Chapter 1

Networks of genetic loci and the scientific literature

J.R. Semeiks, L.R. Grate, I.S. Mian,
Life Sciences Division (MS 74-197),
Lawrence Berkeley National Laboratory,
Berkeley, CA 94720

A largely under-explored and unexploited area is biological information graphs: undirected graphs in which nodes correspond to genetic loci ("genes" for brevity) and an edge signifies that the two connected loci are discussed in the same article(s) in the scientific literature ("documents"). Operations that utilize the graph topology can assist biomedical researchers in the scientific discovery process, for example, a shortest path between two nodes defines an ordered series of genes and documents that can be used to explore the relationship(s) between genes of interest. This work (i) describes how topologies in which edges are likely to reflect genuine relationship(s) can be constructed from human-curated corpora of genes annotated with documents (or *vice versa*), and (ii) illustrates the potential of biological information graphs in synthesizing knowledge in order to formulate new hypotheses and generate novel predictions for subsequent experimental study. In particular, the well-known LocusLink corpus is used to construct a biological information graph consisting of 10,297 nodes and 21,910 edges. The large-scale statistical properties of this gene-document network suggest that it is a novel example of a power-law network. The segregation of genes on the basis of species and encoded protein molecular function indicate the presence of assortativity, the preference for nodes with similar attributes to be neighbors in a network. The practical utility of a gene-document network is illustrated by using measures such as shortest paths and centrality to analyze a subset of nodes corresponding to genes implicated in aging. Each release of a curated biomedical corpus defines a particular static graph. The topology of a gene-document network changes over time as curators add and/or remove nodes and/or edges. Such a dynamic, evolving corpus provides both the foundation for analyzing the growth and behavior of large complex networks and a substrate for examining the sociology and history of biological research.

Introduction

Graph theory provides a formal framework for specifying and modeling the relationships amongst a set of objects. In the real-world, undirected and directed graphs involving large numbers of nodes and edges have been employed to represent and investigate social, information, biological and technological networks (for a recent elegant review, see [4]). Since such graphs may contain $\sim 10^3 - 10^9$ nodes, statistical methods have been developed to quantify and characterize different large-scale properties of such complex networks. In biology, most efforts have focused on metabolic, gene regulatory, and protein interaction networks in which nodes are equated with molecules and edges denote genetic or biophysical associations. Biological information graphs have received less attention despite, for example, their unique capacity to connect nodes representing genes from different species. As shown here, networks capturing the relationships between "genes" and "documents" will emerge as important research topics both from a theoretical perspective and because of their practical value.

A study may uncover genes of interest to a biomedical investigator but with which they are unfamiliar. Such a scenario is increasingly common and one especially associated with high-throughput, large-scale molecular profiling technologies. For example, a cancer-related transcriptional profiling study may highlight the importance of the genes telomerase reverse transcriptase (TERT) and GATA binding protein (GATA1). To elucidate possible relationships between genes, the prevailing approach is simple keyword searches of PubMed, the online biomedical literature database maintained by the National Center for Biotechnology Information (NCBI) and consisting of over 14 million citations from MEDLINE and additional life science journals[1]. However, a PubMed query using the string "TERT GATA1" is uninformative. Thus, the information retrieval problem becomes one of discovering indirect links between genes, *i.e*, that two specific PubMed citations relate TERT to Sp1 transcription factor (SP1) and another relates SP1 to GATA (Figure 1.1).

Given a biological information graph, the information retrieval problem can be recast as a simple, standard convex optimization problem: find the shortest path(s) between the nodes corresponding to TERT and GATA1. The nodes and edges linking these genes define an ordered sequence of genes and published studies that can assist in discovering the relationship(s) between TERT, GATA1, and cancer, *i.e.*, the graph organizes extant information in such a manner that it facilitates the formulation of biological insights and the generation of predictions.

Graphs can provide a visual representation of a corpus of annotated data. Thus, a network where nodes are identified with genes and edges with documents can be constructed from a corpus of documents annotated with genes or,

[1]URLs discussed: PubMed, `http://www.ncbi.nlm.nih.gov/pubmed`; LocusLink, `http://www.ncbi.nlm.nih.gov/LocusLink`; WormBase, `http://www.wormbase.org`; SGD, `http://www.yeastgenome.org`; FlyBase, `http://www.flybase.org`; Gene Ontology (GO), `http://www.geneontology.org`; SAGEKE, `http://sageke.sciencemag.org`; HuGE, `http://www.cdc.gov/genomics/hugenet`; R, `http://www.r-project.org`; Graphviz, `http://www.research.att.com/sw/tools/graphviz`.

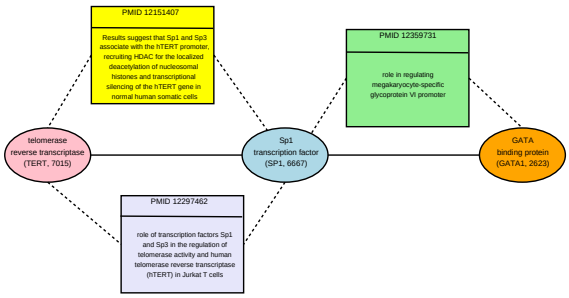


Figure 1.1: A visual depiction of three PubMED citations (squares) that link the three genes TERT, SP1 and GATA (circles). The free-text statement in a box is a synopsis of the article with the stated standard PubMed identifier (PMID).

equivalently, genes annotated with documents. For example, natural language processing and other text analysis techniques have been used to determine associations between given genes/proteins and specific MEDLINE abstracts (reviewed in [2, 9]), *i.e.*, such efforts generate corpora of documents annotated with genes. However, a critical aspect of these "automated" approaches is a requirement to solve the problem of entity recognition, ascertaining whether the abstract contains a reference to a specific gene. This task remains a challenge, not least because of the complexity of biological terminology [8]. The frequency of erroneous edges (false positives) and missed edges (false negatives) in gene-document networks produced automatically, although unknown, is likely to be non-negligible. One consequence of this is a reduction in the reliability and accuracy of shortest path calculations and other analyses. By transferring the burden of annotation to humans, the veracity of edges in a gene-document network should increase but this would be at the expense of added time and labor.

This work shows that by viewing extant widely-available biomedical resources such as LocusLink WormBase, SGD, and FlyBase as human-curated corpora of annotated data, gene-document networks can be produced that emphasize quality rather than quantity. Specifically, a biological information network is derived from LocusLink, examined in terms of its large-scale statistical properties, and exploited to enhance understanding of genes implicated in aging. Finally, open theoretical and applied problems raised by this study are discussed.

Materials and Methods

LocusLink gene-document (GeneRIF) network

LocusLink is a text-based, non-redundant, comprehensive, catalog of information on genetic loci created by the NCBI in 1999. LocusLink entries ("loci") are maintained via a process that includes both automated computational methods and manual data curation. An entry may include any number of GeneRIFs, literature annotations describing the locus' structure or function and assigned by professional NCBI indexers and (to a lesser extent) the scientific community. A GeneRIF consists of the PMID for a relevant article and a free-text synopsis of the pertinent article. Since every locus is associated with zero or more GeneRIFs, LocusLink can be viewed as a corpus of "genes" annotated with "documents". A

GeneRIF network is a LocusLink-derived biological information graph in which nodes correspond to genetic loci and an edge denotes that the connected genes are annotated with a GeneRIF referencing the same PMID. The absence of an edge between genes denotes either a lack of knowledge or no genuine relationship.

The GeneRIF network constructed and analyzed here was derived from the April 2004 release of LocusLink and contained 181,380 loci from 14 model organisms (the most represented were mouse, *M. musculus*; human, *H. sapiens*; fly, *D. melanogaster*; rat, *R. norvegicus*; worm, *C. elegans*; and zebrafish, *D. rerio*). Genes with no GeneRIFs or whose GeneRIF PMIDs were not shared by any other gene(s) were excluded resulting in a final network with no single isolated. All research was performed using LocusLink `LL_tmpl` and GO flat files housed in a custom relational database, the R statistics package, the C++ Boost Graph Library, and GraphViz.

Species and molecular function assortativity

Nodes can be characterized using attributes that take the form of scalars, vectors, graphs and so on. Assortativity refers to the preference of nodes with similar attributes to be neighbors in a network. Here, the focus is such selective linking on the basis of organismal origin of genes (species assortativity) and biophysical properties (functional assortativity). The notion of preferential association can be quantified via an assortativity coefficient, $0 \leq r \leq 1$, computed using a normalized symmetric mixing matrix (Eq (17) in [4]). For species assortativity, an element of the mixing matrix specifies the fraction of edges that connect genes from two given species.

Assessing functional assortativity is more challenging because proteins can have multiple functions, distinct functions are not necessarily independent, and functional annotation is incomplete. The approach used here exploits ongoing efforts to characterize gene products via the assignment of terms from an extensive controlled vocabulary known as the Gene Ontology (GO). Each GO term belongs to one of three orthogonal aspects (`Molecular Function`, `Biological Process`, `Cellular Component`) and the ontology itself is organized as a directed acyclic graph (DAG). The GO terms assigned to a protein are available in the LocusLink entry for the corresponding genetic locus. Given this source of node-specific attributes, the task of calculating functional assortativity becomes one of computing the similarity between nodes in a gene-document network given attributes that themselves take the form DAGs. Establishing similarities among the nodes of GO DAGs should account for all possible paths connecting the nodes and for the lengths of those paths. A simple similarity score was estimated using an approach developed to take into consideration the GO graph structure and the frequency of GO terms assigned to genes [3, 7]. For the GeneRIF network, functional assortativity was assessed by computing pairwise GO molecular function semantic similarity scores using only GO `Molecular Function` terms accompanied by a `Traceable Author Statement` Evidence Code, *i.e.*, GO assignments of high-confidence since they reflect knowledge present in the primary literature.

Gene-document network-based analysis of aging-related genes

To illustrate how a gene-document network may be used to study an arbitrary collection of "interesting" genes and to discover new genes that may have roles in the same phenomenon and/or process, the SAGEKE database of known aging-related genes was analyzed using the following heuristic. The shortest path (SP) between two nodes in a network can be computed using standard algorithms such as breadth-first search [1]. An SP for a pair of genes in a collection defines a sequence of genes in which neighbors are related via knowledge found in the scientific literature. The union of shortest paths (USP) subgraph is the set of nodes and edges defined by the SPs for all distinct node pairs in a collection. Centrality refers to the influence a node has over the spread of information in a network. A simple measure of this notion is stress centrality, the number of SPs between node pairs that pass through a node of interest. Given nodes in the USP subgraph ranked by their stress centrality score, good candidates for genes for additional study are nodes with high scores but that are not present in the original collection. In the absence of sound methods for computing centrality in multi-component networks, stress centrality was computed for SAGKE USP genes in the GeneRIF giant component.

Unweighted and weighted networks

The specific weights assigned to edges affect the SP found in a graph. In a gene-document network with uniform edge-weights (unweighted network), the number of documents common to two genes is ignored. In a network with non-uniform weights (weighted network), both the number of shared documents and the number of other genes annotated to the documents is considered. A weighted GeneRIF network was estimated using an approach proposed for scientific collaboration networks in which nodes correspond to scientists and an edge indicates that two scientists are co-authors [5]. The weight of an edge is the sum, over all common documents, of $1/(n-1)$, where n is the number of genes associated with a document. This document-independent weighting scheme could produce biologically unreasonable weights: a paper describing sequencing of genomic DNA and annotated with 10 genes would contribute the same as a paper annotated with the same genes but defined as the result of a functional genomics study.

Results

Large-scale statistical properties, including assortativity

The GeneRIF network contains 10,297 nodes and 21,910 edges organized into 1229 distinct components (two or more nodes connected transitively to each other but not to other nodes in the graph). The basic statistics of this gene-document network are similar to those networks studied previously (Table 1 in [4]; Figure 1.2). Since the degree distribution follows an approximate a power law, most genes are related to each other by a small cabal of highly-connected genes. These are human TP53, tumor protein p53 (annotated with

GeneRIF component(s)		
	All 1229	Giant
n	10,297	7,167
m	21,910	19,372
z	4.26	5.41
l	5.54	5.54
$C^{(1)}$	0.153	0.146
$C^{(2)}$	0.365	0.401
r	0.141	0.108

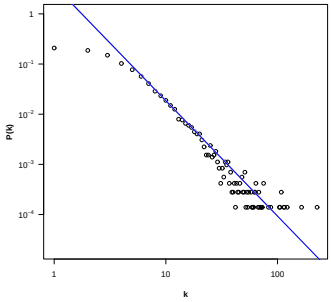


Figure 1.2: Left: Large-scale statistical properties of the unweighted GeneRIF network. n, number of nodes; m, number of edges; z, mean degree (number of edges per node, k); l, SP distance between node pairs (Eq (2) in [4]); $C^{(1)}$, clustering coefficient (mean probability that two nodes that are network neighbors of the same third node are themselves neighbors; Eq (3) in [4]); $C^{(2)}$, alternative clustering coefficient (weighs the contribution of nodes with few edges more heavily; Eq (6) in [4]); r, degree correlation coefficient. Right. Log-log plot of the degree, k, and degree probability (fraction of nodes with degree k), $P(k)$, for the giant component. The tail follows an approximate power law, $P(k) = ck^{-\alpha}$ ($\alpha \approx 2.3$, $c \approx 2.7$ for the line shown).

510 documents/number of edges or degree $k = 226$); human TNF, tumor necrosis factor (319/164); human VEGF, vascular endothelial factor (236/122); human MAPK1, mitogen-activated protein kinase 1 (124/115); human TGFB1, transforming growth factor, beta 1 (176/114); human NFKB1, nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (113/108); mouse Trp53, transformation related protein 53 (148/108); mouse Tnf, tumor necrosis factor (150/105); and human SP1, Sp1 transcription factor (72/104).

The biological properties of signaling and kinase activity distinguish well-connected genes (105 genes having the top 1% of degrees, $k \geq 33$) from all genes (7,167 genes in the giant component). For GO terms assigned to $\geq 10\%$ of genes in both sets, the ratio of the relative frequency of a GO term in well-connected genes to its relative frequency in all genes was computed. GO terms with a ratio greater than 2.0 are `MAP kinase activity`, `cell proliferation`, `cell-cell signaling`, `kinase activity`, `apoptosis`, `signal transduction`, `regulation of cell cycle`, `protein amino acid phosphorylation`, `immune response`, `protein serine/threonine kinase activity`, `protein kinase activity`, `transferase activity`, `ATP binding`, `extracellular`, `cytoplasm`, and `transcription factor activity`.

Nodes in the giant component of the GeneRIF network exhibit extensive segregation on the basis of both species and molecular function. The species assortativity coefficient of $r = 0.73$ suggests that this gene-document network is closer to a perfectly assortative network (if $r = 1$, every edge connects nodes of the same type) than a randomly mixed network ($r = 0$). The functional similarity between pairs of nodes (pairwise GO molecular function semantic similarity score) decreases as the SP distance between them increases (Figure 1.3).

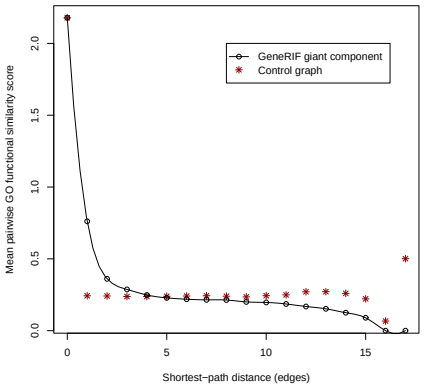


Figure 1.3: Functional assortativity of nodes in the GeneRIF network giant component and a topologically identical network generated by randomly shuffling the identities of nodes in the giant component. At distances of 1 to 4, genes in the real network are more similar to each other than might be expected (the distance of a gene to itself is defined as zero). The fluctuation at large distances is due to small sample sizes.

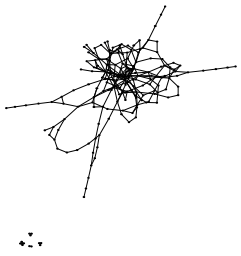
GeneRIF network: known and novel aging-related genes

In order to investigate previously identified aging-related genes and suggest new genes with potential roles in this phenomenon, a weighted GeneRIF network was used to analyze a collection of genes implicated in aging (Figure 1.4). The presence of mouse (Trp53, Igf1, Tnfsf11, Tnfsrf11b, Akt1) and fly (W) genes amongst human genes highlights a unique property of biological information graphs. Because edges can link nodes from different organisms, gene-document networks provide an ability to navigate both intra- and inter-species relationships, a facet absent in protein interaction and other biological networks.

Nodes in the SAGEKE USP subgraph that have high stress centrality scores include some genes that are highly-connected in the giant component of the GeneRIF network (large degree k): human TP53, mouse Trp53, and human VEGF. The higher score and degree of TP53 compared to its mouse homolog (Trp53) reflect the greater scrutiny to which this gene has been subjected (to date, SAGEKE has only linked the mouse gene explicitly with aging). Other highly-connected GeneRIF genes are human VEGF, human TNF and mouse TNF. One heuristic for identifying new genes with possible roles in aging is to equate them with nodes that have high centrality scores in the SAGEKE USP subgraph but are less well-connected in GeneRIF network. Using these criteria, additional studies of IGF1/Ifg1, PTEN, Tnfsf11, Tnfrsf11b/TNFRSF6, CDKN1A, CASP8, and LEP could prove informative.

Discussion

The biological information graph investigated here is similar to the literature citation network first studied four decades ago [6] in that semantic links between nodes come exclusively from published research. Gene-document networks have a variety of strengths and applications. As suggested by the analysis of aging-related genes using a LocusLink-derived GeneRIF network, they have potential as tools in the scientific discovery process. Such networks provide an individual with little expertise in given domain ready and simple access to an extensive body



Gene	Name	Centrality	k
TP53	tumor protein p53 (Li-Fraumeni syndrome)	288.2	13
Trp53	**transformation related protein 53**	136.2	6
VEGF	vascular endothelial growth factor	135.7	10
IGF1	insulin-like growth factor 1 (somatomedin C)	128.5	9
PTEN	phosphatase and tensin homolog	110.3	4
Igf1	insulin-like growth factor 1	93.3	6
W	Wrinkled	77.3	4
Tnfsf11	tumor necrosis factor (ligand) superfamily, member 11	75.5	4
Tnfrsf11b	tumor necrosis factor receptor superfamily, member 11b	74.0	2
Akt1		73.0	3
APP	**amyloid beta (A4) precursor protein (Alzheimer disease)**	71.5	8
CDKN1A	cyclin-dependent kinase inhibitor 1A	64.7	4
CASP8	caspase 8, apoptosis-related cysteine protease	61.0	5
LEP	leptin (obesity homolog)	57.8	6
TNFRSF6	tumor necrosis factor receptor superfamily, member 6	56.9	3

Figure 1.4: Left: The 185 nodes and 268 edges of the weighted GeneRIF network that constitute the union of shortest-paths (USP) subgraph for 51 known aging-related (SAGEKE) genes (note that some SAGEKE genes correspond to nodes that are not in the giant component, small clusters of nodes in the bottom left). Right: Genes in the SAGEKE USP subgraph with the 15 highest stress centrality scores. Genes in bold are SAGEKE genes.

of prior work allowing them to formulate hypotheses and generate predictions for subsquent investigation. Especially noteworthy, although rare, are edges between genes from different species because the associated documents provide a useful bridge for comparative genomics and biology studies of similar processes such as aging. By ascertaining isolated components in the GeneRIF network, NCBI indexers can be alerted to genes that might benefit from systematic efforts to identify publications that link them to other and larger components.

Gene-document networks are limited by a number of factors, not the least of which is the corpus of annotated data used to construct the network. The very origin and nature of LocusLink means that the GeneRIF network is neither comprehensive nor complete. The focus of GeneRIF is papers pertinent to basic gene structure and function rather than, for example, evolution. Because GeneRIF was initiated in 2001 as a manual curation effort by a small group of experts, only those relationships between genes in recent publications are present in the LocusLink. Some of these deficiencies may be overcome by building gene-document networks using multiple manually curated gene-centered corpora, for example, the HuGE database of published epidemiology articles on human genes.

Since the nodes in gene-document, protein interaction and related networks correspond to genetic loci, such sources of heterogeneous data could be fused to yield biological information graphs of greater scope and enhanced utility. Such an operation would provide a simple method for synthesizing disparate information. Since most biomedical resources are ongoing efforts, the size and coverage of such corpora is increasing over time. Building biological information graphs using each release would yield new real-world examples of large, complex, dynamic, networks. In addition to theoretical studies of their properties and behavior, these networks would be useful not only for researchers and clinicians, but also policy makers, historians, and sociologists interested in the evolution of different disciplines in biology.

Acknowledgements

This work was supported by the California Breast Cancer Research Program, National Institute on Aging, National Institute of Environmental Health Sciences, and U.S. Department of Energy.

Bibliography

[1] Cormen, T.H., C.E. Leiserson, R.L. Rivest, and C. Stein, *Introduction to Algorithms* 2nd ed., McGraw-Hill (2001).

[2] H., Shatkay, and Feldman R., "Mining the biomedical literature in the genomic era: an overview", *J. Computational Biology* **10** (2003), 821–855.

[3] Lord, P.W., R.D. Stevens, A. Brass, and C.A. Goble, "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.", *Bioinformatics* **19**, 10 (2003), 1275–83.

[4] Newman, M.E.J., "The structure and function of complex networks", *SIAM Review* **45** (2003), 167–256.

[5] Newman, M.E.J., "Coauthorship networks and patterns of scientific collaboration", *Proc. Natl. Acad. Sci. USA* **101** (2004), 5200–5205.

[6] Price, D.J. de S., "Networks of scientific papers", *Science* **149** (1965), 510–515.

[7] Resnik, P., "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language", *J. Artificial Intelligence Res.* **11** (1999), 95–130.

[8] Rzhetsky, A., I. Iossifov, T. Koike, M. Krauthammer, P. Kra, M. Morris, H. Yu, P.A. Duboue, W. Weng, W.J. Wilbur, V. Hatzivassiloglou, and C. Friedman, "GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data", *Journal of Biomedical Informatics* **37** (2004), 43–53.

[9] Yandell, M.D., and W.H. Majoros, "Genomics and natural language processing", *Nature Reviews Genetics* **3** (2002), 601–610.