topic similarity is observed in the pair GenePaint and COMe (0.1331).

High topic similarity between databases means that there are many terms commonly appearing in articles that mention these two databases. However, the coverage and the variety of information provided by a particular database would also influence the subject area of an article mentioning a database. For example, in Table 5, AGIS and PLACE – showing the highest similarity – are related to plant biotechnology, and the content provided by these databases is subject dependent. AGIS is an information resource of the Arabidopsis promoter sequence, transcription factors, and their target genes. PLACE is a database of motifs found in plant *cis*-acting regulatory DNA elements. In plant biotechnology, the plant model most used in experiments is Arabidopsis thaliana, and it therefore causes high subject cohesiveness. On the other hand, CC+ and UniProt, with the lowest similarity, provide different content related to proteins. CC+ is a detailed searchable repository of coiled-coil assignments. UniProt is the central hub for the collection of functional information on proteins, with accurate, consistent, and rich annotation. In particular, UniProt provides integrated information, which can be used for different research domains. This is attributed to the low similarity between UniProt and CC+.

Table 6
Topic similarity (Path 4 in similarity-weighted dataset citation network).

Source	Target	Similarity	Source	Target	Similarity
Pfam	UniProt	0.1872	ABA	GenePaint	0.1686
UniProt	InterPro	0.1909	GenePaint	COMe	0.1331
InterPro	NCBI Taxonomy	0.1926	COMe	SGD	0.1790
NCBI Taxonomy	PROSITE	0.1875	SGD	STRING	0.2025
PROSITE	BAliBASE	0.1956	STRING	EcoCyc	0.2056
BAliBASE	SCPD	0.2290	EcoCyc	ENZYME	0.1940
SCPD	PlantProm	0.2856	ENZYME	BRENDA	0.2111
PlantProm	TAIR	0.2226	BRENDA	ASAP	0.1697
TAIR	CAGE	0.1969	ASAP	TransportDB	0.1781
CAGE	DBTSS	0.2462	TransportDB	MetaCyc	0.2093
DBTSS	ASC	0.1658	MetaCyc	The CMR*	0.1976
ASC	EPD	0.1635	The CMR*	ProDom	0.2117
EPD	SAGEmap	0.1949	ProDom	CAMERA	0.2139
SAGEmap	HomoloGene	0.2040	CAMERA	Enzyme Nom**	0.2111
HomoloGene	ABA	0.1744	Enzyme Nom**	ApiDB	0.2306

* The Comprehensive Microbial Resource.
** Enzyme Nomenclature.

In addition, the experiment does not confirm the assumption that the topic similarity between two databases is low simply because they belong to two different NRA categories. For instance, in Table 5, the pair SCPD and PlantProm shows the highest topic similarity, even though the databases belong to different NRA categories. On the other hand, the pair GenePaint (Gene Expression) and COMe (Protein Sequence) shows the lowest topic similarity; GenePaint focuses on gene expression patterns in mice and COMe provides information on metalloproteins and other complex proteins, using the concept of bioinorganic motif.

5. Conclusion

In the era of "big data," science – especially biomedical science – is becoming more data-driven and data-intensive. Research breakthroughs are highly dependent upon access to valuable databases, innovative integration of different databases, and advances in sharing and maintaining high-quality curated databases. Understanding how databases are curated, used, reused, archived, and integrated is crucial to developing best practices in creating and sharing databases. Previously, the evaluation of database usage was done primarily using surveys, which are limited and qualitative. Nowadays, with open access to most medical publications, it is possible to study database usage in a quantitative way and to address this issue on a large scale.

This study provides a quantitative way to analyze database usage by identifying main paths in database link networks. A database link network was generated based on whether these databases were mentioned in the full-text methodology sections of PubMed Central articles and their cited PubMed Central references, in the domain of bioinformatics. The edge of the database link network represents the link relationship between two databases. The weight of the edge was based on: link frequency of two databases (i.e., in a link-weighted database link network) and topic similarity between two databases (i.e., in a similarity-weighted database link network). Database usage of top-ranked databases (e.g., their main paths) was examined using modularity-based component analysis, diffusion patterns of subject categories assigned to databases, and keywords extracted from methodology sections of the full-text articles mentioning these databases. The major difference between these two networks is the degree of database overlap between paths. The main-path approach yielded five paths, consisting of 202 databases for the link-weighted database network, where 49 out of the 202 databases are non-overlapping (24.6%). On the other hand, the top five main paths in the similarity-weighted database link network produced 49 unique databases out of 114 (42.1%). This indicates that the similarity-weighted database link network makes it possible to discover different link patterns and avoid the bias of focusing on frequently cited databases. In addition, the similarity-weighted database link network shows better differentiation of network components made by the five paths. Due to little overlapping of paths, clear separation among components is observed, and the link flow between components is also more obvious.

This study demonstrates that utilizing information embedded in publications allows us to trace the usage patterns of databases. Several points are worthwhile to explore in the future: (1) Database usage patterns: We can extend similar analysis to other databases beyond the top five popular databases, to identify common usage patterns; (2) Dynamic evolution data usage: We can divide the time range of bioinformatics articles into several small sections and analyze their evolution patterns; (3) Subgraphs of two databases: We can generate subgraphs of two databases to see how they are connected, which will allow us to identify critical databases that connect these two databases, and (4) to study the dynamics of databases and their subject areas can help us better understand the trajectory of development of a field. Studying the diversity of subject categories in various subgraphs of any two databases will provide useful information to see whether trans-subject-category database link plays an important role in enabling scientific innovation.

Acknowledgements

This work, done as part of the project "Cooperation Analysis of Technology Innovation Team Member Based on Knowledge Network—Empirical Evidence in the Biology and Biomedicine Field" (Grant Number: 71103114), was supported by National Natural Science Foundation of China; and also supported partly by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Ministry of Science, ICT & Future Planning (Grant Number: 2013M3A9C4078138); and also supported by National Science Foundation (Grant Number: NSF 1158670).

References

Blondel, V. D., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, (10), P10008. http://dx.doi.org/10.1088/1742-5468/2008/10/P10008 (12pp)

Bolser, D. M., Chibon, P. Y., Palopoli, N., Gong, S., Jacob, D., Del Angel, V. D., et al. (2012). MetaBase—The wiki-database of biological databases. *Nucleic Acids Research*, *40*(D1), D1250–D1254.

Bornmann, L., & Leydesdorff, L. (2013). Macro-indicators of citation impacts of six prolific countries: In cites data and the statistical significance of trends. *PLoS One*, *8*(2), e56768.

Brazas, M. D., Yim, D. S., Yamada, J. T., & Ouellette, B. F. F. (2011). The 2011 bioinformatics links directory update: More resources, tools and databases and features to empower the bioinformatics community. *Nucleic Acids Research*, *39*, W3–W7.

Carley, K. M., Hummon, N. P., & Harty, M. (1993). Scientific influence—An analysis of the main path structure in the journal of conflict-resolution. *Science Communication*, *14*(4), 417–447.

de la Pena, J. A. (2011). Impact functions on the citation network of scientific articles. *Journal of Informetrics*, *5*(4), 565–573.

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, *5*(1), 187–203.

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L. L., et al. (2013). Entitymetrics: Measuring the Impact of Entities. *PLoS One*, *8*(8), e71416.

Ding, Y., Yan, E. J., Frazho, A., & Caverlee, J. (2009). Page rank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, *60*(11), 2229–2243.
Eccles, K. E., Thelwall, M., & Meyer, E. T. (2012). Measuring the web impact of digitised scholarly resources. *Journal of Documentation*, *68*(4), 512–526.
Fernandez-Suarez, X. M., & Galperin, M. Y. (2013). The 2013 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Research*, *41*(D1), D1–D7.
Hawley, D. R., & Geske, S. (2000). The use of theory in family therapy research: A content analysis of family therapy journals. *Journal of Marital and Family Therapy*, *26*(1), 17–22.
Hu, C. P., Hu, J. M., Deng, S. L., & Liu, Y. (2013). A co-word analysis of library and information science in China. *Scientometrics*, *97*(2), 369–382.
Huang, H., Andrews, J., & Tang, J. (2012). Citation characterization and impact normalization in bioinformatics journals. *Journal of the American Society for Information Science and Technology*, *63*(3), 490–497.
Hummon, N. P., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks*, *11*(1), 36–63.
Jie, T., Ruoming, J., & Jing, Z. (2008). A topic modeling approach and its integration into the random walk framework for academic search. In *Data mining, 2008. ICDM '08. Eighth IEEE international conference on* (pp. 1055–1060).
Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 79–86.
Lu, L. Y. Y., & Liu, J. S. (2013). An innovative approach to identify the knowledge diffusion path: The case of resource-based theory. *Scientometrics*, *94*(1), 225–246.
Lucio-Arias, D., & Leydesdorff, L. (2008). Main-path analysis and path-dependent transitions in HistCite (TM)-based historiograms. *Journal of the American Society for Information Science and Technology*, *59*(12), 1948–1962.
Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods of Information in Medicine*, *40*(4), 346–358.
Medina, C. M. C., & van Leeuwen, T. N. (2012). Seed journal citation network maps: A method based on network theory. *Journal of the American Society for Information Science and Technology*, *63*(6), 1226–1234.
Mons, B., van Haagen, H., Chichester, C., Hoen, P.-B. T., den Dunnen, J. T., van Ommen, G., et al. (2011). The value of data. *Nature Genetics*, *43*(4), 281–283.
Pettigrew, K. E., & McKechnie, L. (2001). The use of theory in information science research. *Journal of the American Society for Information Science and Technology*, *52*(1), 62–73.
Reichhardt, T. (1999). It's sink or swim as a tidal wave of data approaches. *Nature*, *399*(6736), 517–520.
Sun, Y. Z., & Han, J. W. (2013). Meta-path-based search and mining in heterogeneous information networks. *Tsinghua Science and Technology*, *18*(4), 329–338.
Tan, F., Li, L., Zhang, Z. Y., & Guo, Y. L. (2013). Latent co-interests' relationship prediction. *Tsinghua Science and Technology*, *18*(4), 379–386.
Urquhart, C., & Dunn, S. (2013). A bibliometric approach demonstrates the impact of a social care data set on research and policy. *Health Information and Libraries Journal*, *30*(4), 294–302.
Vieira, E. S., & Gomes, J. (2010). A research impact indicator for institutions. *Journal of Informetrics*, *4*(4), 581–590.