Using the Hierarchy of Biological Ontologies to Identify Mechanisms in Flat Networks

William Bechtel
Department of Philosophy
University of California, San Diego

Abstract

Systems biology has provided new resources for discovering and reasoning about mechanisms. In addition to generating databases of large bodies of data, systems biologists have introduced platforms such as Cytoscape to represent protein-protein interactions, gene interactions, and other data in networks. Networks are inherently flat structures. One can identify clusters of highly connected nodes, but network representations do not represent these clusters as at a higher level than their constituents. Mechanisms, however, are hierarchically organized: they can be decomposed into their parts and their activities can be decomposed into component operations. A potent bridge between flat networks and hierarchical mechanisms is provided by biological ontologies, both those curated by hand such as Gene Ontology (GO) and those extracted directly from databases such as Network Extracted Ontology (NeXO). I examine several examples in which by applying ontologies to networks, systems biologists generate new hypotheses about mechanisms and characterize these novel strategies for developing mechanistic explanations.

1. Introduction

In the 19th and 20th centuries biologists in many fields of biology sought to explain phenomena by identifying responsible mechanisms, decomposing them into their parts and operations, and then showing that when appropriately organized and situated, these parts and operations could produce the phenomena for which explanations were sought (Bechtel & Abrahamsen, 2005; Machamer, Darden, & Craver, 2000). This approach starts with the phenomenon to be explained and uses a variety of experimental manipulations to differentiate parts and operations. Typically initial research identifies only a small number of parts and operations and proposes a relatively simple account of how they are organized to generate the phenomenon. Over time more and more parts and operations are identified and the complexity of the mechanistic accounts grows (Bechtel & Richardson, 1993/2010; Craver & Darden, 2013). Systems biology, however, has provided a different strategy for advancing mechanistic accounts. Automated techniques have enabled the collection of vast amounts of data about constituents (e.g., genes and proteins) of living systems. These data have been collected in a large number of databases, many organized around specific species,[1] and then analyzed to put forward hypotheses about mechanisms and their components.

[1] Over the past 15 years numerous large databases based on a variety of interactions between genes or between molecules in cells have been made publically available. A

Some systems biologists (Huang, 2011) and philosophers of biology (Braillard, 2010; Gross, 2011; Huneman, 2010) have viewed the tools for analyzing large-scale data as providing an alternative to traditional mechanistic accounts. Many systems biologists, however, continue to appeal to mechanisms. Instead of rejecting mechanistic explanations, they seek to draw upon this vast amount of data to develop enriched accounts of biological mechanisms. This process often involves developing network representations of a body of data and performing computational operations on these networks to identify clusters of proteins or genes that interact. The researchers view these clusters as candidate mechanisms. In many cases proposed mechanisms can be identified with mechanisms that have been discovered through more traditional biological research. Even when this is the case, though, the clusters often involve many additional genes or proteins that are potential new components of these mechanisms. In other cases the clusters don't correspond to existing mechanisms and represent possible new mechanisms.

To illustrate the basic strategy of using clusters in networks to reason about mechanisms, I begin in Section 2 by focusing on one recent study that produced a map of the cell (Figure 1 below). This study, like many others, employed Gene Ontology (GO) to identify the cell structures or processes in which the genes or proteins are involved. [The use of the term *ontology* in biology is only loosely related to its use in philosophy to refer to an account of the types of entities thought to exist and how they relate to one another. The use in biology draws from artificial intelligence (Gruber, 1995) and refers to a formal representation or taxonomy of the entities and their properties and interrelations invoked in a domain of discourse.] GO is the product of a large-scale effort to organize published biological knowledge both within and across species and in many cases in which there has been research on a gene or its product, this information is encoded in GO. Thus, it provides a valuable tool to make sense of networks.

In terms of drawing upon networks to understand mechanisms, however, GO offers much more. One reason some view network analyses as providing an alternative to mechanistic explanation is that networks are flat structures. One can identify interconnected components in networks, but nodes are not organized hierarchically into higher-level systems as proposed in mechanistic accounts. GO, however, is organized hierarchically—small cell components are identified as parts of larger components and specific biological processes are linked to higher-level biological biological processes. The hierarchies in GO correspond to the hierarchies found in mechanisms in which parts are contained within larger structures and operations of parts contribute to the operations of these larger structures. Thus, GO does more than provide annotations—it anchors network interpretations in a mechanistic framework. I will examine GO as well as NeXO, a related

---

regularly updated compilation of molecular biology databases is maintained at https://www.oxfordjournals.org/our_journals/nar/database/c/. It currently includes 685 databases. Starting with supplementary issues in April 1991 and May 1992 and a regular issue in July, 1993, the journal *Nucleic Acids Research* has regularly reviewed databases. Beginning in 1996, the journal identified its first issue of each year as the database issue.

ontology developed directly from the databases from which networks are built, in Section 3.

In Section 4 and 5 I turn to how GO and NeXO have recently been put to use to analyze networks to provide novel understanding of biological mechanisms. Section 4 examines a procedure referred to as *active interaction mapping* that has been employed to advance understanding of the phenomenon of autophagy. The research generated an ontology specifically for autophagy on the basis of which the researchers advance new hypotheses about parts and operations of the responsible mechanisms. Finally, in section 5, I describe how the incorporation of the information in ontologies enhances the ability of researchers to determine how complicated mechanisms will behave, thereby showing how they explain those behaviors of cells.

To date the most extensive research using ontologies to advance mechanistic explanations on the basis of networks has been done on budding or brewer's yeast, *Saccharomyces cerevisiae*. Budding yeast has been adopted as a model organism due to the extensive molecular tools that have been developed to perform experimental manipulations and its relatively small genome (approximately 6000 genes). Accordingly, I will restrict my focus to research on this model organism.

Although many types of data have been collected and used as the basis for network analysis, I will focus on two. Most people think of proteins as individually catalyzing reactions in cells, but they often do so as part of complexes. These complexes result from the binding of proteins to one another and are generally construed by the biologists as structural elements of the cell. Techniques such as yeast two-hybrid screening and affinity purification followed by mass spectrometry have enabled the generation of large bodies of data about which yeast proteins are able to bind to each other.

Although genes are also structural components, researchers often characterize them functionally in terms of the phenotypic traits to which they contribute. A standard way of figuring out the functions of a gene is to knock it out, observe the deficits, and infer what is missing. Gene interaction studies extend this procedure. About a thousand genes in yeast are essential—the yeast does live when they are knocked out. Knocking out the others individually does not kill the yeast, although it may slow colony growth. Some pairs of non-essential genes are not lethal when individually knocked out but are when jointly knocked out (these are referred to as a *synthetic lethals.*) A lesser effect (known as *synthetic sickness*) arises when a double knockout results in a reduction in growth of a colony that is either lesser or greater than what would be predicted from the effects of single knockouts. When the effect on growth is less than expected, the interaction is considered positive. (More complex procedures have been developed to identify interactions involving essential genes.) Synthetic lethality and sickness are interpreted as showing interactions between genes—for example, the proteins coded by the genes may be able to substitute for each other or perform related operations.

As data about protein-protein interactions and synthetic lethal gene interactions began to be assembled in large databases at the end of the 20th century, researchers, many with

training in computer science as well as biology, began to develop new techniques to analyze the data. They developed network models in which proteins or genes are the nodes and interactions are shown as edges. A given protein may interact with several other proteins, and edges will connect the nodes for each of these proteins. Each protein will in turn interact with different other proteins, resulting in a large interconnected network. Likewise, a given gene may generate synthetic lethality or sickness with numerous genes, and those genes with yet other genes. In many cases, both sorts of interactions will be shown in the same network, where the nodes represent both the genes and the proteins for which they code. Often the interconnections become very dense, resulting in what is derogatorily referred to as a *hairball*. To make sense of these networks, investigators invoke a number of strategies, many of which serve to cluster together nodes that are especially highly interconnected.

2. Identifying Mechanisms with Clusters in Network Representations

Whether the data involves protein-protein interactions or gene interactions, the clusters that are identified are interpreted mechanistically as either parts of mechanisms or mechanisms themselves. To show how this is done, and how researchers attempt to draw mechanistic insights from such analyses, I focus on one of the largest gene-interaction studies that has been conducted. Costanzo, Baryshnikova, Bellay et al. (2010) targeted 1712 genes (constituting about 30% of the yeast genome) that they took to be representative of the whole genome. They examined survival and colony growth for each possible double mutant, finding approximately 170,000 synthetic lethal interactions among the approximately 5.4 million gene pairs they examined. They employed Cytoscape, a platform that supports construction of network representations and allows for multiple forms of analysis, to construct the network in Figure 1. This involved several procedures. First, the researchers identified those gene pairs in the overall connectivity matrix that were most highly correlated (Pearson coefficient > 0.2) and inserted edges between them. Second, using an edge-weighted, spring-embedded network layout algorithm they generated a network. The algorithm caused nodes for genes that interact in a similar way to be clustered together. Finally, they annotated the network using Gene Ontology (GO), which provided information about what is known about the biological processes to which particular genes contribute. (GO is discussed further in the next section.) Nodes with the same GO annotation are colored the same. The clustering of colors in Figure 1 illustrates that common gene interaction profiles is predicative of common function. The way the network was constructed not only positioned together nodes for genes that exhibited the same profile, but situated clusters that are engaged in related cell processes near each other. For example, mitosis and chromosome repair, DNA replication and repair, and cell polarity and morphogenesis are positioned near each other.

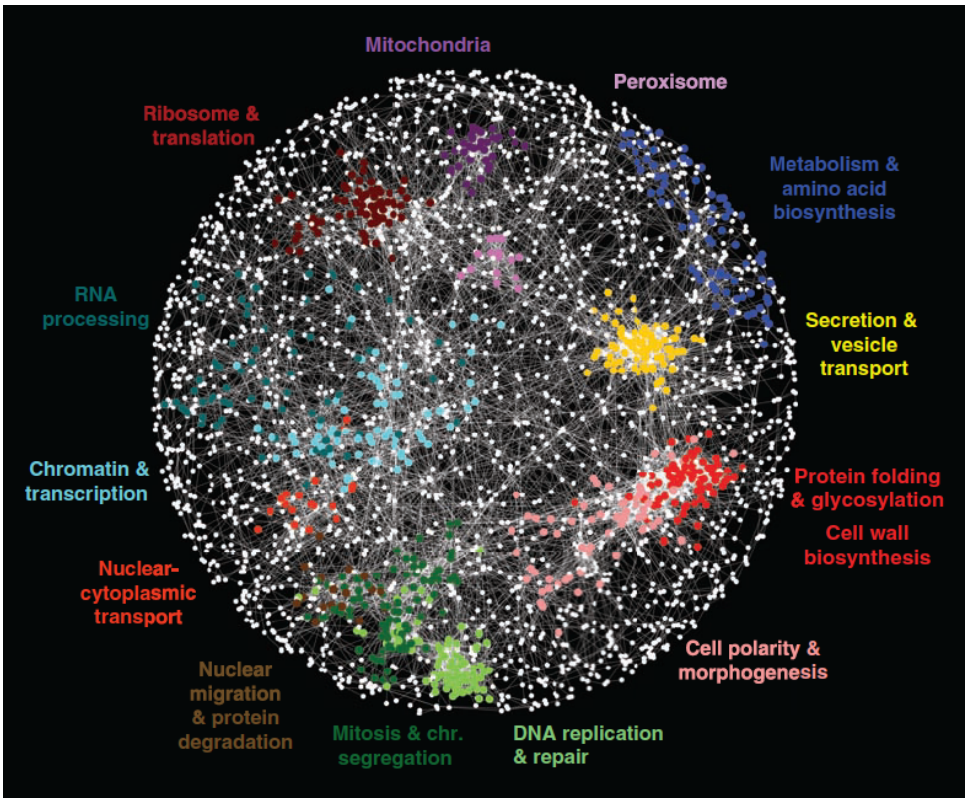

Figure 1. Functional map of a cell generated from interactions among 1712 yeast genes. Genes with similar interaction profiles are located near each other. Color-coding indicates the biological processes associated with genes in GO. From Constanzo et al. (2010), Figure 1, reprinted with permission from AAAS.

Viewing the network with greater resolution reveals connections between genes that interact in performing more specialized functions, interacting in specific biological pathways and or protein complexes. Figure 2A extracts sub-networks of the network shown in Figure 1, with panels B-D further isolating three sub-networks: those for amino acid biosynthesis and uptake, the endoplasmic reticulum (ER) and Golgi apparatus, and tRNA modification. Some edges correspond to positive interactions (green), others to negative ones (red). Diamonds correspond to essential genes, circles to non-essential genes, and yellow to those whose function is not characterized in GO. At this scale one can identify subcomplexes within each complex that correspond to components of the given mechanism.

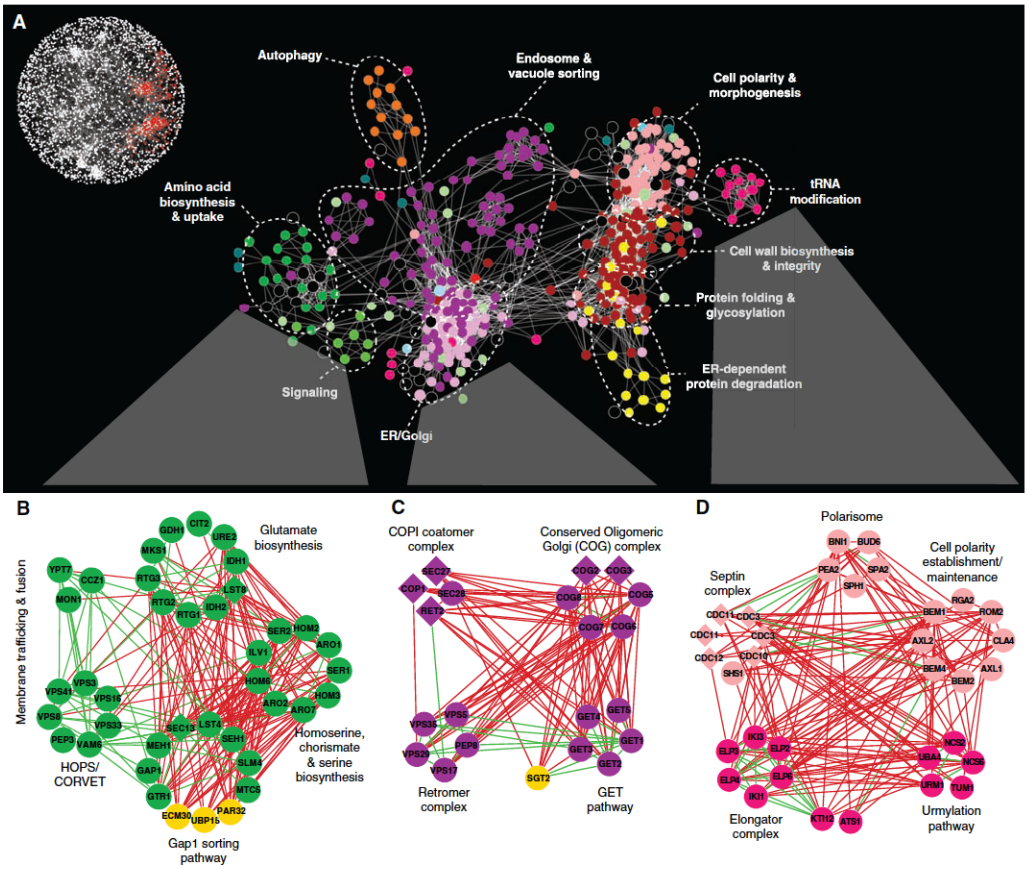

Figure 2. Subnetworks of network shown in Figure 1. Panels B-D isolated specific networks within those shown in Panel A and focus in on the positive or negative interaction between clusters within three of the subnetworks. From Constanzo et al.'s (2010), figure 2, reprinted with permission from AAAS.

One of the advantages of using network analyses to identify mechanisms is that it can reveal components of mechanisms not identified by more traditional strategies that targeted only genes hypothesized to be involved. If a gene not previously identified with a mechanism clusters with others that are part of the mechanism, researchers infer that it is involved in the same mechanism and carries out the same function (a strategy known as *guilt by association*). For example, three genes, *PAR32*, *ECM30*, and *UBP15*, clustered with the genes already assigned to the Gap1-sorting modules. Accordingly, the researchers imputed the same role to them (Figure 2B). They supported this assignment experimentally by showing that when each was deleted, Gap1 sorting and transport deficits appeared. The researchers likewise inferred from the fact that SGT2 exhibits similar connectivity with the GET pathway that it performed a similar function. They supported this inference by showing that when mutated, it, like the GET pathway genes, resulted in the mislocalization of Pex15, a tail-anchored protein.

One virtue of identifying mechanisms within networks is that connections between clusters/mechanisms may identify operations through which one mechanism regulates another. Based on their synthetic lethal interactions, the researchers proposed that the

umylation pathway plays a role in regulating the elongator complex (Figure 2D). In addition to specific regulatory connections between mechanisms, Costanzo et al. also made a number of observations about overall organization of the network, through which mechanisms are coordinated with each other. As in many biological networks, they found that most genes participated in few interactions, but a small number were hubs engaged in large numbers of interactions. Hubs tended also to be genes that exhibited more severe deficits when mutated. Moreover, they also had a greater number of GO attributions. They concluded that hubs are more likely to be pleiotropic and involved in a greater variety of phenotypic traits, suggesting that they "play key roles in the integration and execution of morphogenetic programs."

Costanzo et al.'s study is exemplary of many similar analyses from which researchers are developing new hypotheses about mechanisms in yeast cells. Inference procedures such as guilt by association are discovery heuristics. On their own, they do not suffice to confirm or falsify the hypotheses. Rather, the resulting hypotheses need to be tested by more traditional tools of molecular biology. What is important is that the network approaches are generating many new hypotheses, most of which would not have been advanced otherwise.

2. Constructing Ontologies that Represent Mechanistic Knowledge

To interpret the clusters in their network as cell processes, Costanzo et al., like many other researchers, appealed to Gene Ontology (GO). But in many of these studies, GO is simply a tool for annotating nodes based on what has been reported in the published literature about the biological process to which they contribute or the cell structure in which they are active. But GO offers much more—it represents a form of theory insofar as it not only provides stable definitions of terms but also identifies relations among them (Leonelli, 2010, 2012). In particular, by organizing the terms used to describe cell structures and processes hierarchically, it provides a framework for conceptualizing how components are organized into mechanisms as higher-level entities. In this section I will introduce GO as well as an alternative ontology, NeXO, and show how they represent knowledge about the hierarchical organization of biological mechanisms.

GO was developed, starting in 1998, with the goal "to produce a dynamic, controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is accumulating and changing" (Ashburner, Ball, Blake et al., 2000). The perceived need for a common vocabulary arose as a by-product of the sequencing of the genes of several model organisms. Sequencing of budding yeast, the round worm (*Caenorhabditis elegans*), and the fruit fly (*Drosophila melanogaster*), had just been completed, and sequencing of the flowering plant (*Arabidopsis thaliana*) and fission yeast (*Schizosaccharomyces pombe*) were underway. Finding a large number of likely orthologs in the genetic sequences of budding yeast and worms, Chervitz, Hester, Ball et al. (1999) found they could infer the biological roles of about 12% of the proteins in worms from putative orthologs in yeast. Rubin, Yandell, Wortman et al. (2000) extended this finding to fruit flies. This suggested that a reasonable hypothesis when orthologous genes are found in two species is that they perform the same biological function. At the time, however, there

was not a common vocabulary in which functions were described, making inference between species difficult. As a result, the investigators responsible for three online model organism databases, FlyBase, Mouse Genome Informatics (MGI), and the Saccharomyces Genome Database (SGD), joined forces to create the Gene Ontology Consortium.[2]

The consortium set out not just to produce a standardized vocabulary with clear definitions for each term, which could facilitate communication across model organisms, but also to define relations between terms—to construct an ontology. The challenge its developers faced was how to systematically relate the vocabulary terms they adopted. One part of the challenge was the recognition that knowledge is often incomplete. So a goal was to "organize, describe, query and visualize biological knowledge at vastly different stages of completeness." Another part of the challenge was to incorporate what seemed very different types of information about genes/proteins: information about the molecular function of proteins (i.e., the reactions they catalyzed), the biological processes in which they figured, and the cellular component in which they were active.[3] This led to the construction of three different ontologies for Molecular Function, Biological Process, and Cellular Component. Within each they organized terms hierarchically from low-level instances to high-level instances. For example, "translation" and "cAMP biosynthesis" were low-level specifications of biological process, while "cell growth and maintenance" and "signal transduction" were high-level specifications. A tree structure in which each node has one or more decedents would provide a simple form of a hierarchy, but the designers recognized that individual genes might be active in multiple tissues and their products may contribute to multiple molecular processes and biological processes. Accordingly, each ontology was organized as a directed acyclic graph in which the edges represented "is_a" or "part_of" relations ("has_part" and "regulates" were subsequently added).

Figure 3 shows portions of each ontology from a very early version of GO. In each panel GO terms are shown in black and genes associated with the term are shown in a color representing the species in which they occurred. Panel A shows a section of the Biological Process ontology for DNA metabolism. DNA repair and DNA recombination are both high-level children of DNA metabolism. DNA ligation is a lower-level process in three different processes: DNA repair, DNA recombination, and DNA-dependent DNA replication. Accordingly, it has three different parents. Many of the same genes are shown in more than one ontology and in some cases appear in multiple locations in the same ontology. For example, *Mcm2-7* are shown under the term pre-replicative complex formation and maintenance in the Biological Process ontology, under chromatin binding in the Molecular Function ontology and as active in two locations (the cytoplasm and the pre-replicative complex) in the Cell Component ontology.

---

[2] MGI was itself the product of integrating two mouse databases. In 2000 the Arabidopsis Information Resource (TAIR) and the Caenorhabditis elegans group joined GO.

[3] "These particular classifications were chosen because they represent information sets that are common to all living forms and are basic to our annotation of information about genes and gene products" (Ashburner, Ball, Blake et al., 2001).

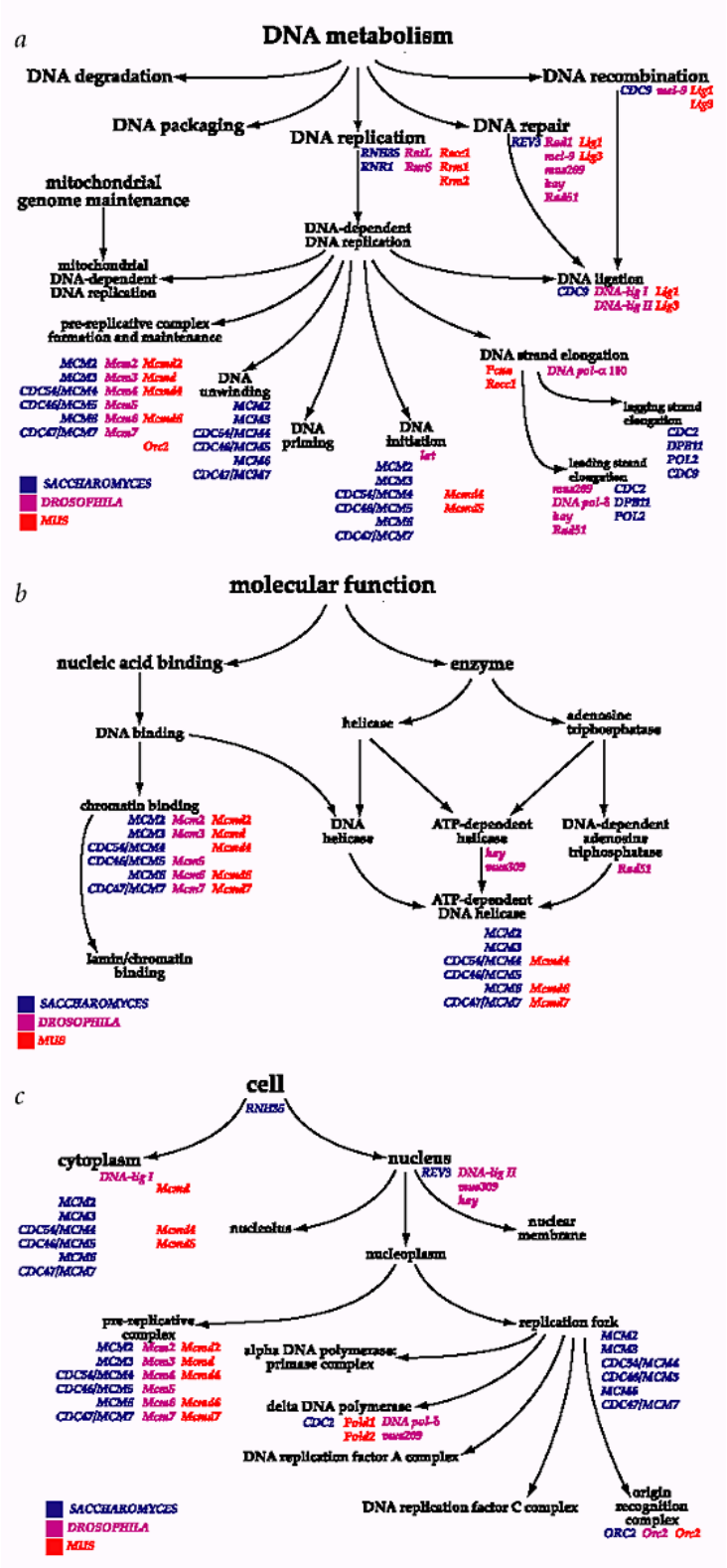

Figure 3. Parts of each of the three ontologies in GO from an early version of GO. Reprinted by permission from Macmillan Publishers Ltd: *Nature Genetics* from Ashburner et al. (2000).

The knowledge represented in GO is mechanistic knowledge. The Cell Component ontology identifies different components of the cell that are associated with specific cell activities and so the locus of one or several mechanisms. The part_of relation between components at different levels of the hierarchy corresponds to the fact that parts, which may themselves be mechanisms, are structural constituents of particular mechanisms. The Biological Process hierarchy characterizes the activities performed by individual mechanisms within the larger system while the Molecular Function identifies the specific operations in which specific components are engaged. Identifying a gene as active in a part of the cell that can be situated among other parts, as engaging in a molecular reaction that is categorized among the various reactions in a cell, and as contributing to a specific biological process that falls under more general biological process, specifies its role in a mechanism.

Moreover, GO is not static but is constantly evolving to include new mechanistic knowledge. From the outset GO's developers recognized that knowledge about cellular mechanisms would grow and change, and this would involve not just locating new genes under various terms but also altering the structure itself. In fact the number of terms in GO has expanded dramatically. One vehicle for expansion has been merging with other ontologies, such as the Subcellular Anatomy Ontology that was one of the ontologies the Neuroscience Information Framework Standard developed for neuroscience. This required examination of each term to determine whether it matched one in GO (perhaps using a different name) or represented additions to GO (Gene Ontology Consortium, 2015).

Leonelli, Diehl, Christie et al. (2011) describe a number of major changes that were made to the structure of GO in its first decade. Some of these resulted from detecting anomalies. For example, when "serotonin secretion" was first incorporated in the Biological Process ontology it was an is_a child of "hormone secretion" and "neurotransmitter secretion." When "serotonin secretion during acute inflammatory response" was added as an is_a child of "serotonin secretion" a problem was recognized: serotonin secretion was not strictly a sub-type of "neurotransmitter secretion" since it operates on other cells than neurons. Accordingly, GO curators made several revisions: they added a new term "neurotransmitter secretion," which they situated at the same level as "serotonin secretion." They made "serotonin secretion during acute inflammatory response" a child only of the latter. Other changes resulted from expanding the scope of GO, for example, to include host-parasite relations or to include prokaryotes. The latter, for example, required considerable adjustment since a cell component such as the "tricarboxylic acid cycle enzyme complex" that resides in the mitochondrion in eukaryotes is situated in the cytoplasm in prokaryotes. This resulted in, among other changes, adding a specific term for "mitochondrial tricarboxylic acid cycle enzyme complex" as a child of both "tricarboxylic acid cycle enzyme complex" and "mitochondrial matrix," and locating the original term as a child of "cytoplasmic part." These various changes were often the product of extended discussion among the curators and reflect an attempt to synthesize the developing knowledge into a coherent mechanistic framework that provides the categories needed to describe both the structural and functional hierarchies involved in mechanistic accounts of various cell phenomena.

While GO is a valuable resource reflecting mechanistic understanding of cells as it is developed in published scientific studies, it faces limitations. First, a constant challenge for the Gene Ontology Consortium is manually curating all new findings. Given the increasing rate of publication, this has quickly become extremely costly and the Consortium is regularly investigating ways to automate the expansion of GO (e.g., using TermGenie to add new terms). There is a second problem, however, which has motivated the development of an alternative approach to creating ontologies—as a result of being built from established scientific research, GO is focused around the cell components and processes that have been investigated: "there has been a strong bias in coverage within GO toward processes that are well-studied, and a corresponding lack of coverage of processes that have been more recently identified" (Dutkowski, Kramer, Surma et al., 2013). To overcome this limitation, Dutkowski et al. set out to create an ontology directly from large-scale data sets that include information not yet codified into mechanistic understanding of cells.

Most network studies based on cluster analysis generally result in what Dutkowski et al. characterize as a flat network, not hierarchically organized as GO is. The challenge Dutkowski et al. took up was to extract a hierarchical organization directly from networks derived from three datasets that represented protein-protein interactions, gene interactions, co-expressed genes datasets, and from YeastNet, an integrated probabilistic functional gene network that combines evidence from multiple experimental sources to provide a weighted functional relationship between genes (Lee, Li, & Marcotte, 2007). They annotated each network using GO, which revealed a substantial agreement between them and the GO hierarchy. Dutkowski et al. then integrated the four networks and applied a probabilistic model for detecting hierarchically organized communities or complexes (Park & Bader, 2011). This generated a binary tree in which genes were progressively grouped into larger clusters (Step 1 in Figure 4). Forcing the data into a binary tree prevented nodes from having more than two children or more than one parent. Since terms in ontologies can have multiple parents and children, Dutkowski et al. revised the graph to allow terms to have multiple parents and children when doing so would increase the fit to the original network data (fit was judged by a probability score that takes into account edges in the original network that cross between subtrees) (Step 2). The resulting graph constituted the NeXO ontology. The researchers then aligned it with the GO Cellular Component ontology by matching terms with the same (or very similar) assignment of genes and located at similar points in the hierarchy (Step 3).

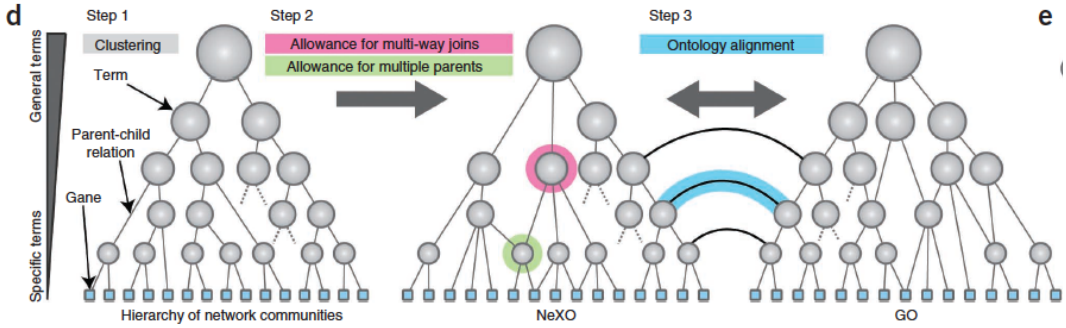

Figure 4. Steps in the construction of NeXO. Reprinted by permission from Macmillan Publishers Ltd: *Nature Biotechnology*, from Dutkowski et al. (2013), Figure 1.

Aligning NeXO with GO served multiple ends: transferring names and definitions in GO to NeXO, identifying terms in NeXO that are not in GO, and registering conflicts between NeXO and GO. Figure 5A shows the NeXO ontology as a tree with three main branches for the intracellular compartment, the membrane, and the mitochondrion. The size of nodes indicates the number of genes assigned to the term while the color indicates the degree of correspondence to a term in GO. The names of the high-level term assignments in GO are indicated. Overall NeXO shows a high correlation with GO. Altogether, a third of NeXO terms map to terms in one or more of the three GO ontologies. The percentage is greatest (~60%) for terms in the GO Cellular Component ontology, and about 25% for terms in the Molecular Function and Biological Process ontologies.

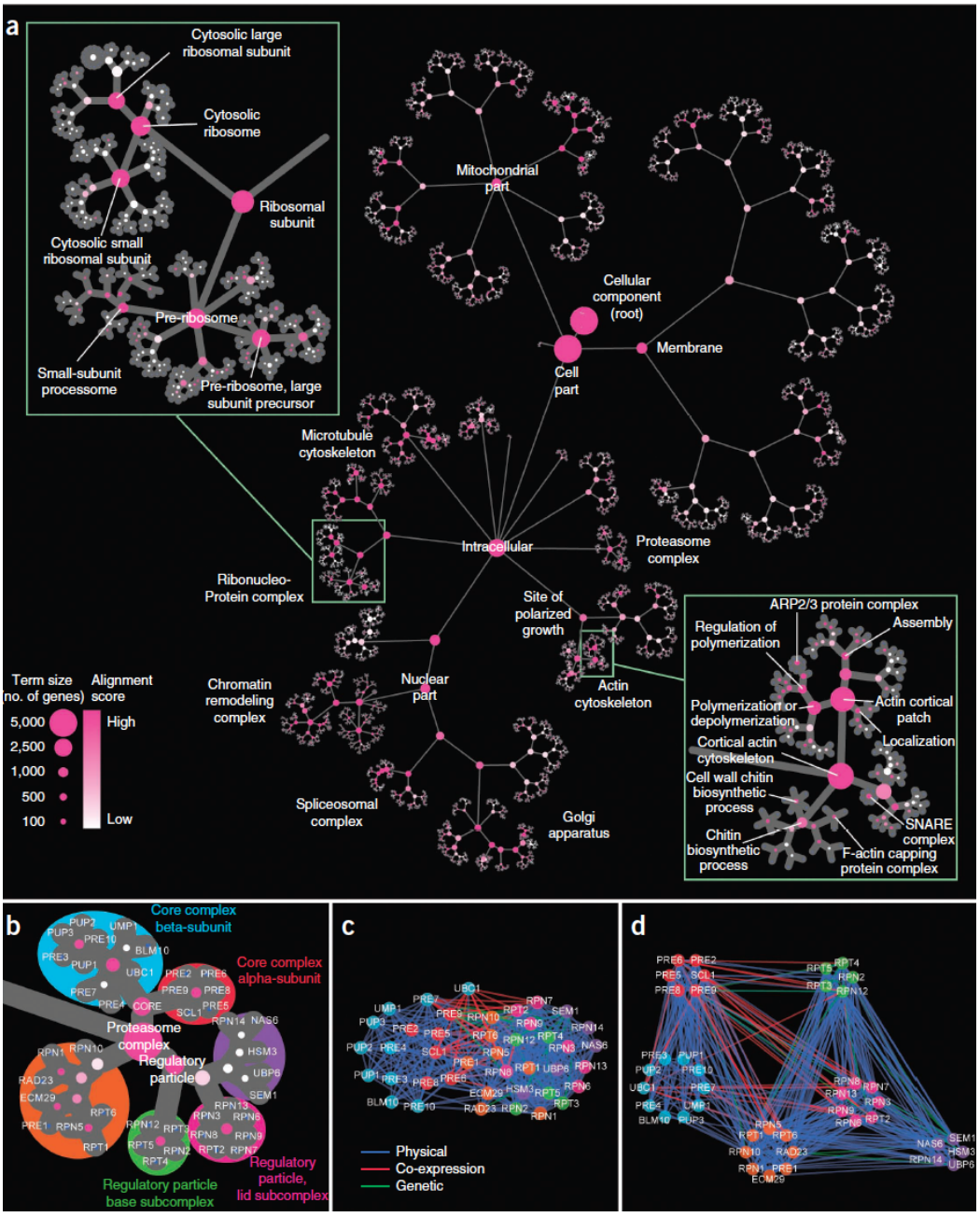

Figure 5. A. The NeXO ontology after alignment with the Cellular Component of GO. The color of nodes indicates the degree of alignment. B-D. The region of the ontology for the proteasome complex. C and D show different network representations of the raw data while B, the product of aligning with GO, reveals the detailed mechanistic organization of the proteasome complex. Reprinted by permission from Macmillan Publishers Ltd: *Nature Biotechnology*, from Dutkowski et al. (2013), Figure 2.

As well as viewing the global structure of NeXO, one can examine it at a more fine-grained level. Panel 5B focuses on the portion of the ontology for the proteasome complex (a complex of proteins that degrades unneeded or damaged proteins) and reveals a hierarchy of subcomponents. Graphing the raw interaction data for the proteasome complex with a

force directed layout (Figure 5C) and a layout distinguishing interactions within and between subcomponents (Figure 5D), revealed a division of the proteasome complex into two components. Yet, the alignment with GO in panel 5B most clearly brings out the hierarchical structure of the proteasome complex. The proteasome is revealed as a mechanism with multiple parts performing a variety of different operations.

Since the point of developing a data-driven ontology was to go beyond what is provided in GO, Dutkowski et al. focused primarily on terms not in GO. One group of terms consisted of those for which strong support could be found in the literature but had not yet been incorporated into GO. Dutkowski et al. submitted these to GO and many were accepted. One example is that the NeXO term NeXO:6164 groups together *BLS1*, *SNN1*, *CNL1*. Recent studies had identified these as members of a complex referred to as BROC in yeast, and the GO curators revised GO by introducing the term BROC-1 complex and identifying these genes as members. NeXO also identified several relations between terms not captured in GO, such as that the Piccolo NUA4 complex, which it presents as a part of the NUA4 histone acetyltransferase complex. The curators also incorporated these relations into GO. In these cases NeXO extends the mechanistic knowledge in GO by finding published results that the curators of GO had not yet identified or incorporated. But the strategy employed is rather different. Rather than taking specific pieces of information about a part or operation developed in a traditional study of a cell mechanism, NeXO starts with relationships identified in networks built from large-scale databases. Through application of the algorithms that generate NeXO, the researchers identify components that fit into mechanisms and then return to the published data to identify what had already been discovered about them.

The power of the strategy is revealed when NeXO supports new hypotheses about mechanisms and their parts and operations. For example, the researchers identified 73 genes that annotated to the NeXO term that mapped to the GO term “Golgi apparatus” but were not so annotated in GO. Zeroing in on where some of these genes are located in NeXO, Figure 6 shows a subnetwork under the new NeXO term NeXO:9763. Because of its similar connectivity, NeXO:9763 is positioned next to the retromer, known to regulate recycling transmembrane receptors, and the HOPS and Corvet complexes, which capture endosomal vesicles. The researchers inferred that NeXO:9763 is also involved in endosomal and Golgi regulation. Within NeXO:9763 there are two pairs of highly correlated genes which are each assigned a term in NeXO: *NNF2* and *YEL043W* fall under the term NeXO:8060 while *MTC1* and *SFT2* are assigned to NeXO:9270. These terms represent potential new components of a mechanism, but unlike terms in GO, which generally can be referred to with common English names that fit their role in a mechanism, many of the new terms identified by NeXO lack common English names. This reflects the fact that NeXO picks out previously unstudied units in the cell and is advancing hypotheses about how they fit into the mechanistic accounts of cell structure and function.

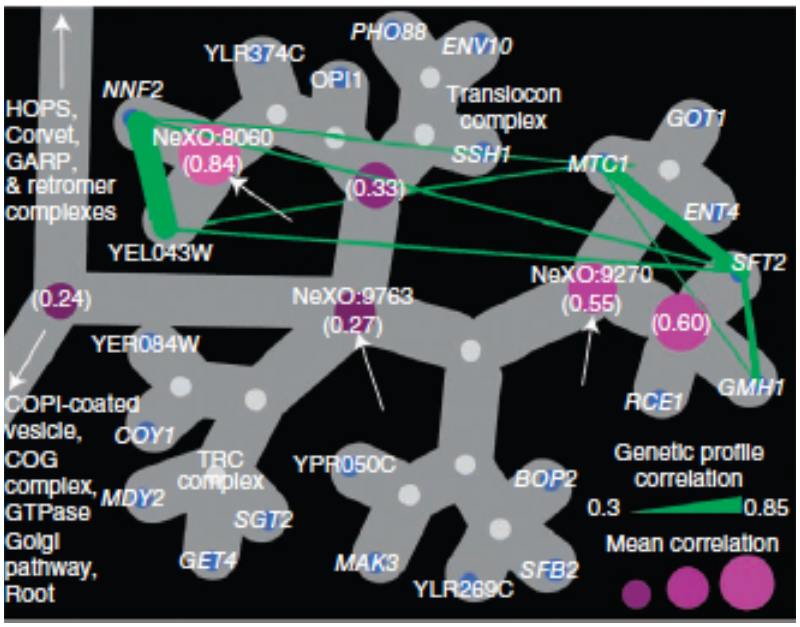

Figure 6. The location of the NeXO term 9763 near the HOPS and Corvet complexes supported the hypothesis that the components organized under this term are involved in endosomal and Golgi regulation. Reprinted by permission from Macmillan Publishers Ltd: *Nature Biotechnology*, from Dutkowski et al. (2013), Figure 4b.

The two ontologies discussed in this section, GO and NeXO, provide a hierarchical perspective that is typically lacking in network analyses. They thus provide an important bridge between network analysis and mechanistic hypotheses. The relations from which the hierarchies are constructed correspond to component relations between parts or operations in mechanisms. In the case of GO the alignment with mechanisms is not surprising since GO is constructed by curating existing mechanistic research. NeXO, an ontology constructed directly from the data collected in various databases, incorporates terms and genes not included in GO. When new genes are included under an existing term they suggest possible new parts of known mechanisms. When new terms are identified under higher-level terms, they suggest new mechanisms that interact with known ones. In these ways, organizing the data from databases into ontologies supports new mechanistic hypotheses.

3. Applying Ontologies to Networks to Develop Mechanistic Hypotheses

Having discussed in general how ontologies create hierarchical representations that correspond to existing mechanistic understanding or advance new mechanistic hypotheses, I turn now to a specific network study that employed ontologies to develop new mechanistic understanding. Kramer, Farre, Mitra et al. (2017) introduce what they call Active Interaction Mapping (IA-Map) and employed it to advance mechanistic understanding of the phenomenon of autophagy—a process in which cells respond to starvation and other stresses by degrading macromolecules into amino and fatty acids

(procuring needed ATP in the process). The amino and fatty acids are then utilized to synthesize new proteins.

To construct their network model, Kramer et al. combined data from 76 networks built using diverse types of data (protein-protein interactions, gene interactions, gene similarity, etc.). They selected 492 genes (20 identified as having a core function in autophagy, 102 annotated to autophagy in GO, and 370 that manifest similar network connectivity to the core genes). They used a clustering algorithm, CliXO (Clique Extracted Ontology), that Kramer, Dutkowski, Yu et al. (2014) had developed to create a directed acyclic graph in which nodes can have multiple parents and children (avoiding some of the steps originally required to construct NeXO). They termed the resulting graph of 218 terms and 310 relations between terms an initial Autophagy Ontology (Atg 1.0) and aligned it with GO. Figure 7A-C shows the ontology with terms in rectangles—red for terms that did not align with GO, and varying shades of blue for terms that did align (darker colors signify stronger alignment). The size of the rectangles corresponds to the number of genes receiving the annotation. Panels B and C present two parts of the ontology in which genes are shown in ovals, with light shading indicating that the gene was not previously identified with the annotation, medium shading indicating that the gene was so annotated in GO, and dark shading indicating core genes. Panel B shows the ontology beneath the term AtgO:15, which does not align with any GO term, and C the part beneath AtgO:18 that aligns to the GO annotation "macrophagy." Panel D shows the corresponding GO ontology. Many of the genes not annotated to autophagy in GO exhibited significant enrichment for other cell processes including the cell cycle, cellular response to stress, and general vesicle transport, suggesting links between these processes and autophagy. In this and other respects, AtgO 1.0 already has the potential to advance new mechanistic understanding of autophagy, but before exploring this, the researchers developed a revised ontology, AtgO 2.0.

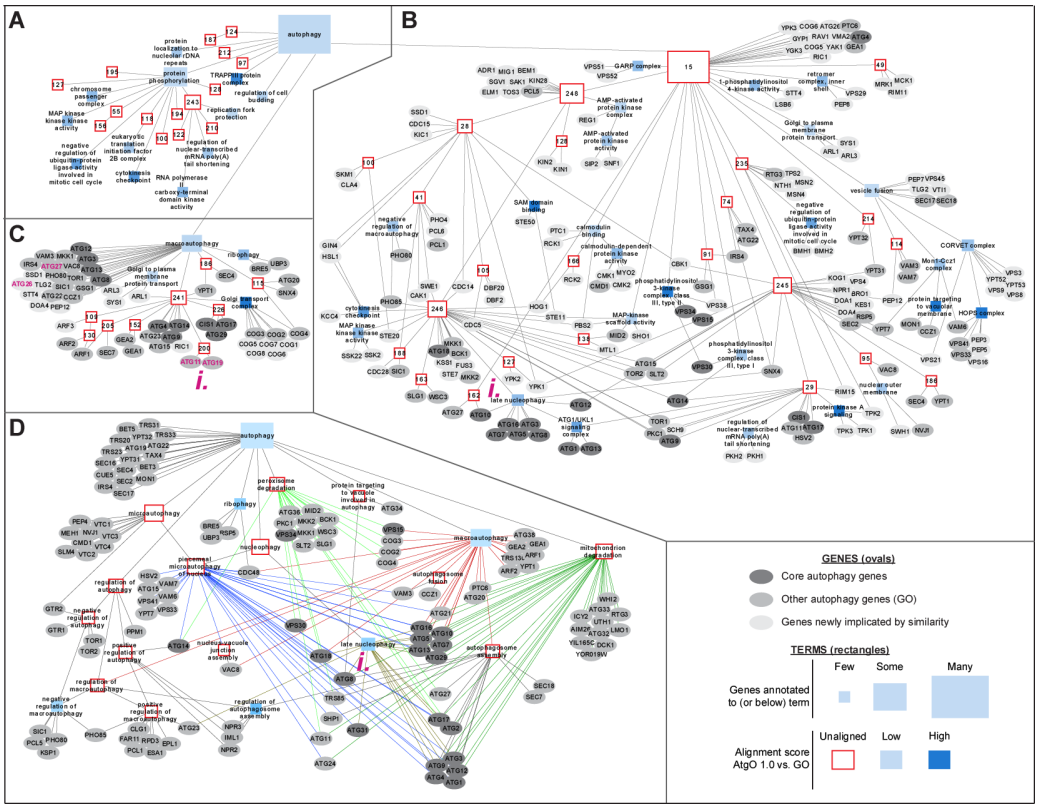

Figure 7. A-C. AtgO 1.0 ontology for autophagy. D. GO ontology for autophagy. Reprinted from Kramer et al. (2017), Figure 2, with permission from Elsevier.

To develop a revised ontology, Kramer et al. identified gene interaction studies as making the largest contribution to AtgO 1.0. These gene interaction studies were all conducted by growing yeast colonies under normal conditions. Other gene-interaction studies, however, compared the networks generated for colonies grown in rich media with those generated for colonies grown under various stress conditions (Luscombe, Babu, Yu et al., 2004; Ideker & Krogan, 2012; Guénolé, Srivas, Vreeken et al., 2013). Kramer and colleagues pursued this strategy, conducting a new gene-interaction study pairing 52 autophagy genes with 3,007 other genes (approximately two-thirds of the non-essential genes in yeast) in two conditions known to induce autophagy (exposure to rapamycin, which pharmacologically induces autophagy, and nitrogen starvation, which metabolically induces autophagy) and an untreated control condition. As in the Costanzo et al. study above, interaction is demonstrated when the affect on growth differs, positively or negatively, from what would be expected with no interaction. The researchers constructed networks for each condition, from which they generated differential networks by subtracting the strengths of specific interactions found in one condition from those found in another. They also constructed a network that integrated all three conditions. From the integrated network they constructed an ontology (using CliXO) that correlated better with GO than AtgO 1.0. The correlation was even better when they constructed the ontology after combining the original network on which AtgO 1.0 had been constructed and the new integrated network. They designated the resulting ontology AtgO 2.0.

The total number of terms increased only slightly between AtgO 1.0 and AtgO 2.0. However, the number of genes assigned to the terms increased significantly (Figure 8). As in the case

of NeXO, the researchers identified terms not in GO but for which there was evidence in the published literature. These were submitted to the GO curators and were accepted. They also identified numerous genes associated with AtgO terms that mapped to GO terms that had not been so annotated in GO. These were also accepted into GO. In this way they used AtgO to develop new mechanistic hypotheses for which evidence already existed in the published literature but had not been incorporated into the GO ontology.

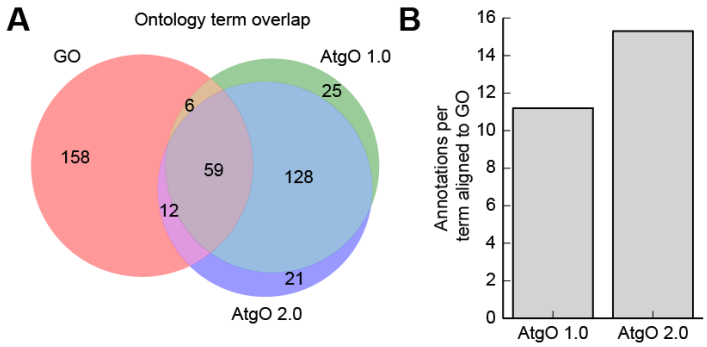

Figure 8. A. Euler diagram showing overlap of terms in GO, AtgO 1.0, and AtgO 2.0. B. Average number of genes annotated to terms aligned with GO in AtgO 1.0 and 2.0 From Kramer (2016), Chapter 4, Figure 6.

Much more interesting is a case in which AtgO 2.0 facilitated advancing an important new hypothesis about autophagy mechanisms. This involved Atg26, which together with Atg27 was added to AtgO:185. In AtgO 1.0, AtgO:185 only included Atg11 and Atg19 (Figure 9). Atg11, Atg19 and Atg27 had previously been established to figure in the cytoplasm-to-vacuole targeting pathway that figures in the transport of aggregates of the aminopeptidase precursor prApe1 to the vacuole where it is processed into mature Ape1. Guilt by association would suggest a similar role for Atg26, but previous studies in budding yeast in which Atg26 was knocked out had failed to reveal any effects on generation of Ape1 or in any autophagy pathway. However, in a different species of yeast, *Pichia pastoris*, the ortholog of Atg26 had previously been shown to be required for the degradation of large peroxisomes in pexophagy. To try to detect a role for Atg26, the researchers compared wild-type and Atg26 mutants in a condition in which prApe1was overexpressed. This condition resulted in larger prApe1 aggregates and now Atg26 mutants exhibited decreased processing. They also demonstrated larger prApd1 aggregates in the mutants with microscopy. Kramer et al. conclude "these data support a new role for Atg26 in the processing of large prApe1 aggregates and validate term AtgO:185, which we hereby name 'prApe1 aggregate processing.'"

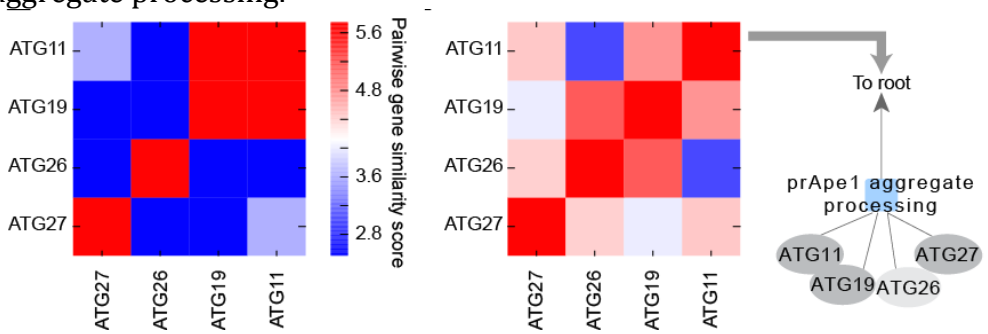

Figure 9. Similarity between gene pairs in AtgO 1.0 on the left and AtgO 2.0 on the right. Only Atg11 and Atg19 cluster in 1.0, and in 2.0 they are joined by Atg26 and Atg27. Reprinted from Kramer et al. (2017), Figure 7, with permission from Elsevier.

In the context of the phenomenon of autophagy, Kramer et al. have shown the potential of ontologies built up from gene and protein interaction data to identify new components of mechanisms and to advance new understanding of the operation of these mechanisms. They suggest, moreover, that this is a strategy that can be applied more generally to other cell phenomena.

4. Putting Ontologies to Work to Map from Genotypes to Phenotypes

A major interest in molecular biology is to determine the phenotypic effects of genetic variations—to be able to map from genotype to phenotype. Of particular interest is to understand why different mutations have the same phenotypic effect. Whereas a focus on individual genes may fail to explain this, the recognition that different mutations all affect the same mechanism may provide the desired explanation. The reason focusing on individual genes often fails is that the effect of perturbing an individual gene often depends on the other genes that are expressed in the cell. Networks such as the one discussed in section 2 provide a first step towards being able to understand the effects of different mutations, but as noted above, networks tend to be flat and so do not recognize the hierarchical of components of the cell into mechanisms. Yu, Kramer, Dutkowski et al. (2016, p. 77) characterize the limitation:

> In reality, however, genotype is transmitted to phenotype not only through gene-gene interactions but through a rich hierarchy of biological subsystems at multiple scales: genotypic variations in nucleotides (1 nm scale) give rise to functional changes in proteins (1–10 nm), which in turn affect protein complexes (10–100 nm), cellular processes (100 nm), organelles (1 **m**m), and, ultimately, phenotypic behaviors of cells (1–10 **m**m), tissues (100 **m**m to 100 mm), and complex organisms (>1 m).

The hierarchy Yu et al. are describing is a hierarchy of mechanisms—perturbing a part of a mechanism affects the mechanism's behavior, and that of yet higher-level mechanisms of which the first is a part. Since gene ontologies provide relevant information about these hierarchies, Yu et al. drew upon them to characterize what they term an *ontotype*. The ontotype is intended to reflect how mutations to particular genes are mediated by the hierarchical mechanistic organization within the cell, "representing variation at intermediate scales between nanoscopic changes in genes and macroscopic changes in phenotype."

The strategy for developing an ontotype and using it to reason about phenotypes is illustrated in Figure 10. On the left in panel A, a toy gene ontology is shown, with the root at the bottom. Mutations are indicated for genes B and D. At each level in the ontology the number of genes associated with a node or its children are indicated by shading. This characterization of mutated genes in terms of an ontology constitutes the ontotype. Drawing on the experimental evidence about growth under each combination of mutations,

Yu et al. employed a learning algorithm to develop rules to predict phenotypes from the features of the ontotype. Panel B shows the ontotype for five combinations of mutations. For example, row 3 is for a mutation to gene *f* alone. In the lowest row of the gene ontology only T4 has *f* as a child and is assigned -1. At the next two levels, T6 and T7 each have one affected child and are also assigned -1. The learning algorithm develops a rule to associate that row of the table to the empirically determined growth value of 0.9. The set of rules constitutes what the authors term a *functionalized ontology*. In mechanistic terms, the rules constituting a functionalized ontology capture how mechanistic organization of cells mediates between alterations of parts and altered behavior of the whole cell.

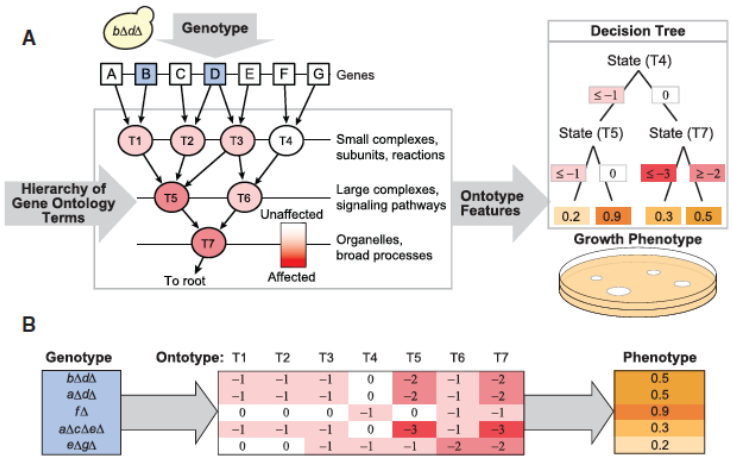

Figure 10. A. Procedure for generating an ontotype to represent effects of gene mutations in an ontology and then applying a learning rule to discover relations to growth phenotype. B. Genotypes mapped onto ontotypes that are then mapped to phenotypes. Reprinted from Yu et al. (2016), Figure 1, with permission from Elsevier.

The authors created functionalized ontologies from both GO (F_{GO}) and NeXO (F_{NeXO}) and each did significantly better than state of the art algorithms at predicting growth for the various genotypes. The authors attributed this success partly to the hierarchical organization of the ontologies, which made apparent to the learning algorithm how different mutations resulted in similar effects to mechanistic components higher in the ontology:

> From the perspective of the ontology, all mutations or variants in a genotype coalesce to the same cellular module, provided one looks at a high enough level. A genotype may include some mutations that map to the same gene, others to the same protein complex, still others to the same broad process or organelle, with all mutations falling within the highest scale represented by the cell itself (pp. 84-5).

To demonstrate further the value of ontotypes, the researchers employed F_{GO} to predict growth for all 12,512,503 pairwise deletions of non-essential genes in budding yeast. 41,605 genetic interactions were predicted, concentrated within and between specific terms. These are shown in Figure 12, using the color of the nodes and edges to indicate whether the connections within or between terms are enriched for positive (blue) or

negative (red) interactions. Some of these actions were between distantly related terms in GO such as negative connections between “intron homing” and “Phosphatidylinositol-3-kinase complex.” These would not have been identified from just examining the ontology and represent potential mechanistic connections to be investigated further.

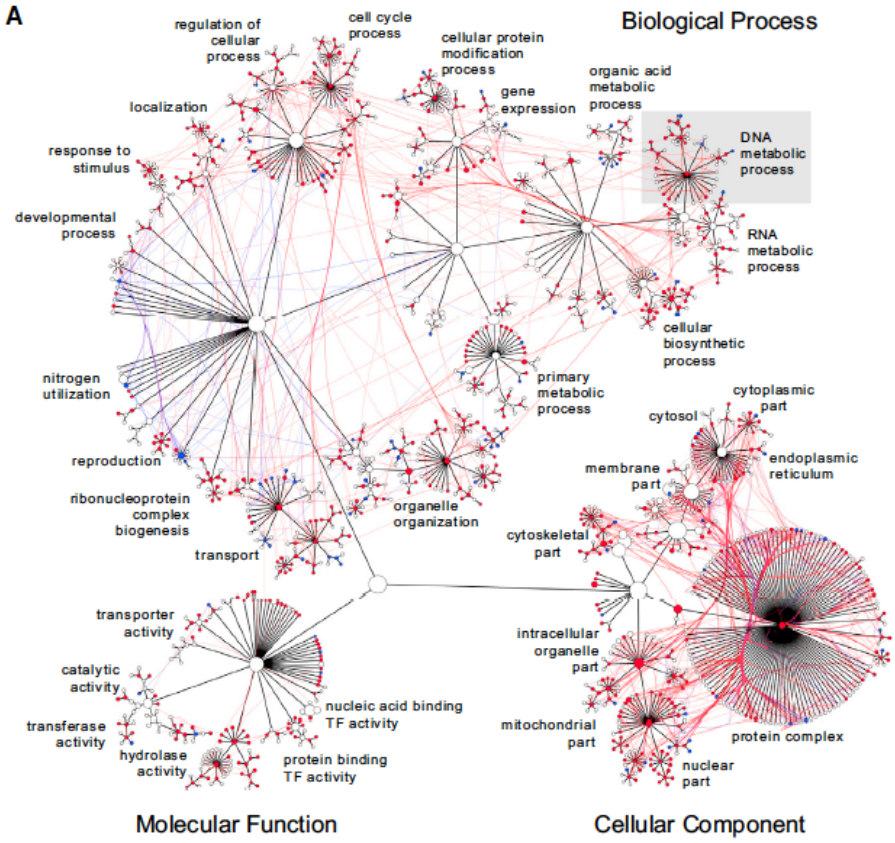

Figure 11. Hierarchy structure of F_{GO}. Red nodes indicate terms with negative enrichment with while blue nodes exhibit positive enrichment. Red edges indicate negative enrichment between terms, blue positive enrichment. Reprinted from Yu et al. (2016), Figure 4, with permission from Elsevier.

The introduction of ontotypes and functionalized ontologies provides a novel perspective on the underlying biological mechanisms and how alterations in mechanisms result in phenomena. Studying the effects of mutations has long being an important strategy for investigating the contributions of individual genes within a mechanism, but individual mutations only provide limited insight. Examining the effects of double mutants reveals associations between genes, allowing network analysis to advance insights into the larger mechanism. Building an ontology provides a hierarchical perspective of mechanisms that allows for interpreting the effects of mutations in light of the mechanisms that are altered. There has been little discussion in the mechanism literature of how one can use mechanistic knowledge to predict how specific alterations in parts of mechanisms will affect the phenomena that are generated, but the construction of ontotypes provides an example of how mechanistic knowledge can be applied to this problem. At the same time,

this research shows how ontotype construction can result in identifying new interactions that enrich the understanding of mechanisms themselves.

5. Conclusions

In this century systems biologists have invoked network representations of vast bodies of data to advance new understanding of biological mechanisms. Costanzo et al.'s analysis based on gene-interaction networks, discussed in section 2, illustrates how cluster analysis performed on networks can guide new discoveries of mechanisms and their parts. But network analyses alone do not capture the hierarchical organization of living cells. GO, based on human curation, captures much of what has been discovered about cellular mechanisms through more traditional mechanistic research. Drawing on GO provides researchers a tool to represent and use the hierarchical organization of biological mechanisms in their network analyses. By building up an ontology directly from databases (albeit relying importantly on alignment with GO) NeXO reveals new hierarchical organization that corresponds to mechanisms. Application of both GO and NeXO provide ways to apply mechanistic organization to understanding biological networks, resulting in new hypotheses about mechanisms and their parts and operations. In section 4 I showed this in the case of research on the mechanisms involved in autophagy. The ontologies constructed for autophagy supported advancing hypotheses of new parts of known mechanisms and of new mechanisms that interact with existing mechanisms. Section 5 then showed a strategy being employed to apply this hierarchical information to explain mechanistically why different mutations may produce the same effects.

The literature on mechanistic explanation offers detailed accounts of the approach of decomposing mechanisms into components and applying reasoning strategies to build up mechanistic explanations of various biological phenomena (Bechtel & Richardson, 1993/2010; Craver & Darden, 2013). These strategies still play an important role in biology and have not been supplanted by the development of new tools for large-scale data collection, representation, and analysis in systems biology. But systems biologists are providing new strategies for advancing hypotheses about mechanisms. The new tools for representing networks, developing and applying ontologies, and creating onotypes, are enabling systems biologists to advance new hypotheses about mechanisms and their components. Once these hypotheses are advanced, more traditional research is employed to evaluate them. But just as philosophical accounts of mechanistic explanations in cell and molecular biology have emphasized heuristic strategies for developing mechanistic explanations, the research in systems biology provides a basis for significantly expanding our understanding of how mechanisms are discovered.

Acknowledgement

I thank Michael Kramer for sharing his dissertation (Kramer, 2016), including what became Kramer et al. (2017), with me. I also thank Trey Ideker, who invited me to attend his lab meetings at which some of the work discussed here and much related research was presented and discussed.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics, 25*, 25-29.

Ashburner, M., Ball, C. A., Blake, J. A., Butler, H., Cherry, J. M., Corradi, J., Dolinski, K., Eppig, J. T., Harris, M., Hill, D. P., Lewis, S., Marshall, B., Mungall, C., Reiser, L., Rhee, S., Richardson, J. E., Richter, J., Ringwald, M., Rubin, G. M., Sherlock, G., Yoon, J., & Consortium, G. O. (2001). Creating the gene ontology resource: Design and implementation. *Genome Research, 11*, 1425-1433.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*, 421-441.

Bechtel, W., & Richardson, R. C. (1993/2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Cambridge, MA: MIT Press. 1993 edition published by Princeton University Press.

Braillard, P.-A. (2010). Systems biology and the mechanistic framework. *History and philosophy of the life sciences, 32*, 43-62.

Chervitz, S. A., Hester, E. T., Ball, C. A., Dolinski, K., Dwight, S. S., Harris, M. A., Juvik, G., Malekian, A., Roberts, S., Roe, T., Scafe, C., Schroeder, M., Sherlock, G., Weng, S., Zhu, Y., Cherry, J. M., & Botstein, D. (1999). Using the Saccharomyces Genome Database (SGD) for analysis of protein similarities and structure. *Nucleic Acids Research, 27*, 74-78.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L., Toufighi, K., Mostafavi, S., Prinz, J., St Onge, R. P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahr, S., Brost, R. L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z. Y., Liang, W., Marback, M., Paw, J., San Luis, B. J., Shuteriqi, E., Tong, A. H., van Dyk, N., Wallace, I. M., Whitney, J. A., Weirauch, M. T., Zhong, G., Zhu, H., Houry, W. A., Brudno, M., Ragibizadeh, S., Papp, B., Pal, C., Roth, F. P., Giaever, G., Nislow, C., Troyanskaya, O. G., Bussey, H., Bader, G. D., Gingras, A. C., Morris, Q. D., Kim, P. M., Kaiser, C. A., Myers, C. L., Andrews, B. J., & Boone, C. (2010). The genetic landscape of a cell. *Science, 327*, 425-431.

Craver, C. F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. Chicago: University of Chicago Press.

Dutkowski, J., Kramer, M., Surma, M. A., Balakrishnan, R., Cherry, J. M., Krogan, N. J., & Ideker, T. (2013). A gene ontology inferred from molecular networks. *Nature Biotechnology, 31*, 38-45.

Gene Ontology Consortium. (2015). Gene Ontology Consortium: going forward. *Nucleic Acids Research, 43*, D1049-D1056.

Gross, F. (2011). What Systems Biology Can Tell Us about Disease. *History and Philosophy of the Life Sciences, 33*, 477-496.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies, 43*, 907-928.

Guénolé, A., Srivas, R., Vreeken, K., Wang, Z. Z., Wang, S. Y., Krogan, N. J., Ideker, T., & van Attikum, H. (2013). Dissection of DNA Damage Responses Using Multiconditional Genetic Interaction Maps. *Molecular Cell, 49*, 346-358.

Huang, S. (2011). Systems biology of stem cells: three useful perspectives to help overcome the paradigm of linear pathways. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 366*, 2247-2259.

Huneman, P. (2010). Topological explanations and robustness in biological sciences. *Synthese, 177*, 213-245.

Ideker, T., & Krogan, Nevan J. (2012). Differential network biology. *Molecular Systems Biology, 8*, 565.

Kramer, M. H. (2016). *Transformation of high-throughput data into hierarchical cellular models enables biological prediction and discovery.* Ph.D., Biomedical Science, University of California, San Diego, LaJolla, CA.

Kramer, M. H., Dutkowski, J., Yu, M. K., Bafna, V., & Ideker, T. (2014). Inferring gene ontologies from pairwise similarity data. *Bioinformatics, 30*, i34-42.

Kramer, M. H., Farre, J.-C., Mitra, K., Yu, M. K., Ono, K., Demchak, B., Licon, K., Flagg, M., Balakrishnan, R., Cherry, J. M., Subramani, S., & Ideker, T. (2017). Active interaction mapping reveals the hierarchical organization of autophagy. *Molelcular Cell, 65*, 761-774 e765.

Lee, I., Li, Z., & Marcotte, E. M. (2007). An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae. *PLoS One, 2*, e988.

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.

Leonelli, S. (2012). Classificatory Theory in Data-intensive Science: The Case of Open Biomedical Ontologies. *International Studies in the Philosophy of Science, 26*, 47-65.

Leonelli, S., Diehl, A. D., Christie, K. R., Harris, M. A., & Lomax, J. (2011). How the gene ontology evolves. *Bmc Bioinformatics, 12*.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., & Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature, 431*, 308-312.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1-25.

Park, Y., & Bader, J. S. (2011). Resolving the structure of interactomes with hierarchical agglomerative clustering. *Bmc Bioinformatics, 12*.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Miklos, G. L. G., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., Cherry, J. M., Henikoff, S., Skupski, M. P., Misra, S., Ashburner, M., Birney, E., Boguski, M. S., Brody, T., Brokstein, P., Celniker, S. E., Chervitz, S. A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R. F., Gelbart, W. M., George, R. A., Goldstein, L. S. B., Gong, F. C., Guan, P., Harris, N. L., Hay, B. A., Hoskins, R. A., Li, J. Y., Li, Z. Y., Hynes, R. O., Jones, S. J. M., Kuehl, P. M., Lemaitre, B., Littleton, J. T., Morrison, D. K., Mungall, C., O'Farrell, P. H., Pickeral, O. K., Shue, C., Vosshall, L. B., Zhang, J., Zhao, Q., Zheng, X. Q. H., Zhong, F., Zhong, W. Y., Gibbs, R., Venter, J. C.,

Adams, M. D., & Lewis, S. (2000). Comparative genomics of the eukaryotes. *Science, 287*, 2204-2215.

Yu, Michael K., Kramer, M., Dutkowski, J., Srivas, R., Licon, K., Kreisberg, Jason F., Ng, Cherie T., Krogan, N., Sharan, R., & Ideker, T. (2016). Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems. *Cell Systems, 2*, 77-88.