Data Journeys Beyond Databases in Systems Biology: Cytoscape and NDEx

William Bechtel

Abstract This chapter investigates how data travels beyond databases in cell biology by focusing on Cytoscape, a platform that has been developed to represent networks, and NDEx, a database that allows for the reuse of network representations. I begin with a brief review of the databases that have been developed for data involving, for example, protein-protein interactions, that are relational and hence productively represented in networks. Given the amount of data stored in modern databases, raw network representations are typically hairballs that provides researchers little useful information other than that lots of things interact. Cytoscape was created by systems biologists to facilitate moving beyond hairballs to informative representations. It provides tools for clustering nodes and annotating them according to what is known about the objects represented. I provide examples of how Cytoscape has been deployed to develop new knowledge about biological mechanisms. Cytoscape has been made freely available, and I describe how a large interational community of researchers has created Apps that enable researchers to make a number of more specialized inferences. NDEx, created by members of the same research lab, serves as an Expo for networks—researchers can share networks they have developed and other researchers can search for networks and made them the basis for further incorporation of data or analyses.

1 Introduction

As in many fields, contemporary biologists generate vast amounts of data. Increasingly, this data is stored in large, on-line databases that procure data from curation of published literature and from high-throughput experiments. There it is accessed by researchers distinct from those who produced the data. Leonelli (2016; this volume) has developed the useful metaphors of *data travel* and *data journeys* to characterize this process of data movement. Much of the work on data journeys to

W. Bechtel (✉)
Department of Philosophy, University of California, San Diego, CA, USA
e-mail: bill@mechanism.ucsd.edu

© The Author(s) 2020

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_7

date has focused on the preparation and travel of the data themselves, with less attention paid to the resources that are employed to analyze the data after they travel.[1] When the data specifies relations (causal, co-occurrence, etc.) between entities, this analysis often involves the construction of network diagrams in which entities are represented as nodes and relations between them as edges.[2] In the course of research, network diagrams are subject to various manipulations designed to reveal additional patterns in the data. Beyond their use in individual research projects, these networks themselves travel, providing the foundation for yet other research projects in which they are subject to further manipulation. Network diagrams are one format in which data are physically instantiated and subject to mutation as they are incorporated into network diagrams and passed on to other researchers (see Leonelli, this volume, for discussion of how data are mutated in the course of data journeys).

My focus will be on the tools that systems biologists have created to construct and operate on network diagrams and to enable networks themselves to travel. Anyone could construct a network diagram by hand from a body of data using a standard graphics package. However, such a process is laborious and the product is frozen—the researchers cannot then integrate data from additional sources or transform the diagram to reveal new patterns. Accordingly, researchers have developed software tools for creating, analyzing, and disseminating network diagrams. In Sect. 4 I will discuss Cytoscape, the most widely used platform for constructing network diagrams in systems biology. While developed in a systems biology framework, Cytoscape has itself traveled to and is actively used in numerous other scientific fields. Cytoscape provides a platform on which researchers with specific analytic needs can develop their own add-ons, referred to as apps. In Sect. 5 I will describe several apps and, using them, illustrate some of the analysis strategies employed in systems biology. In Sect. 6 I will describe the recent development of NDEx, which serves as an online exposition (expo) to which networks themselves can travel so as to be viewed by others and selectively taken up for additional journeys. As a background for focusing on network diagrams, I begin in Sect. 2 by introducing the types of data used to construct network diagrams in systems biology and in Sect. 3 describe the public databases and ontologies from which researchers extract data to create and analyze networks.

[1] Leonelli (2016, chapter 6) provides a pioneering examination of reuse. See chapters by Tempini, Chap. 13, Morgan, Chap. 6, and Griesemer, Chap. 8 in this volume, for other aspects of reuse. Tempini addresses the reuse of data for different objectives than that for which it was collected, and in particular focuses on how this often involves linkage of data from different sources such as between weather, environment, and health data. As he demonstrates, this requires manipulations that attenuate the differences due to where the data originated.

[2] Networks are just one mode of downstream analysis of data. See Cambrosio et al., this volume, for an account of knowledgebases that tailor large datasets for particular clinical applications.

2 Data Production: From Individual Experiments to High-Throughput Experiments

Through most of the twentieth century, experiments in fields like cell and molecular biology were conducted one at a time. But many of the procedures used in these experiments lent themselves to automation so that multiple variants on an experiment could be conducted in parallel. For example, when Sanger first developed techniques for sequencing amino acids in the 1950s or nucleic acids in the 1970s, he applied them to one protein or gene at a time. By the late 1980s these techniques were automated and by the 1990s automation made possible the sequencing of whole genomes of numerous species. Sequencing data identifies proteins and genes, but not what they do. Automated procedures enabled procuring other types of data related to function such as techniques that reveal whether proteins form complexes either with other proteins or with DNA or whether genetic mutations interact epistatically. I discuss only techniques detecting whether proteins can form complexes.

Much of the early twentieth century research focused on the reactions individual proteins catalyze, but in the second half of the twentieth century it became increasingly evident that proteins form complexes with each other and these are important to their catalytic function. Two techniques have proven especially useful in enabling high-throughput studies of protein-protein interactions (PPIs). The first, the yeast two-hybrid technique introduced by Fields and Song (1989), begins by transfecting yeast cells with two plasmids, each attaching to a different protein. One serves as the bait and the other as the prey and when the proteins to which they are attached interact with each other, the two domains are united and form a functional transcription factor that initiates transcription of a reporter gene. This technique identifies pairs of proteins that *can* interact, but many pairs do not do so in a given cell type. An alternative technique, affinity purification followed by mass spectrometry, starts with proteins that are actually bound into a complex in a cell and uses mass spectrometry to determine their identity (Rigaut et al. 1999). This approach identifies stable multi-protein interactions that actually occur in the cell. On the other hand, it misses more transient interactions that form and dissolve as cells carry out activities. As a result, both approaches to obtaining PPI data are actively employed.

High-throughput techniques for performing PPI studies were created shortly after automated gene sequencing was introduced and provided a means to study many of the novel genes they revealed. In the first high-throughput attempt to identify PPIs in yeast, Uetz et al. (2000) chose 192 proteins to use as baits and mated them with 6000 prey proteins. They identified 957 interactions between 1004 proteins. The following year Ito et al. (2001) performed an even larger-scale study, identifying 4549 interactions between 3278 proteins. Surprisingly, there was little overlap with the interactions identified in these two studies. I return to the Uetz et al. and Ito et al. studies to show how they were used in a pioneering network study in the next section.

3 Data Travels in Systems Biology: Databases and Ontologies

As biologists generated increasing volumes of data, they established publicly accessible databases to make this data accessible. The first databases were created for protein and gene sequence data. Dayhoff created the *Atlas of Protein Sequence and Structure* (Dayhoff and Eck 1965-1972) which she published in book form. Shortly after her death in 1984 it was made available electronically as the Protein Information [originally Interaction] Resource's Protein Sequence Database. It eventually merged into UniProt, which continues as a major source of information about proteins (The UniProt Consortium 2017). GenBank was developed in the same period for gene sequence data. Many additional databases for different types of biological data soon appeared—in 1989 the Listing of Molecular Biological Databases identified 50 databases (Lawton et al. 1989) and the number has continued to grow ever since. Annually, the first issue of *Nucleic Acids Research* reviews new and updated databases. On its website it provides a compilation of current databases, totaling 1613 in 2019. As Leonelli (this volume) notes, this process is both uncontrolled and unsustainable. In fact, each year the *Nucleic Acids Research* compilation annually eliminates discontinued URLs, including 147 in 2019.

Two of the early databases to include PPI data were the Yeast Proteome Database (YPD) and the Martinsried Institute for Protein Sequences (MIPS) database of protein interactions. A study by Schwikowski et al. (2000) illustrates how these databases were employed to construct a network from which new knowledge about yeast was extracted. They combined data from YPD and MIPS with data from the two high throughput studies noted at the end of the last section, yielding information on 2709 interactions involving 2039 proteins. Employing hierarchical clustering based on functional assignments found in the YPD and a layout procedure that located similarly connected nodes near each other, Schwikowski et al. identified one large connected network, shown in Fig. 1, plus 203 much smaller networks. In cases in which YPD contained information about a protein's cellular role, the researchers encoded it using the color of nodes: blue for membrane fusion, grey for chromatin structure, green for cell structure, yellow for lipid metabolism, and red for cytokinesis. By zooming in on parts of the network, as in panel B, they could focus on interactions between proteins that performed similar cellular roles, in this case membrane fusion, lipid metabolism, and cell structure.

An important question about any network diagram is whether the patterns it reveals are informative or simply an artifact of the representation strategy the researchers employed. To investigate this, Schwikowski et al. started with a given node to which a cellular role was assigned and asked how often one of the nodes with which it was connected in the network was assigned the same cellular role. This happened 72% of the time, (compared with, on average, 12% for scrambled networks). The authors present this as vindicating the network—had they not known the cellular role of the initial protein, they could have predicted it correctly 72% of the time based on the roles of its neighbors.

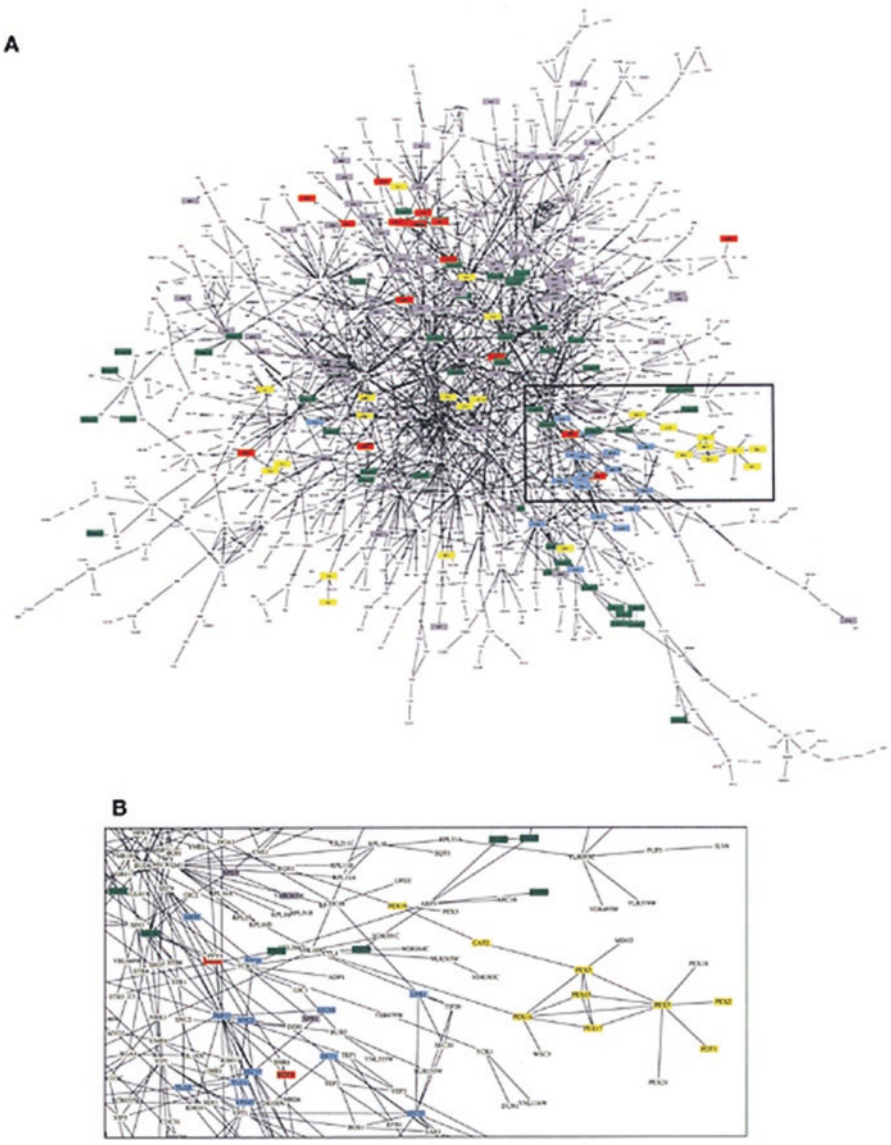

Fig. 1 Network diagram of protein interactions in yeast constructed by Schwikowski et al. 2000 drawing both upon results of high-throughput yeast two-hybrid studies and data from low-throughput studies collected in the MIPS and YPD databases. Reprinted by permission from Springer Nature: *Nature Biotechnology*, A network of protein-protein interactions in yeast, Schwikowski et al. 2000

As researchers recognized the usefulness of drawing upon large datasets in their research, many researchers created their own databases, tailored to their interests, and made them publicly available. These included the Database of Interacting Proteins (DIP) (Xenarios et al. 2000), MINT (Zanzoni et al. 2002), BIND (Alfarano et al. 2005), HPRD (Peri et al. 2003), BioGRID (Breitkreutz et al. 2003a), and IntAct (Hermjakob et al. 2004b). The infrastructure for each was relatively small—on average, they employed two full-time curators who read published papers and entered the data. In addition to primarily serving the interest of a particular laboratory, each database developed its own data structures and procedures for downloading and curating data. No single database could keep up with the rapid appearance of new datasets. As a result, researchers who wanted to use PPI data often combined data from multiple databases, developing their own tools (parsers, etc.) to do so. Recognizing the problem users faced, the curators of several databases collaborated to develop a standardized format (Hermjakob et al. 2004a). A standard format, however, made another problem even more salient. In reporting data, journal articles often failed to supply sufficient information about the entities studied or the experimental procedure used. This information is crucial for others to use and interpret the data (see Leonelli 2016, chapter 4; Rogers and Cambrosio 2007; and Boumans and Leonelli, this volume). Accordingly, the consortium generated guidelines as to the minimal information required in reporting a PPI experiment (Orchard et al. 2007). Several of the databases also began to work directly with journals so that data in new publications could be directly added to the databases. These efforts ultimately led to the development of the International Molecular Exchange (IMEx) Consortium, which among other initiatives introduced a deep curation standard aiming "to capture the full experimental detail provided in the interaction report, as this is often essential to assess interaction context and confidence" (Orchard 2012, p. 347). The initiative also sought to address another problem, that of maintaining funding for the various databases. The IMEx consortium also provided that if a member can no longer curate its databases, its records would be turned over to another member. Accordingly, when MPIDP ceased its curation efforts in 2012, it turned its records over to IntAct, which has subsequently maintained and updated them.

PPI databases have provided the data for constructing networks, but another database created during the same period, Gene Ontology (GO), has played a crucial role in allowing biologists to interpret networks. The motivation for developing GO was to develop "a structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products" (Ashburner et al. 2000, p. 26) represented in the databases that had been developed for different model organisms (initially yeast, fruit fly, and mouse). GO comprises three ontologies, one for biological processes, another for molecular functions, and a third for cellular components, each providing general terms, organized hierarchically, that can be used to annotate individual genes. These ontologies are themselves undergoing continual revision and development (Leonelli 2010, 2016).

By 2000 systems biologists had a rich set of databases on which they could draw. Some, such as GenBase and UniProt, emphasized structural knowledge, but many focused on relational information, including PPI data. GO provided a common lan-

guage for annotating the entries in the different databases. These are the raw materials from which systems biologists constructed network diagrams with the goal of developing new biological knowledge.

4 Cytoscape: A Platform for Generating and Analyzing Network Diagrams

Tables in databases are great for storing and organizing data, but it is often difficult for humans to examine data tables directly and draw biologically meaningful inferences or even figure out what algorithms they might employ to generate inferences.[3] For this reason, most of the databases include a self-developed program to display the results of searches as network diagrams. These, however, typically employ a fixed format designed by the curators of the database.[4] Individual network formats support some inferences but not others. In order for users to leverage the vast amount of data contained in these databases, they need to generate network representations appropriate for their needs (see Leonelli, this volume, for a discussion of the relational nature of data).

Although several programs for creating network diagrams, including Osprey, VisANT, Gephi, and GraphViz, were developed in the first decade of the twenty-first century, Cytoscape (Shannon et al. 2003) has emerged as the most widely used. Ideker and his collaborators at the Institute for Systems Biology began developing Cytoscape in late 2001 for their own research and publicly released Cytoscape 0.8 as an opensource platform in June 2002. When Ideker moved to the University of California, San Diego, it became the center for Cytoscape development. The local team of 3–5 developers collaborates with numerous other developers at other institutions (currently including the Academic Medical Center in Amsterdam, the Institute for Systems Biology, the Institute Pasteur, the Gladstone Institute, the University of California, San Francisco, and the University of Toronto).

Although it is hard to measure actual use, in 2018 Cytoscape was downloaded on average 17,600 times per month and started on users' computers about 5000 times each day. According to Google Scholar, the standard reference used to acknowledge Cytoscape, Shannon et al. (2003), has been cited more than 14,750 times as of September 2019, most often by papers that include a network diagram generated with Cytoscape. These numbers likely significantly underestimate how frequently Cytoscape is used since many users do not explicitly acknowledge it (just as most people do not acknowledge Microsoft Excel or Adobe Illustrator even if they made extensive use of these in their research).

[3] Tables, though, sometimes enable viewers to visualize data. See Müller-Wille and Porter (this volume) for examples.

[4] The exception is BioGRID, whose developers also created Osprey, a network visualization program (Breitkreutz et al. 2003b). However, development of Osprey has ended and its webpage suggests researchers use Cytoscape.

Cytoscape, now in version 3.7.1, is an open-source, freely available java-based software package that runs on individual computers. It is a key platform of the National Resource for Network Biology and its development team continues to add new features to facilitate investigations directed at a range of topics such as representing networks at multiple scales and representing dynamic changes in cellular network organization in disease. An even larger community of computationally oriented biologists from around the world generates apps (initially referred to as plug-ins) that extend Cytoscape's capacities for analyzing networks. These are made available through the Cytoscape App Store, hosted on the Cytoscape website (http://cytoscape.org). In this section I will describe how Cytoscape is used to construct and modify network diagrams. In the subsequent section I will discuss apps and how they support analyses of networks.

Figure 2 provides a schematic overview of the Cytoscape architecture. The Cytoscape Window contains both the tables of node and edge attributes, from which Cytoscape constructs the network diagram, and the network diagram itself. Other components operate on the tables and graphs. I will not elaborate on the Graph Editing and Selection component. It performs functions much like those contained in the File and Edit components of word processing programs: opening stored networks or creating new ones, selecting, deleting or hiding, or copying nodes or edges, etc.

Visual Mapper (later termed VizMapper and in Cytoscape 3.5 renamed Style) and the Layout Engines take their input from the Node and Edge Attribute Tables. An Edge Attribute Table is shown in the screenshot in Fig. 3; a similar table defines

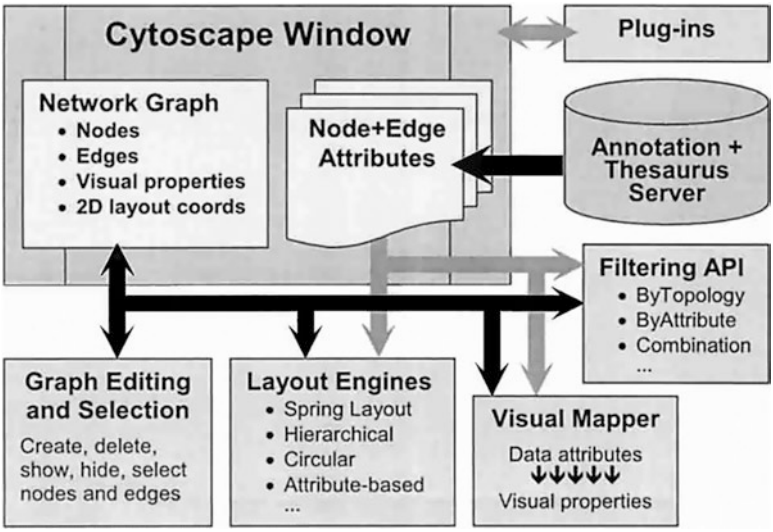

Fig. 2 Schematic overview of the Cytoscape architecture reprinted from Shannon et al. 2003. Although the labels for some of the components have changed, the overall architecture has not. Reprinted with permission of Trey Ideker

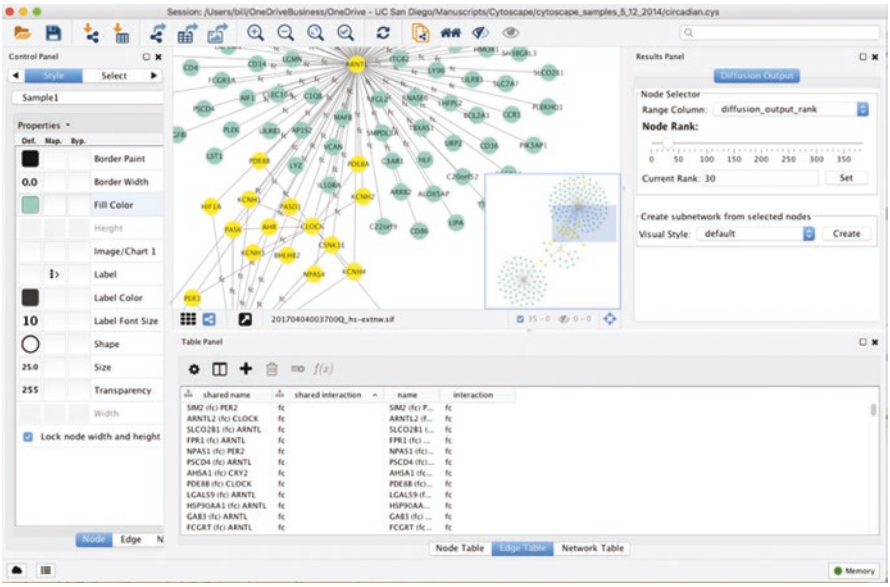

Fig. 3 Screenshot of Cytoscape 3.5. The window at the bottom shows the Edge Table from which the diagram in the upper window is generated. The window on the left shows the assignments of visual properties to nodes in Style. Screenshot used with permission of Trey Ideker

the nodes. A researcher can generate these tables based on data he or she has collected or from data downloaded from one or more of the databases discussed in the previous section. At a minimum, these tables must identify the entities to be represented by the nodes and the relations to be represented by edges, but they may also identify a variety of attributes of the entity (e.g. its concentration) or relation. The tables can also include annotations (e.g., cell location or cell function) procured from sources such as GO.

Style, shown on the left in Fig. 3, maps features specified in the table unto visual properties of nodes and of edges. Thus, an investigator can map attributes or annotations specified in the node and edge tables to labels or to visible features such as shape, size, and color. If color, for example, is used to indicate biological processes as specified in GO and size is used to represent the level of expression of a gene, the viewer can quickly see patterns in how these attributes and annotations vary.

There are many ways to lay out nodes in a 2-dimensional representation—nodes can be positioned randomly, around a circle, in a grid, or in a hierarchical arrangement. It is often useful to group nodes by their values on a particular annotation such as biological process or cellular component. When used with a circular layout, this results in nodes that share an attribute being located close together around the circle. There is great flexibility in how nodes are laid out and the choice affects what patterns the researcher can identify. For example, it is easier to see that several nodes are highly interconnected or are all connected to another set of nodes when they are positioned near each other. Spring-embedded layouts do this by treating edges like

springs (Eades 1984): connected nodes that are far apart are drawn together, but if they get too close, they are repelled a bit. For each of these strategies for laying out nodes there are a variety of algorithms, each of which generates a somewhat different result. After an algorithm is applied, the user can also manually move one or a selected group of nodes. Researchers find it useful to try out different layout strategies to find one that generates interpretable patterns.

Since the goal of network analysis is to generate biologically interpretable results, researchers derogatorily refer to networks such as shown in Fig. 4a as *hairballs*. Although the data is represented, it is not presented in a manner that can be interpreted biologically. Merico et al. (2009) illustrate how, by altering visual features and layout in Cytoscape, to transform this hairball into an informative network diagram revealing components of mechanisms involved in chromosome maintenance and duplication in yeast (Fig. 4b). Figure 4a was generated from curated data of PPIs (represented as edges) from both low- and high-throughput experimental studies retrieved from BioGRID. The nodes represent proteins and their colors indicate their location in the chromosome: red, replication fork; green, nucleosome; blue, kinetochore; yellow, other chromosome components. The use of color in Fig. 4a is already a step away from a pure hairball, but the network diagram offers no mechanistic insight. By applying a spring-embedded layout in which edges are assigned forces so as to draw highly connected nodes closer together and yet keep them from getting too close, the authors transformed Fig. 4a into 4b. Being highly connected, the nodes for proteins in the kinetochore, nucleosome, and replication fork are now situated adjacent to each other. VizMapper (Style) used data about how much gene expression changes over the cell cycle to determine node size. In addition, the width of the edges is determined by the Pearson correlation between transcript profiles. Looking at the network diagram one can readily see that many green

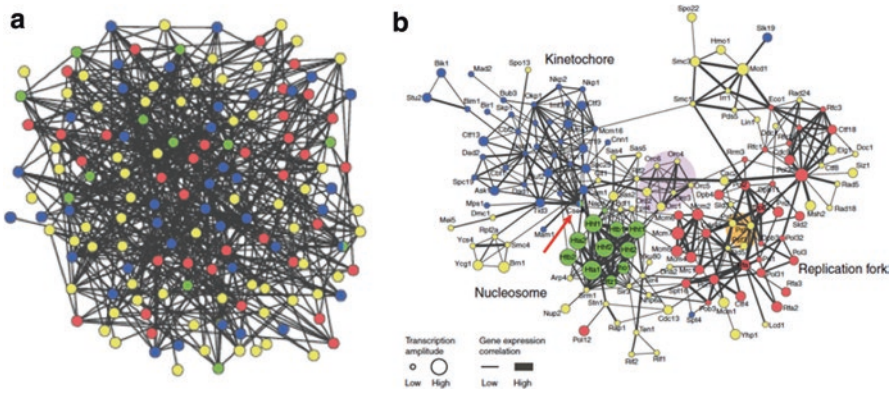

Fig. 4 (**a**) A hairball network diagram based on PPIs among proteins involved in chromosome maintenance and duplication in *Saccharomyces cerevisiae*. (**b**) The network has been transformed into an informative network diagram. Reprinted by permission from Springer Nature: *Nature Biotechnology*, How to visually interpret biological data using networks, Merico et al. 2009

nodes are large and connected with numerous thick edges, indicating that the expression of proteins in the nucleosome is changing together during the cell cycle.

Now that the nodes are laid out in an informative manner, a researcher can zoom in to local regions and make his or her own inferences about parts and operations. A commonly used inference strategy is guilt-by-association—if neighbors of a node without an annotation share a common annotation (in this case, for a cellular component), the researchers infer that the unannotated node should receive the same annotation. The three proteins shown in the region shaded in orange in Fig. 4b, Psf1, Psf2 and Psf3, are colored yellow since GO did not assign them a cellular component annotation below the level of chromosome. The layout procedure, however, situated them among the red nodes that have the replication fork annotation. Employing guilt-by-association, the researchers inferred these proteins should be assigned that annotation as well. Merico et al. report that although these proteins are not so annotated in GO, research already published showed that they belonged to the GINS complex in the nucleosome that is responsible for assembling the DNA replication machinery. Guilt-by-association led the network researchers to make a correct assignment.

The layout algorithm also enables the identification of new mechanisms. The nodes labeled Orc1, Orc2, Orc3, Orc4, Orc5 and Orc6 are located together (in a region shaded in violet) apart from the three regions of nodes annotated to cellular components. The authors infer that they form a distinct mechanism and report that although these nodes lacked specific annotations in GO, "they are known members of the yeast origin recognition complex (ORC), responsible for the loading of the replication machinery onto DNA" (p. 922). In this case again the inference is supported.

Cytoscape thus provides researchers the ability to transform tables into network diagrams, assign visible features to attributes and annotations of entities and their relations, and determine how the nodes and edges will be laid out. Exploration with different approaches (e.g., changing whether an attribute is represented by the shape or color of nodes) is often important to finding informative patterns. This would be very cumbersome if researchers had to construct each network diagram by hand but relatively easy with Cytoscape.

5 Further Analyzing Networks: Cytoscape's App Store

As I have noted, Cytoscape provides a platform for other researchers to construct apps to perform specific analyses for their own purposes but also make the resulting apps available to others. In this way Cytoscape serves multiple groups of users who have different research agendas and require different tools for their execution. Many of the apps are the focus of journal publications that describe the procedures employed in the app and one or more examples of its use (I have identified such publications for several of the apps discussed below). In Spring 2017 there were

more than 180 apps in the Cytoscape App Store that work with Cytoscape 3.X.[5] Some apps support the import and integration of data from specific databases that researchers might wish to represent in networks. For example, KEGGScape, GeneMania, ReactomeFIViz, and STRING, draw results from these different databases into Cytoscape. Bisogenet integrates and imports data from multiple databases such as DIP, BIOGRID, BIND, MINT, and IntAct. AgilentLiteratureSearch allows users to directly query published literature for PPIs and incorporate the results into a Cytoscape network. Apps such as BiNGO and ClueGO facilitate annotation of nodes and edges using Gene Ontology.

Yet other apps provide layout and visualization algorithms that extend beyond what is offered in the core. For example, Cy3D generates three-dimensional views of networks while CyAnimator supports the construction of animations. With respect to layout, GOlorize enables the use of GO annotations to direct the layout of nodes so that the network is interpretable in terms of biological functions while DeDaL facilitates using principal components analysis in developing layouts, aligning one network with another, and morphing between selected layouts so as to find ones that are biologically interpretable.

Yet other apps support particular analyses of networks useful for specific lines of research. I will first discuss two classes of analysis apps: those used to compute a variety of standard network measures and those designed to identify clusters or modules in a given network. I will then offer two illustrations of how particular apps contribute to a better understanding of biological processes.

Apps for Computing Network Measures Graph theorists have developed an extensive set of measures to characterize networks. For purposes of this exposition, I will focus only on networks with undirected edges. Some of the most common measures are *mean shortest path length*, the *clustering coefficient*, and *node degree distribution*. The length of a path between two nodes is the number of edges that are traversed in going from one to the other; the mean shortest path length is the mean for all pairs of nodes of the shortest (or characteristic) path lengths between them. It provides a measure of how quickly effects can travel through the network. The nodes to which any given node is connected are its neighbors and the clustering coefficient characterizes the degree to which the neighbors of a node are connected to one another. Finally, node degree refers to the number of connections a given node has to other nodes. Of particular interest are networks in which node degree is not distributed normally but according to a power law. In such a case, some nodes are highly connected to other nodes, and serve as hubs, whereas most nodes have few connections. NetworkAnalyzer (Assenov et al. 2008) computes these and many other statistics that are used to characterize networks, displaying the results in histograms or scatterplots. Apps such as CytoHubba identify hubs.

[5] Another 132 Apps were written for Cytoscape 2.X but have not been recoded to work with Cytoscape 3.X. This was a serious cost of completely revising the Cytoscape's program interface in 2013, which was done in part to improve the architecture through which apps interact with the core program.

Apps for Identifying Clusters For many research objectives it is valuable to identify nodes that are especially highly interconnected. These clusters, sometimes referred to as *modules*, often reflect groups of components that perform a common activity—that is, work as a mechanism. The apps Molecular Complex Detection (MCODE) (Bader and Hogue 2003) and ClusterMaker2 (Morris et al. 2011) identify clusters. Modules may be organized hierarchically, sometimes with different types of connections at different levels. When Bandyopadhyay et al. (2008) developed a network based on both PPI and genetic interactions they found that PPIs tended to link nodes in modules while genetic interactions generated higher level clusters. Srivas et al. (2011) implemented the procedure Bandyopadhyay et al. employed in the app PanGIA.

5.1 Applying an App for Identifying Active Modules

Most clustering algorithms view networks as static structures, but Ideker et al. (2002) sought to identify nodes that organize into clusters or mechanisms only in specific circumstances such as when particular genes are mutated or yeast are grown on specific media. In an earlier paper, Ideker et al. (2001) has investigated the galactose (GAL) utilization mechanism in yeast. They started with PPI and protein-DNA interaction data to construct a network of 348 genes with 362 interactions. They grew colonies of wild-type and nine mutant strains, each lacking one known GAL gene, on media containing or lacking 2% galactose, measured global mRNA changes and protein concentration changes across the conditions, and plotted these on the network. As Cytoscape had not yet been developed, they used the LEDA toolbox developed at the Max-Planck-Institut für Informatik (Mehlhorn and Näher 1999) to construct the network shown in Fig. 5a. Arrows represent protein-DNA interactions and straight edges PPIs. The nodes are shown in clusters corresponding to genes that exhibited similar changes in expression over all perturbations and the clusters are labeled by their biological functions. Darker shading of nodes indicates increased and lighter shading decreased expression. The size of the nodes reflects the magnitude of change in the case in which gal4 (the node colored in red) is knocked out in the presence of galactose. The network diagram reveals that the expression changes resulting from the perturbation is more correlated in connected proteins than among randomly selected proteins, a result Ideker et al. further confirmed with statistical analysis.

In the 2002 study, Ideker et al. sought to identify modules in which expression changed the most in specific conditions. Having developed Cytoscape, they represented the network in it and developed an analysis strategy that became one of the first Cytoscape apps, jActiveModules. The analysis first computes a z-score for the degree of change in expression of each gene in a particular condition, indicated by the shading of the nodes in Fig. 5b. It then identifies subnetworks of genes under or over expressed and rank-orders them in terms of activity. The top five subnetworks are indicated in Fig. 5b by common coloring of the node border and the attached

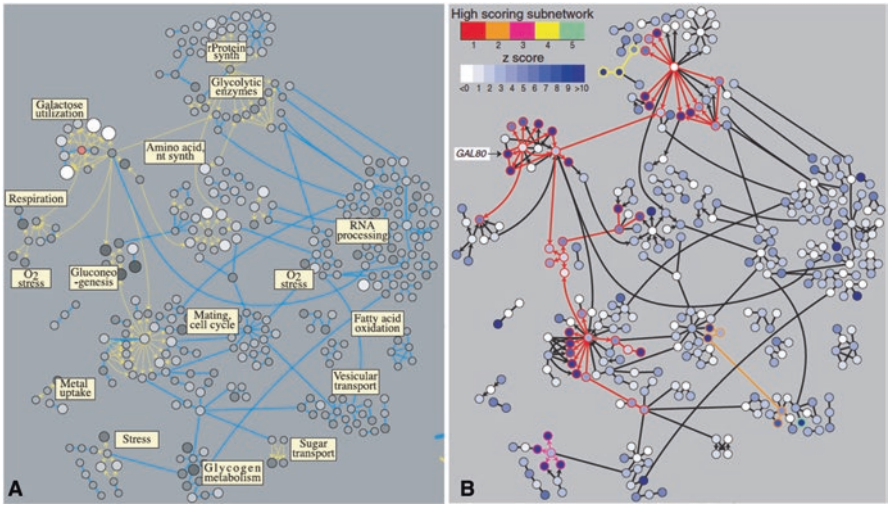

Fig. 5 Comparative network diagrams: (**a**) from Ideker et al. 2001 and (**b**) from Ideker et al. 2002. Both show the same 362 associations between genes whose expression was increased or decreased when grown with or without 2% galactose. In the diagram on the left, darker nodes indicate increased expression when gal4 (shown in red) is knocked out. The edges shown in color other than black in the diagram on the right indicate the subnetworks that were most altered when gal80 was knocked out. A. From Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., & Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science, 292*, 929–934. Reprinted with permission from AAAS. B. reprinted from Ideker, T., Ozier, O., Schwikowski, B., & Siegel, A. F., Discovering regulatory and signalling circuits in molecular interaction networks, *Bioinformatics,* 2002, Volume 18 Suppl 1, S233–240, by permission of Oxford University Press

edges. Ideker et al. interpret the subnetworks active in a particular condition as mechanisms involved in transmitting signals and performing regulatory functions. In the example shown, GAL80 (the only labeled node) is deleted. The adjacent node, GAL4, is a hub with protein-DNA connections to seven other genes. This suggests the hypothesis that GAL80 influences these genes through its effect on GAL4, a hypothesis for which there was already empirical support (Lohr et al. 1995).

5.2 Applying an App for Modeling Diffusion

Whereas jActiveModules was one of the first apps developed for Cytoscape, Diffusion (Carlin et al. 2017) is one of the most recent. Diffusion implements a distinctive strategy for discovering underlying clusters that correspond to mechanisms that has proven effective in fields such as cancer research in which researchers confront extremely heterogeneous data. For example, in The Cancer Genome Atlas study of 500 tumors of various types, individual tumors exhibited from 20 to 300 somatic mutations, with the genes mutated varying substantially across samples of

the same type of tumor. This made it difficult to determine which mutations might play a causal role. To address this problem, Vandin et al. (2011) developed a strategy of mapping mutated genes onto a PPI network and treating them as hot spots from which simulated heat could diffuse. In many cases, heat diffusing from different nodes would converge on the same cluster of nodes. These nodes were hypothesized to represent a mechanism or pathway that, when disrupted through any of the mutations, leads to cancer. The approach was further developed by Hofree et al. (2013), who used propagation in networks to stratify cancer populations in ways that corresponded to patient survival. Heat diffusion algorithms are computationally extremely demanding. Thus, the designers of Diffusion linked the app locally installed on an individual researcher's computer to an internet service that performs the computation. Using Diffusion within Cytoscape, the user can visually select nodes as heat sources, invoke the service, and then visualize the diffusion results.

Carlin et al. employed Diffusion to better understand why one melanoma cell line responds to the drug Vemurafenib (LOX-IMVI) while another is resistant. They use a network generated from the NCI Pathway Interaction Database (an amalgamation of expert-curated cancer pathways) and initiated diffusion from six genes with known relations to the drug: *BRAF*, *PDGFRB*, *NRAS*, *HGF*, *MAP 2 K1*, and *MAPK1*. Diffusion identified a subnetwork of 53 nodes and 448 edges. Cytoscape was then used to filter the top 10% of nodes activated after diffusion. Based on combining the results of multiple queries followed by filtering, Carlin et al. determined that *TSC2* and *BLNK* are mutated in the resistant but not the sensitive cell lines and proposed that this might explain the difference.

6 Network Expo: NDEx

In the previous two sections I have characterized how tools like Cytoscape allow for data that has traveled to databases to travel one step further and be used in network analyses. But is that the end of the line? In this section I show how network diagrams themselves can also travel. Traditionally, network diagrams have been distributed as static visual representations and those who wanted to analyze them further had to recreate them for themselves. But networks generated with Cytoscape and similar programs can be stored in structured data formats in which they can then be distributed to other users, who may then incorporate additional data into the network or perform a different type of analysis (e.g., a different clustering procedure) to the existing network. While such sharing can be carried out informally by authors,[6] the Network Data Exchange (NDEx) is providing a platform for doing this on a large scale.

[6] A collaboration between Elsevier and Cytoscape created the Interactive Network Viewer which allowed authors to make networks available in online publications in a viewer with some capacities for readers to further explore the network or download it to Cytoscape. This project is no longer active.

NDEx was introduced in 2014 as "an online commons" (Pillich et al. 2017) or expo that functions much like World Expos. In this case, the exhibits are the networks that provide original interpretations of data. By uploading their networks, researchers can showcase them and others can download them for use in their own work. The developers further characterize NDEx as "a step toward an ecosystem in which networks bearing data, hypotheses, and findings flow easily between scientists" (Pratt et al. 2015). The project employs its own group of developers in Ideker's lab at UC San Diego and is supported by the National Cancer Institute, the National Resource for Network Biology, the California Stem Cell Agency, Pfizer, Janssen, and Roche.[7]

At its core, NDEx functions much like Google Docs or Dropbox. Networks are added to NDEx either from other online sources such as Pathway Commons, which draws data from a wide range of databases including BIND, DIP, and BioGRID that were discussed in Sect. 2, or by individual users via either direct file import or from Cytoscape. Individual users store their own networks and have control over who can access them—they can keep them private, share them with designated others, or make them public. Sharing with a group of researchers allows a group to collaborate in further developing a network. If made public, other users might use the network as the basis for their own work and upload new versions for others to access. Each network that is added to NDEx is assigned a Universally Unique Identifier (UUID) so that it can be easily referenced. If someone modifies a public network and saves it, it is assigned a new UUID. NDEx is distinct from other online network repositories such as KEGG and Pathway Commons in that users manage their own networks rather than the networks being managed by the organization that maintains the resource. To facilitate visualizing and indexing networks as well as interactions with Cytoscape, NDEx employs the Cytoscape Cyberinfrastructure network exchange format, CX, to store information. CX, however, maintains the semantics of the format employed by the creator of the network.[8]

For networks to be useful to others, it is important that depositors provide sufficient information about how they were created and the data that was used (databases are updated regularly and attempts to reconstruct networks will not necessarily yield the same results unless the same iteration of the database is used). Accordingly, NDEx maintains a provenance history that contains this information. The history also includes information about other networks that were used in constructing a particular network.

For NDEx to provide a useful expo, other users must be able to find networks that are relevant to them. Thus, when networks are uploaded, NDEx indexes text strings for network descriptions, the user and group that manages the network, the

[7] Legally, the Cytoscape Consortium, a 5.0.1cs corporation, owns Cytoscape and NDEx, along with NeXO and Cytoscape.js. It contracts with the various pharmaceutical companies and subcontracts with UC San Diego.

[8] WikiPathways provides a useful comparison case with NDEx. WikiPathways is based on the Wiki model in which everyone collaborates on a common public document. It is also limited to small networks and allows for content that is not represented in a network.

genes or proteins represented by the nodes, the relations represented by the edges, and references cited. Users can initiate searches from NDEx homepage by entering names of cell processes or names of genes or proteins. This will bring up a table listing a number of networks. Figure 6 shows the results of a search for three circadian genes, *per2*, cry2, and *bmal1*. This returned 165 networks in which at least one of these genes is included. The table shows the name of the network, the number of nodes and edges, whether the network is public or private, the owner, and the date it was last modified. When one hovers a mouse over the name of a network, a popup window appears with a description of the network if one has been provided. If there is an icon in the Ref. column, it links to a publication in which the network appeared. One can proceed to download the network by selecting the icon with a white downward arrow.

Clinking on a network name brings it up in a window (if there are too many edges, a sample of 500 edges will be displayed). Users can choose instead to see a listing of the edges in a table view. The screen also shows either network info (e.g., when it was created, its UUID address) or the provenance history. A search box enables users to query particular nodes and select a number of edges out from those nodes. The network selected in Fig. 6 has 195 nodes and 4534 edges. Entering CRY2 and distance 1 returns the more restricted network shown in Fig. 7. Selecting the nodes PER2, CRY2, and the two edges connecting them, brings up information about the nodes, including links to UniProt, GenBank, and publications providing evidence for the edges.

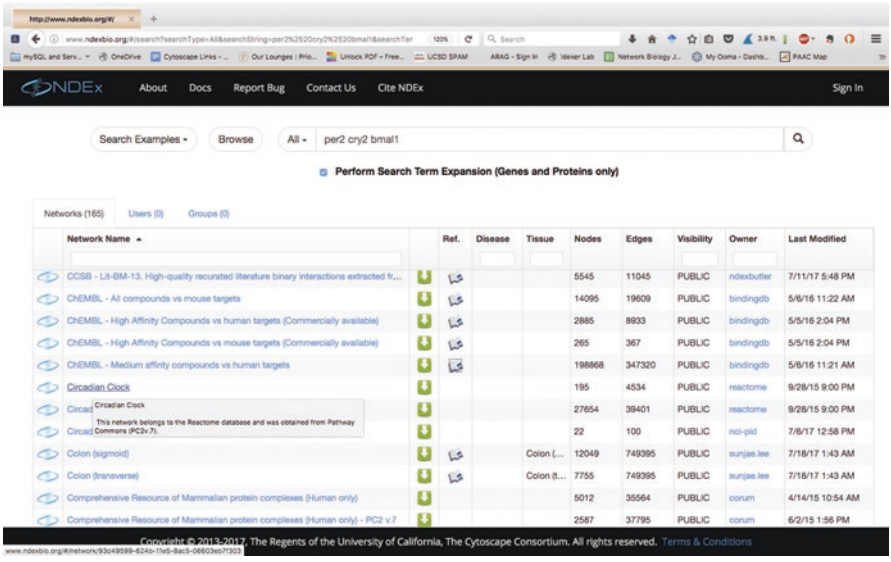

Fig. 6 Screen shot of NDEx after search for networks that include *per2*, *cry2*, or *bmal1*, three prominent mammalian circadian genes

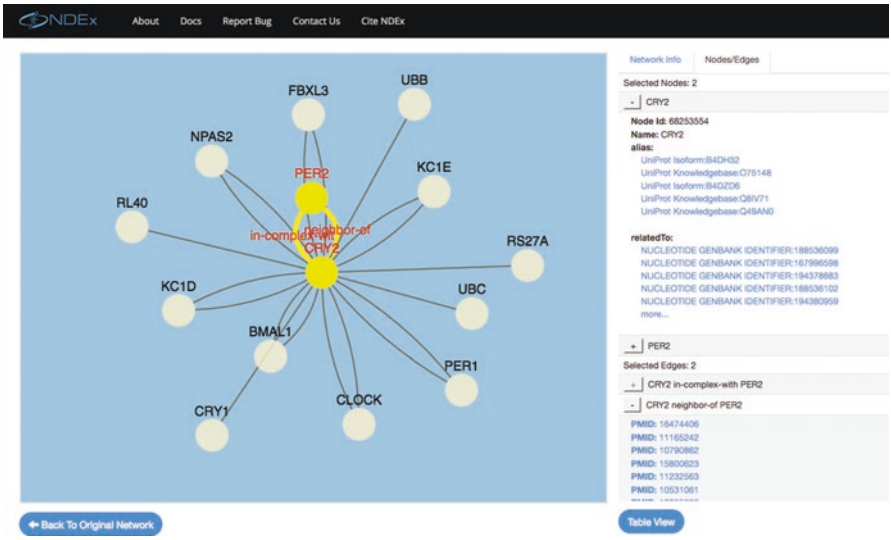

Fig. 7 Screen shot of the network selected in Fig. 6 after a query requesting nodes directly connected to CRY2

NDEx has been designed to integrate smoothly with Cytoscape. From within Cytoscape, one can use the app CyNDEx to query networks in NDEx and import selected ones. CyNDEx also allows users to export networks developed or modified in Cytoscape to NDEx. Once a network has been imported from NDEx to Cytoscape, a researcher can use it to continue the inquiry for which it was originally designed by carrying out additional analyses or accept the analysis offered and incorporate further data into the network.

The developers of NDEx have advanced a bold vision of how NDEx can provide "new models of scientific publication." It provides an expo "in which live data structures replace static diagrams and supplemental files." Drawing upon these live data structures, other biologists can create new networks that serve their own ends and create new expositions in NDEx. For NDEx to realize these goals, network biologists must be willing to share their networks. There is evidence that they will as use of NDEx is showing steady growth. From July 2015 until March 2016 the number of unique visitors per month increased from 151 to over 1200. As of July 2017 there were 3190 public networks, 810 registered users and 37 groups, although not all of these have uploaded networks to NDEx. The developers are pursuing a number of strategies to encourage greater use such as making NDEx a platform on which authors may make networks in their papers available to reviewers. To the extent that NDEx is successful as an expo of networks, network diagrams will be both products of inquiry and inputs for future inquiries.

7 Conclusions

In systems biology and many other fields, relational data travel from individual researchers to publicly accessible databases, from which they are accessed and employed by subsequent researchers. I have focused on the resources that systems biologists have created to enable further data journeys. These resources are allowing researchers both to represent and extract interpretations from the data and to share the products of their research so that other researchers can build upon them. These tools enable data and the analyses constructed from them to continue to travel far beyond the initial database to which they were uploaded.

My focus has been on the increasingly popular use of network representations of relational data. Networks are not just an attractive format in which to represent data. As I have developed in earlier publications, they are employed in novel ways to make discoveries about biological mechanisms. In recent decades, philosophers of biology have characterized the research strategies by which biologists in a variety of fields search for mechanisms to explain phenomena of interest (Bechtel and Richardson 1993/2010; Craver and Darden 2013). Most of these strategies start with hypothesized mechanisms and decompose them to find their constituents. Network biology pursues a different strategy, starting with data about how biological entities are related to each other (e.g., which proteins interact), identifying mechanisms as local clusters within the network and appealing to them to explain biological phenomena (Bechtel 2017, 2019).

Key to network biology is the construction of network representations and the application of tools to analyze these representations. Since its introduction in 2002, Cytoscape has emerged as a freely available and widely used platform for creating and analyzing network representations. The core of Cytoscape allows researchers to import databases of relational data and generate network representations employing a variety of different layouts that enable specific inferences from the data and different ways to annotate the representation to incorporate yet additional information. A user can, for example, quickly switch between different layouts until he or she finds one that provides insight into the data. Of central importance are algorithms used to find clusters of nodes that are then interpreted as potential mechanisms.

The construction of a revealing network representation is often just the starting point for further analysis. The core of Cytoscape provides a range of tools intended for use on a wide variety of network studies (extending, for example, to the social sciences). But Cytoscape also provides a platform for other researchers, often with interests limited to specific domains, to develop their own analytic tools in the form of apps. By providing an App store, the developers of Cytoscape have encouraged researchers to make these available to yet other researchers.

Cytoscape and its apps are powerful tools for researchers to reuse data that has been deposited into the growing number of databases developed by biologists. A particularly valuable feature is allowing researchers to readily integrate data from a

variety of different databases into a single network that can then be analyzed in different ways. Until recently, however, these network representations and the analyses performed on them represented the end of data journeys—they might be published, but anyone who wanted to carry on the inquiry would have to procure the network in a useable format from the researchers or reconstruct it for themselves. By providing an easily searchable expo of networks that other users can access, add data to, and further analyze (using Cytoscape or another platform), NDEx enables data to travel yet further. Since users can both download networks and upload their revised network, data can be recirculated potentially indefinitely.

Resources such as databases, Cytoscape and its apps, and NDEx, constitute important infrastructures that are increasingly relied upon by contemporary biologists. These tools supplement traditional experimental tools, allowing results to travel widely and to be analyzed by multiple researchers using different techniques for network analysis. They thereby contribute in novel ways to the development of scientific knowledge.

Acknowledgements I thank the editors, Sabina Leonelli and Niccolò Tempini for helpful comments on this manuscript. Further, I thank Benjamin Sheredos, Rebecca Hardesty, Jason Winning, and other members of the Philosophy of Science in Practice Study Group at UC San Diego as well as participants in the workshop on Varieties of Data Journeys at Exeter University in November 2017 for their valuable comments and suggestions on earlier drafts of this paper. In addition, I thank Barry Demchak, former Project Manager for Cytoscape, and Dexter Pratt, Director of Software Development in the Ideker Lab, for their comments and suggestions and Trey Ideker for inviting me to sit in on his lab meetings.

References

Alfarano, C., C.E. Andrade, K. Anthony, N. Bahroos, M. Bajec, et al. 2005. The Biomolecular Interaction Network Database and Related Tools 2005 Update. *Nucleic Acids Research* 33: D418–D424.

Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, et al. 2000. Gene Ontology: Tool for the Unification of Biology. *Nature Genetics* 25: 25–29.

Assenov, Y., F. Ramirez, S.E. Schelhorn, T. Lengauer, and M. Albrecht. 2008. Computing Topological Parameters of Biological Networks. *Bioinformatics* 24: 282–284.

Bader, G.D., and C.W. Hogue. 2003. An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks. *BMC Bioinformatics* 4: 2.

Bandyopadhyay, S., R.M. Kelley, N.J. Krogan, and T. Ideker. 2008. Functional Maps of Protein Complexes from Quantitative Genetic Interaction Data. *PLoS Computational Biology* 4: e1000065.

Bechtel, W. 2017. Using the Hierarchy of Biological Ontologies to Identify Mechanisms in Flat Networks. *Biology and Philosophy* 32: 627–649.

———. 2019. Analyzing Network Models to Make Discoveries About Biological Mechanisms. *British Journal for the Philosophy of Science* 70: 459–484.

Bechtel, W., and R.C. Richardson. 1993/2010. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Cambridge, MA: MIT Press. 1993 edition published by Princeton University Press.

Boumans, Marcel, and Sabina Leonelli. this volume. From Dirty Data to Tidy Facts: Clustering Practices in Plant Phenomics and Business Cycle Analysis. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Breitkreutz, B.J., C. Stark, and M. Tyers. 2003a. The GRID: The General Repository for Interaction Datasets. *Genome Biology* 4: R23.

———. 2003b. Osprey: A Network Visualization System. *Genome Biology* 4: R22.

Cambrosio, Alberto, Jonah Campbell, Etienne Vignola-Gagné, Peter Keating, Bertrand R. Jordan, and Pascale Bourret. this volume. 'Overcoming the Bottleneck': Knowledge Architectures for Genomic Data Interpretation in Oncology. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Carlin, D.E., B. Demchak, D. Pratt, E. Sage, and T. Ideker. 2017. Network Propagation in the Cytoscape Cyberinfrastructure. *PLoS Computational Biology* 13: e1005598.

Craver, C.F., and L. Darden. 2013. *In Search of Mechanisms: Discoveries Across the Life Sciences*. Chicago: University of Chicago Press.

Dayhoff, M.O., and R.V. Eck. 1965-1972. *Atlas of Protein Sequence and Structure*. Silver Spring: National Biomedical Research Foundation.

Eades, P. 1984. A heuristic for graph drawing. *Congressus Numerantium* 42: 149–160.

Fields, S., and O. Song. 1989. A Novel Genetic System to Detect Protein-Protein Interactions. *Nature* 340: 245–246.

Griesemer, James. this volume. A Data Journey Through Dataset-Centric Population Genomics. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Hermjakob, H., L. Montecchi-Palazzi, G.D. Bader, J. Wojcik, L. Salwinski, et al. 2004a. The HUPO PSI's Molecular Interaction Format--A Community Standard for the Representation of Protein Interaction Data. *Nature Biotechnology* 22: 177–183.

Hermjakob, H., L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, et al. 2004b. IntAct: An Open Source Molecular Interaction Database. *Nucleic Acids Research* 32: D452–D455.

Hofree, M., J.P. Shen, H. Carter, A. Gross, and T. Ideker. 2013. Network-Based Stratification of Tumor Mutations. *Nature Methods* 10: 1108–1115.

Ideker, T., V. Thorsson, J.A. Ranish, R. Christmas, J. Buhler, et al. 2001. Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. *Science* 292: 929–934.

Ideker, T., O. Ozier, B. Schwikowski, and A.F. Siegel. 2002. Discovering Regulatory and Signalling Circuits in Molecular Interaction Networks. *Bioinformatics* 18 (Suppl 1): S233–S240.

Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, et al. 2001. A Comprehensive Two-Hybrid Analysis to Explore the Yeast Protein Interactome. *Proceedings of the National Academy of Sciences of the United States of America* 98: 4569–4574.

Lawton, J.R., F.A. Martinez, and C. Burks. 1989. Overview of the LiMB Database. *Nucleic Acids Research* 17: 5885–5899.

Leonelli, S. 2010. Documenting the Emergence of Bio-Ontologies: Or, Why Researching Bioinformatics Requires HPSSB. *History and Philosophy of the Life Sciences* 32: 105–125.

———. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago: University of Chicago Press.

———. this volume. Learning from Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Lohr, D., P. Venkov, and J. Zlatanova. 1995. Transcriptional Regulation in the Yeast GAL Gene Family: A Complex Genetic Network. *The FASEB Journal* 9: 777–787.

Mehlhorn, K., and S. Näher. 1999. *Leda: A Platform for Combinatorial and Geometric Computing*. New York: Cambridge University Press.

Merico, D., D. Gfeller, and G.D. Bader. 2009. How to Visually Interpret Biological Data Using Networks. *Nature Biotechnology* 27: 921–924.

Morgan, Mary S. this volume. The Datum in Context: Measuring Frameworks, Data Series and the Journeys of Individual Datums. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Morris, J.H., L. Apeltsin, A.M. Newman, J. Baumbach, T. Wittkop, et al. 2011. clusterMaker: A Multi-Algorithm Clustering Plugin for Cytoscape. *BMC Bioinformatics* 12: 436.

Müller-Wille, Staffan. this volume. Data, Meta Data and Pattern Data: How Franz Boas Mobilized Anthropometric Data, 1890 and Beyond. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Orchard, S. 2012. Protein Interaction Data Curation: The International Molecular Exchange (IMEx) Consortium. *Nature Methods* 9: 345–350.

Orchard, S., L. Salwinski, S. Kerrien, L. Montecchi-Palazzi, M. Oesterheld, et al. 2007. The Minimum Information Required for Reporting a Molecular Interaction Experiment (MIMIx). *Nature Biotechnology* 25: 894–898.

Peri, S., J.D. Navarro, R. Amanchy, T.Z. Kristiansen, C.K. Jonnalagadda, et al. 2003. Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans. *Genome Research* 13: 2363–2371.

Pillich, R.T., J. Chen, V. Rynkov, D. Welker, and D. Pratt. 2017. NDEx: A Community Resource for Sharing and Publishing of Biological Networks. *Methods in Molecular Biology* 1558: 271–301.

Porter, Theodore M. this volume. Most Often, What Is Transmitted Is Transformed. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Pratt, D., J. Chen, D. Welker, R. Rivas, R. Pillich, et al. 2015. NDEx, the Network Data Exchange. *Cell Systems* 1: 302–305.

Rigaut, G., A. Shevchenko, B. Rutz, M. Wilm, M. Mann, et al. 1999. A generic Protein Purification Method for Protein Complex Characterization and Proteome Exploration. *Nature Biotechnology* 17: 1030–1032.

Rogers, S., and A. Cambrosio. 2007. Making a New Technology Work: The Standardization and Regulation of Microarrays. *The Yale Journal of Biology and Medicine* 80: 165–178.

Schwikowski, B., P. Uetz, and S. Fields. 2000. A network of Protein-Protein Interactions in Yeast. *Nature Biotechnology* 18: 1257–1261.

Shannon, P., A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, et al. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* 13: 2498–2504.

Srivas, R., G. Hannum, J. Ruscheinski, K. Ono, P.L. Wang, et al. 2011. Assembling Global Maps of Cellular Function Through Integrative Analysis of Physical and Genetic Networks. *Nature Protocols* 6: 1308–1323.

Tempini, Niccolò. this volume. The Reuse of Digital Computer Data: Transformation, Recombination and Generation of *Data Mixes* in Big Data Science. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

The UniProt Consortium. 2017. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Research* 45: D158–D169.

Uetz, P., L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, et al. 2000. A Comprehensive Analysis of Protein-Protein Interactions in Saccharomyces cerevisiae. *Nature* 403: 623–627.

Vandin, F., E. Upfal, and B.J. Raphael. 2011. Algorithms for Detecting Significantly Mutated Pathways in Cancer. *Journal of Computational Biology* 18: 507–522.

Xenarios, I., D.W. Rice, L. Salwinski, M.K. Baron, E.M. Marcotte, et al. 2000. DIP: The Database of Interacting Proteins. *Nucleic Acids Research* 28: 289–291.

Zanzoni, A., L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, et al. 2002. MINT: A Molecular Interaction Database. *FEBS Letters* 513: 135–140.

William Bechtel is Distinguished Professor of Philosophy and a Member of the Center for Circadian Biology and the Interdisciplinary Program in Cognitive Science at the University of California, San Diego. His research focuses on philosophical issues in cell and molecular biology, systems biology and circadian biology. In his book *Discovering Complexity* (1993/2010, with Robert Richardson), he argued that explanations in many fields of biology take the form of identifying a mechanism responsible for a selected phenomenon. He developed this perspective in detail for cell biology in *Discovering Cell Mechanisms* (2006) and for cognitive science and neurobiology in *Mental Mechanisms* (2008). In subsequent work with Adele Abrahamsen, he has explored the use of computational modelling to understand the functioning of interactive biological mechanisms such as the circadian clock that exhibit complex dynamic behaviour. He has also investigated strategies for network representation in systems biology and their use in discovering mechanisms within larger interactive systems. Most recently, he has expanded his focus on how productive biological mechanisms (e.g. muscles and metabolic pathways) are controlled by heterarchically organized control mechanisms within cells and multicellular organisms and how such control enables production mechanisms to support the autonomy of these organisms.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.