Published in final edited form as:
IET Syst Biol. 2008 September ; 2(5): 206–221. doi:10.1049/iet-syb:20070075.

Network integration and graph analysis in mammalian molecular systems biology

A. Ma'ayan
Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, One Gustave Levy Place, Box 1215, New York, NY 10029-6574, USA

Abstract

Abstraction of intracellular biomolecular interactions into networks is useful for data integration and graph analysis. Network analysis tools facilitate predictions of novel functions for proteins, prediction of functional interactions and identification of intracellular modules. These efforts are linked with drug and phenotype data to accelerate drug-target and biomarker discovery. This review highlights the currently available varieties of mammalian biomolecular networks, and surveys methods and tools to construct, compare, integrate, visualise and analyse such networks.

1 Introduction

The exponential accumulation of molecular-biological intracellular data has the promise to advance biomedical sciences into a stage so as to understand and track most important events that regulate mammalian cells under normal and pathophysiological conditions. This would permit the development of a new generation of personalised therapeutics [1] as well as open doors in synthetic biology [2]. Physicists, mathematicians and engineers are increasingly engaging in systems biology. In this trend, experts bring tools from different disciplines to model intracellular complexity. Modelling efforts can be divided into three categories: network inference, dynamical modelling and graph analysis [3]. In contrast to graph analysis, network inference and dynamical modelling need quantitative details and/or large data sets to build models [4]. The main challenge with network inference and dynamical modelling methods is that many model realisations can fit the same data. Hence, the question commonly asked is: 'how do we really know whether the model represents the real system under investigation since there could be many alternative models that can fit the same data?' Much of current available and rapidly accumulating experimental data attempting to capture intracellular regulation is qualitative, noisy, inaccurate and incomplete. Hence, validating models from an assemblage of possible models is difficult. Alternatively and complementarily, network integration and graph analysis highlighted in this review represent a practical alternative to network inference and dynamical simulations. Some of the challenges within this research domain are: 'how to project lists of genes or proteins identified in multivariate experiments onto large-scale known intracellular interaction networks?', 'how to integrate different networks so they can be used as background knowledge to fill in missing gaps not captured experimentally?' and 'how to develop heuristics to overcome the NP-hardness of the graph search problem?' Effective data integration with filtering, graph querying and visualisation tools are key components for success in this subfield of systems biology [5].

The abstract representation of interactions within a cell to networks (formally graphs) is becoming a conventional approach to deal with the large volume of data collected through

Email: avi.maayan@mssm.edu.

emerging high-throughput technologies [6–8], and low-throughput studies reported in the research literature integrated and abstracted into networks. Different biological networks can be represented by different types of graphs (Fig. 1) [9]. For example, protein–protein interaction networks can be represented as undirected graphs where nodes are proteins and edges represent direct physical interactions. Gene-regulatory networks can be abstracted to directed graphs where nodes are genes encoding transcription factors (or other types of proteins) and links represent transcriptional regulation. Metabolic networks can be represented as bipartite graphs where nodes are separated into two sets: enzymes and substrates [10]. Although different graphs are used for different networks, the abstraction to networks helps with data integration [11]. For example, Tanay *et al.* [12] used a bipartite graph to integrate different 'omics' data by using yeast genes as anchors. Most efforts in reconstructing in-silico regulatory networks are for model organisms, but it is recognised that mammalian cellular networks are critically needed to facilitate biomedical breakthroughs. This review will focus mostly on data sets and tools that deal with mammalian biomolecular intracellular networks.

2 Different types of networks

2.1 Mammalian protein–protein interaction networks

Several repositories collect, store and share experimentally determined mammalian protein–protein interactions. Such networks, stored as undirected graphs [13], include, for example: HPRD [14–16], MINT [17,18], IntAct [19,20], Reactome [21], DIP [22,23] and BioGrid [24]. Other similar databases infer mammalian protein–protein interactions using orthologs. Here, protein–protein interactions identified in lower organisms are inferred to also exist in mammalian cells. Such efforts include OPHID [25], HPID [26], IntNetDB [27], STRING [28] and POINT [29]. Data warehouses consolidate different protein–protein databases by merging networks stored in different formats. These include, for example, UniHI [30,31] which integrates both experimentally and computationally determined interactions; and cPath [32] or BioNetBuilder [33] which are Cytoscape [34] plug-ins that integrate interaction data stored in PSI-MI format [35]. Systems such as Atlas [36], BioWarehouse [37], BIOZON [38,39], INTEGRATOR [40] and the Gaggle [41] provide integration and querying capabilities with other types of biological data in addition to protein–protein interactions, metabolic, gene-regulation and cell-signalling networks. Several studies compared different mammalian protein–protein interactions databases to assess their overlap and coverage [31,42].

2.2 Considering structure in protein– protein interaction networks

Early efforts in reconstructing in silico protein–protein interaction networks did not consider that most interactions are mediated through structural domains. The fields of structure biology and network biology are coming closer as more attention is given to structural elements of protein– protein interaction networks [43]. Once structural domains are considered, the domains network underneath the protein–protein interaction network can be delineated [44]. Complementarily, databases and algorithms have been developed to predict protein–protein interactions based on structure and sequence [45]. Additionally, databases of experimentally determined protein–protein interactions specific for a structural domain also exist [46]. Careful analysis of protein–protein and domain-domain interactions in yeast showed that hubs can be divided into two distinct groups: single- and multi-binding hubs [47]. Considering structure with protein–protein interactions can be used to reconstruct macro-molecular complexes. For instance, Takamori *et al.* [48] reconstructed synaptic vesicles after carefully measuring the size of most components. Besides synaptic vesicles, the molecular organisation of macro-molecular complexes is still mostly unknown. A first step towards assembling this 3D puzzle is identifying protein–protein interactions, their localisation and the functional relationship among the components [5].

2.3 Signalling networks

In contrast to protein–protein interaction networks, cell-signalling representation as a network captures functional relationships. Signalling networks are commonly represented as directed graphs with three types of links: activation, inhibition and neutral. Besides proteins, signalling networks also include small molecules such as calcium and cAMP. Some examples are science signalling (previously STKE) [49], the Cancer Cell Map (http://cancer.cellmap. org/cellmap/), KEGG [50,51], GOLD.db [52] and BioCarta (http://biocarta.com). Signalling networks regulate the response of cells to changes in the extracellular environment where signals, received at the cell surface by receptors, transduce information to effector proteins through cascades of coupled biochemical reactions. The most common signalling reaction is phosphorylation. Databases that record phosphorylation sites (predicted computationally or determined experimentally) are critical for tracking information-flow through cell-signalling pathways. Recent efforts made substantial progress in this area [53,54]. For example, NetworKIN [54] is a web-based resource providing access to predicted as well as experimentally identified phosphorylation sites and the kinases responsible for the phosphorylations. Cell-signalling networks can be inferred directly from multivariate data using network inference methods such as Bayesian networks [55,56]. Bayesian networks are used for constructing acyclic graphs based on statistical interdependencies among measured variables [57]. Since Bayesian networks are part of the network inference modelling genre, describing them in detail is out of the scope of this review.

2.4 Gene-regulatory networks

Directed graphs with activation/inhibition links can also be used to represent gene-regulatory networks. Here, genes are translated to proteins that function as transcription factors (or more distal regulators) regulating the expression of other genes. Gene networks function at longer time scales compared with cell-signalling, metabolic or protein–protein interaction networks. Biotechnologies that can experimentally map gene-regulatory networks in high-throughput are rapidly emerging. Network inference approaches are instrumental for building networks from perturbation or time-series data. These methods are typically applied to gene expression microarrays. ChIP-chip [58] and chIP-seq [59], comparative genomics (identifying conserved non-coding sequences as potential binding sites), or purely computational approaches that use known consensus DNA binding motifs [60,61] can also be used to reconstruct in silico gene-regulatory networks. Several tools are developed to integrate such data to allow novice users easy access to identify gene-regulatory interactions For example, MYBS [62], YEASTRACT [63], SGD [64,65], SCPD [66], TRANSFAC [67], MAPPER [68], TRANSCompel [69], TRRD [70,71] and SWISS-REGULON [72] are web-based tools providing a user interface for an underlying gene-regulatory network database.

2.5 Metabolic networks

Metabolic networks are in general more complete and rich in quantitative information as compared with protein–protein, cell-signalling and gene-regulatory networks. BioCyc [73] and its subset database MetaCyc [74-76] are comprehensive resources for metabolic networks in many organisms. The *Escherichia coli* metabolic network is the most complete metabolic network from an experimentally, in silico predicted, and computationally analysed perspectives. It was extensively mapped by Palsson and co-workers [77–79]. A metabolic network for yeast was later similarly reconstructed by Förster *et al.* [80] and was compared with the *E. coli* metabolic network.

2.6 MicroRNA networks

Besides protein–protein, cell-signalling, gene-regulation and metabolic networks there is a growing appreciation for non-canonical metabolites, non-protein biomolecules and non-

conventional post-translational modifications that function in intracellular regulation. One example is miRNA networks. miRNAs are short (~22 nucleotide) transcripts that pair with (full-length) mRNAs of transcribed and translated genes and thereby suppressing their translation into proteins [81]. Since these transcripts have known sequence, it is computationally simple to identify the network of interactions between miRNAs and the expressed genome. Shalgi *et al.* [82] developed and analysed a network of transcription factors and miRNAs. Cui *et al.* [83] used a large-scale cell-signalling network extracted manually from research literature [84] to assess how endogenous miRNAs target and regulate components in the cell-signalling system.

3 Data acquisition and representation

3.1 Automatic extraction of interactions from literature

A large amount of knowledge about functional regulatory interactions and the components involved in these interactions is embedded in the biomedical research literature from the past 30–40 years [85]. Text mining is used to extract interactions using natural language processing (NLP) and information retrieval technologies [86]. The first step in this process consists of extracting biological terms [87–89]. Protein and gene names and other biological entities are organised into dictionaries [90]. It is important to resolve ambiguities in entity naming, for example, resolving synonymous names for proteins and genes [91–95]. Unique terms in biomedical text often represent a biological entity [92], whereas co-occurrence is often used to resolve ambiguity in names [96]. A related effort is to automatically assign gene ontology (GO) annotation for biological terms [97,98]. Taggers, such as ABNER [99], can be used to highlight different entities in biomedical text, and systems such as AliBaba [100] make use of tagged text to build networks from key terms in abstracts. iHOP [101,102] tags biomedical text with the ability to navigate on the web from one highlighted term to another by clicking on hyperlinked terms. The most sophisticated systems, for example, GeneWays [103] and PathwayStudio [104,105], use NLP to extract interactions.

Automatic literature search can be combined with data analysis of microarray gene expression profiling by identifying literature-based relationships between co-regulated genes [106]. Text mining of disease phenotypes can be automatically linked to protein–protein interaction networks in order to identify enrichment of human phenotypes that correlate with disease genes [107]. OMIM's morbid map is commonly used as a resource for text mining tools that attempt to integrate diseases with biomolecular networks. OMIM is an NCBI resource that mines relationships between genes and disease phenotypes [108]. Text mining cannot be completely covered in this review. For additional information on this topic, readers may find the review by Krallinger and Valencia [109] as a helpful start.

3.2 Visualisation and interoperability

Efforts of building biomolecular networks entail the challenges of visualisation and interoperability. There is a rapid emergence of desktop and web-based applications for pathway and network visualisation [110,111]. For example, consider the systems VisANT [112,113], PATIKAweb [114], Cytoscape [34], CellDesigner [115] and AVIS, a light-weight viewer that uses the Google Gadget API to automatically visualise cell-signalling pathways [116], for network visualisation. Visualisation tools support different network storage formats. For example, PIMWalker [117] supports visualising data stored in PSI-MI [35,118]; The systems biology markup language (SBML) was extended to provide information needed for network visualisation. Standard storage schemas such as SBML are important for interoperability.

Interoperability efforts attempt to develop standards for data sharing and exchange between isolated data sets and analysis tools. Many schemas are used to represent biomolecular

intracellular networks. These include SBML [119], CellML [120], BioPAX [121], KGML [50] and PSI-MI [35]. All these formats use XML [122] which provides a flexible way to store data in a structured format with semantics about the data captured within the storage schema. Each storage schema listed above is geared towards handling different types of biomolecular networks [121]. For example, PSI-MI is mostly useful for describing details about experiments, SBML is useful for directly exporting networks into quantitative modelling tools such as the SBMLToolbox [123] or others [124]. BioPAX does not require quantitative information. BioPAX is useful for network visualisation as well as data exchange. Some databases and their tools develop their own XML schemas. These include, for example, VisML developed for VisANT [112] and KGML [125] developed for KEGG [51]. Although, many standards exist, attention is still given to improving them as well as standardising and expanding existing standards [126].

3.3 Gene ontology

It is realised that interoperability efforts will greatly benefit from the development of ontologies. Ontologies are a set of terms that describe entities with encoded conceptual relationships between entities assembled and organised for specific knowledge domains. The gene ontology (GO) consortium attempts to provide controlled vocabulary and hierarchical relations for knowledge representation of function, cellular component and involvement in biological process for categorising genes and proteins [127–129]. GO is a part of a greater effort towards developing open biomedical ontologies [130] for a variety of biomedical and biological domains. GO has been useful for organising and extracting functional relationships between groups of genes. Genes and proteins identified experimentally are classified based on their common annotated functions. GO is now integrated with most leading gene and protein databases [131,132].

4 Tools and methods for network analysis

4.1 GO tools

The annotation of genes and proteins into GO terms was conducive to the development of many GO analysis tools. The general theme of these tools is the effort of identifying common functions for group of genes. GOLEM [133], GOlorize [134], DyGO [135], GObar [136], WEGO [137] and GoSurfer [138] are tools for visually exploring and analysing groups of genes using GO. BiNGO [139], GOToolBox [140], GOstat [141] and GO::TermFinder [142] can be used to identify over-representation of GO terms in groups of genes. For example, the most popular approach is to apply GO analysis to groups of genes that are either up-regulated or down-regulated in microarray experiments [143–146]. Tools such as DAVID [147] and PathExpress [148] go a step further and provide linkage to pathways using the KEGG database. The Blast2GO tool is an example of linking sequence with GO annotations [149].

Developing ontologies for genes was done first. Efforts are now shifting to developing ontologies for the relations between genes. GO is being extended to include interaction/pathway ontology. Interaction-ontologies are less developed but are needed. Lu *et al.* [150] proposed an ontology for classifying interactions. Science signalling database developed a database named CMADES which contains controlled hierarchical vocabularies that cover most types of reactions and their functional effects. This information can be readily converted into a standard ontology. BioPAX [121] is a leading interaction-exchange standard in the field that uses ontologies. Once ontological relations and ontological events for links are specified, these can become a set of logical models and potentially dynamical models. For example, the INOH pathway database (http://www.inoh.org/ontology-viewer/) attempts to convert static pathways into dynamical event systems through the use of event ontologies. Systems such as HyBrow (hypotheses browser) use ontologies and experimental data to build dynamical testable

hypotheses that can be used to design experiments [151]. Although such systems are currently at a prototype level, it is expected that they will improve in the future. BioSigNet [152] and PathwayLogic (http://www.csl.sri.com/projects/pathwaylogic/) are two other examples of developing a logical languages to describe relationships between cellular components in the context of regulatory networks. Alternatively, unified modelling language (UML) is a method from software engineering that is well-accepted for developing complex software systems. To handle the inherent complexity of large-scale software systems, UML unify eight different design views of a system before the software is implemented [153]. This approach has been suggested as a potential language for representing and modelling biochemical networks [154]. The above-mentioned efforts link network integration and graph analysis towards dynamical modelling. The advantage of these approaches is their ability to embed complex reasoning through knowledge representation in a standard way to describe biomolecular regulatory networks in standard formulation.

4.2 Prediction of GO terms based on network topology

Many proteins and genes do not have GO annotation. This means that these genes were identified experimentally but do not yet have an assigned function. Most computational approaches to predict function for genes use sequence similarity; but these efforts are gradually augmented by network topology-based approaches. For an excellent review on this endeavour, see [155]. Predicting function for genes with unassigned function using protein–protein interaction networks is based on the observation that proteins that are known to interact, often share GO terms. The network-based prediction of protein function can be categorised into two groups: direct or module assisted [155]. The direct method explores the protein–protein interaction neighbourhood around the uncategorised gene to assess the functional category most prevalent in the node's immediate neighbourhood. If a certain functional category is highly over-represented, it is used as a prediction for the function of the gene with the unassigned function. There are several algorithms that can be used for this purpose, for example, Markov random fields (MRFs). MRFs are used to identify neighbourhoods around a gene by applying a Markov random walk. A random walker, starting from the gene with unassigned function, is travelling randomly on edges and nodes from the protein–protein interaction network to visit nearby nodes with already assigned function [156,157]. A simpler approach is to look at the enrichment of GO terms in the first-level neighbours [158]. Chua *et al.* [159] showed that looking at different combinations of sets of first and second neighbours can significantly improve the functional prediction quality.

4.3 Comparing different networks

Concerns were raised when it was found that two full-genome high-throughput yeast-2-hybrid screens that attempted to characterise protein–protein interactions in yeast showed little overlap [160]. Protein–protein interaction networks from different sources have been compared and evaluated to identify biases, overlap and sources for false-positives and false-negatives [161]. Comparing networks across species is useful for predicting interactions and function for proteins using orthologs [162]. Several tools and algorithms have been developed for comparing and aligning different networks. These include NetAlign [163,164], PathBLAST [165] and others [166,167]. These tools combine interaction topology and sequence similarity to identify conserved network substructures across organisms, across networks and within the same network.

4.4 Network motifs and graphlets

Network motifs are small circuits of interacting components in directed graphs that are found to be highly overrepresented in real networks compared with counts of the same motifs in shuffled networks created from the original networks. The different possibilities for links

among few nodes (i.e. 3–6) define different types of network motifs [168,169]. Alon and co-workers were the first to introduce the network motif concept for analysing the topology of intracellular gene-regulatory networks. They analysed the gene-regulatory networks of *Saccharomyces cerevisiae* and *E. coli* to identify signature patterns of motifs in those networks. Przulj *et al.* [170] used a similar approach to analyse protein–protein interaction networks. Instead of counting motifs, Przulj and co-workers searched for graphlets. Graphlets are similar to network motifs but are defined for undirected graphs. Network motifs identified in mammalian cell-signalling networks showed some similarities with motif patterns identified in gene-regulatory networks [84]. The bifan motif [171], a four-node motif connecting two source nodes to two targets nodes, was found to be the most highly abundant motif in most intracellular networks studied so far. This is probably due to the large number of isoforms resulting from the process of duplication–divergence.

The concept of network motifs was found to be useful for making predictions. Albert and Albert [172] combined motif search algorithms with the SUGGEST machine-learning algorithm to predict interactions. Similarly, Yu *et al.* [173] used ‘defective cliques’ to predict interactions. Bu *et al.* [174] used network motifs identified in protein–protein networks to predict the function of proteins with unassigned GO functional classification.

4.5 Identifying modules in networks

Many real-world networks [175,176], including intracellular networks, were shown to be organised in modules [177–180]. These modules can be identified with network clustering algorithms such as betweenness centrality clustering [175,178,181–184]. Betweenness centrality is computed for each node or link by counting the number of times shortest paths go through the node or link [185]. This method allows identification of clusters by finding the nodes and links that connect clusters. Such nodes and links have relatively low connectivity but many shortest paths go through them. Simpler methods for network clustering use the shortest path length or the number of shared neighbours as the distance measure needed for finding clusters [177]. For example, Rives and Galitski [177] defined the distance between all pairs of nodes as a transformed shortest path $1/d^2$, where d is the shortest path distance. The reason for not using d directly is to emphasise shorter distances. Many more complicated methods for finding clusters in networks exist. For example, Frey and Dueck developed a method using message passing [186]. Sen *et al.* [187] used eigenmodes of the connectivity matrix and applied it to cluster the yeast protein–protein interaction network. Another approach is to compare real networks to randomly wired networks [176]. With these methods, deviation from random connectivity towards modular structure can be identified.

Once clusters have been identified, their strength can be quantified. Radicchi *et al.* [183] defined a strong community structure if nodes in the community have more connections within the community than with other nodes in the network. This measure was inspired by the analysis of web communities [188]. After modules have been identified with only considering the topology, the modules can be validated by observing if the components in the module share similar GO terms [178,189]. Instead of using GO just for validation, Lubovac *et al.* [190] combined GO terms and network connectivity for module identification. Several software systems can be used to assist non-specialists to identify modules in networks. For example, MoNet [191] is a Java implementation of the Girvan– Newman [175] betweenness algorithm. MCODE [192] is a tool that uses the concept of clustering coefficient to identify network clusters [193]. MCL [194] uses a Markov clustering algorithm. As different network clustering algorithms are developed, benchmarks to evaluate their performance are important for their evaluation [195].

4.6 Expanding the neighbourhood around seed lists

Using background knowledge encoded into biological intracellular networks such as metabolic, protein–protein interactions, gene-regulation or signalling pathways, it is possible to expand interactions around the neighbourhood of seed lists of genes or proteins, as long as they exist as nodes in the large background network [27,196]. This concept was applied to analyse metabolic networks [197]. Canonical signalling pathways were enriched using information from a protein–protein interaction network [198]. The same concept was also applied to identify disease genes modules [199]. Asthana *et al.* [200] used protein–protein interaction networks from multiple sources to fill in gaps and reconstruct protein complexes. A similar concept was applied to rank seed lists of genes based on their importance. The famous PageRank algorithm was applied to analyse lists of seed genes identified in microarray experiments in context of a background protein–protein interaction network where the genes are ranked based on their 'importance'. Importance is defined based on gene degree of connectivity and also based on the connectivity of the gene's direct neighbours [201]. Morrison *et al.* [202] used a similar method with a background network created using the GO database instead of using interaction networks.

Projecting lists of seed nodes onto a background network can be applied using different algorithms. For example, finding all shortest paths between all pairs of nodes in the seed list [203], finding the Steiner tree [204,205], finding the minimum spanning tree, or expanding the neighbours around seed nodes using a random walker. Scott *et al.* [196] implemented Steiner trees to connect seed lists of genes shown to be altered in microarray experiments using a background protein–protein interaction network. Steiner trees, commonly used to design telecommunication networks, are minimal spanning trees where intermediate nodes can be used to connect terminal (seed) nodes.

4.7 Seed lists from genome-wide association studies

Genome-wide association studies are used to identify mutations in human genes that can be linked to disease propensity [206–208]. By comparing single-nucleotide polymorphisms (SNPs) from healthy individuals originated from the same demographic background, with individuals with a known common disease, SNPs in the disease group can be identified. Lists of genes with SNPs for a particular disease are rapidly emerging. For example, The Wellcome Trust Control Consortium analysed seven groups of 2000 patients having common diseases including: bipolar disorder, coronary artery disease, Crohn's disease, rheumatoid arthritis and type 1 and 2 diabetes [206]. They found 24 genes with mutations that correlate with disease susceptibility. The cooperation between groups that engage in genome-wide association studies is critical for identifying SNPs because of the high cost of sequencing and the statistical power that requires large samples. The International HapMap consortium is a large-scale collaboration project that records and shares, through a central database, data on SNPs from four different demographically homogeneous populations [207,208]. Databases reporting genes with their disease associations are emerging [209]. Once disease genes are identified, disease gene lists can be projected onto biomolecular networks to identify modules that are perturbed in disease. In other words, a list of genes associated with a disease could serve as seed for graph analysis. This approach can be used to identify additional disease genes and potentially novel drug targets [210].

4.8 Linking human diseases, disease genes, drugs and drug targets

The rapid identification of disease genes allows for the construction of a human–disease/human–disease-gene bipartite network [211]. This network is useful for identifying global relationship between different diseases. This type of network analysis would lead to the identification of functional modules of interacting disease genes and be used to predict additional disease gene candidates [199,210,212]. At the same time, networks of drugs and

drug targets can be developed [213,214]. Combining and analysing such networks is a valuable initial step towards finding novel ways to reuse approved drugs and to better understand side effects.

5 Conclusions

Network integration involves data mining and data standards at a bottom layer which addresses the need for interoperability. Graph analysis tools utilise the fused data sets at a top layer. Network integration and graph analysis of intracellular networks is only one effort in a broader biomedical science revolution that involves breakthroughs in genomics, structural biology and imaging [215] (Fig. 2). Although network integration and graph analysis deal with qualitative data, and as such, the analysis and the representation miss many important aspects of cellular regulation, this approach provides means to handle diverse and massive data sets more easily, and can produce useful predictions that can be validated experimentally [216]. Network representation of intracellular complex systems can be integrated with networks of drugs, disease phenotypes and side effects. Identifying additional novel components that participate in pathways in mammalian cells can be fruitful for understanding disease mechanisms, for identifying new biomarkers and for discovering novel drug targets.

Acknowledgments

This research was supported by NIH Grant No. 1P50GM071558-01A27398 and start-up fund from Mount Sinai School of Medicine to A.M. A.M. would like to thank the anonymous reviewers for their very helpful comments and suggestions.

References

1. Hood L, Heath JR, Phelps ME, Lin B. Systems biology and new technologies enable predictive and preventative medicine. Science 2004;306(5696):640–643. [PubMed: 15499008]
2. Endy D. Foundations for engineering biology. Nature 2005;438(7067):449–453. [PubMed: 16306983]
3. Albert R. Network inference, analysis, and modeling in systems biology. Plant Cell 2007;19(11):3327–3338. [PubMed: 18055607]
4. Bornholdt S. Systems biology. Less is more in modeling large genetic networks. Science 2005;310 (5747):449–451. [PubMed: 16239464]
5. Ma'ayan A, Iyengar R. From components to regulatory motifs in signalling networks. Brief Funct Genomic Proteomic 2006;5(1):57–61. [PubMed: 16769680]
6. Zhu X, Gerstein M, Snyder M. Getting connected: analysis and principles of biological networks. Genes Dev 2007;21(9):1010–1024. [PubMed: 17473168]
7. Albert R. Scale-free networks in cell biology. J Cell Sci 2005;118(21):4947–4957. [PubMed: 16254242]
8. Carter GW. Inferring network interactions within a cell. Brief Bioinf 2005;6(4):380–389.
9. Ma'ayan A, Blitzer RD, Iyengar R. Toward predictive models of mammalian cells. Ann Rev Biophys Biomol Struct 2005;34(1):319–349. [PubMed: 15869393]
10. Patil KR, Nielsen J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. PNAS 2005;102(8):2685–2689. [PubMed: 15710883]
11. Joyce AR, Palsson BO. The model organism as a system: integrating ‘omics’ data sets. Nat Rev Mol Cell Biol 2006;7(3):198–210. [PubMed: 16496022]
12. Tanay A, Sharan R, Kupiec M, Shamir R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. PNAS 2004;101(9):2981–2986. [PubMed: 14973197]
13. Mathivanan S, Periaswamy B, Gandhi TKB, et al. An evaluation of human protein–protein interaction data in the public domain. BMC Bioinf 2006;7(suppl 5):S19.
14. Mishra GR, Suresh M, Kumaran K, et al. Human protein reference database–2006 update. Nucl Acids Res 2006;34(suppl 1):D411–D414. [PubMed: 16381900]

15. Peri S, Navarro JD, Kristiansen TZ, et al. Human protein reference database as a discovery resource for proteomics. Nucl Acids Res 2004;32(suppl 1):D497–D501. [PubMed: 14681466]
16. Peri S, Navarro JD, Amanchy R, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 2003;13(10):2363–2371. [PubMed: 14525934]
17. Chatr-Aryamontri A, Ceol A, Palazzi LM, et al. MINT: the molecular INTeraction database. Nucl Acids Res 2007;35(suppl 1):D572–D574. [PubMed: 17135203]
18. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a molecular INTeraction database. FEBS Lett 2002;513(1):135–140. [PubMed: 11911893]
19. Hermjakob H, Montecchi-Palazzi L, Lewington C, et al. IntAct: an open source molecular interaction database. Nucl Acids Res 2004;32(suppl 1):D452–D455. [PubMed: 14681455]
20. Kerrien S, Alam-Faruque Y, Aranda B, et al. IntAct – open source resource for molecular interaction data. Nucl Acids Res 2007;35(suppl 1):D561–D565. [PubMed: 17145710]
21. Joshi-Tope G, Gillespie M, Vastrik I, et al. Reactome: a knowledge base of biological pathways. Nucl Acids Res 2005;33(suppl 1):D428–D432. [PubMed: 15608231]
22. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: the database of interacting proteins. Nucl Acids Res 2000;28(1):289–291. [PubMed: 10592249]
23. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. Nucl Acids Res 2004;32(suppl 1):D449–D451. [PubMed: 14681454]
24. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucl Acids Res 2006;34(suppl 1):D535–D539. [PubMed: 16381927]
25. Brown KR, Jurisica I. Online predicted human interaction database. Bioinformatics 2005;21(9):2076–2082. [PubMed: 15657099]
26. Han K, Park B, Kim H, Hong J, Park J. HPID: the human protein interaction database. Bioinformatics 2004;20(15):2466–2470. [PubMed: 15117749]
27. Xia K, Dong D, Han JD. IntNetDB v1.0: an integrated protein–protein interaction network database generated by a probabilistic model. BMC Bioinf 2006;7(1):508.
28. Von Mering C, Jensen LJ, Kuhn M, et al. STRING 7 – recent developments in the integration and prediction of protein interactions. Nucl Acids Res 2007;35(suppl 1):D358–D362. [PubMed: 17098935]
29. Huang TW, Tien AC, Huang WS, et al. POINT: a database for the prediction of protein –protein interactions based on the orthologous interactome. Bioinformatics 2004;20(17):3273–3276. [PubMed: 15217821]
30. Chaurasia G, Iqbal Y, Hanig C, Herzel H, Wanker EE, Futschik ME. UniHI: an entry gate to the human protein interactome. Nucl Acids Res 2007;35(suppl 1):D590–D594. [PubMed: 17158159]
31. Futschik ME, Chaurasia G, Herzel H. Comparison of human protein–protein interaction maps. Bioinformatics 2007;23(5):605–611. [PubMed: 17237052]
32. Cerami E, Bader G, Gross B, Sander C. cPath: open source software for collecting, storing, and querying biological pathways. BMC Bioinf 2006;7(1):497.
33. Avila-Campillo I, Drew K, Lin J, Reiss DJ, Bonneau R. BioNetBuilder: automatic integration of biological networks. Bioinformatics 2007;23(3):392–393. [PubMed: 17138585]
34. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13(11):2498–2504. [PubMed: 14597658]
35. Hermjakob H, Montecchi-Palazzi L, Bader GC, et al. The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data. Nat Biotech 2004;22(2): 177–183.
36. Shah S, Huang Y, Xu T, Yuen M, Ling J, Ouellette BFF. Atlas – a data warehouse for integrative bioinformatics. BMC Bioinf 2005;6(1):34.
37. Lee T, Pouliot Y, Wagner V, et al. BioWarehouse: a bioinformatics database warehouse toolkit. BMC Bioinf 2006;7(1):170.
38. Birkland A, Yona G. BIOZON: a system for unification, management and analysis of heterogeneous biological data. BMC Bioinf 2006;7(1):70.

39. Birkland A, Yona G. BIOZON: a hub of heterogeneous biological data. Nucl Acids Res 2006;34 (suppl 1):D235–D242. [PubMed: 16381854]
40. Chang A, Mcdermott J, Frazier Z, Guerquin M, Samudrala R. INTEGRATOR: interactive graphical search of large protein interactomes over the web. BMC Bioinf 2006;7(1):146.
41. Shannon P, Reiss D, Bonneau R, Baliga N. The gaggle: an open-source software system for integrating bioinformatics software and data sources. BMC Bioinf 2006;7(1):176.
42. Chaurasia G, Herzel H, Wanker EE, Futschik ME. Systematic functional assessment of human protein–protein interaction maps. Genome Inf 2006;17(1):36–45.
43. Beltrao P, Kiel C, Serrano L. Structures in systems biology. Curr Opin Struct Biol 2007;17(3):378–384. [PubMed: 17574836]
44. Moon HS, Bhak J, Lee KH, Lee D. Architecture of basic building blocks in protein and domain structural interaction networks. Bioinformatics 2005;21(8):1479–1486. [PubMed: 15613386]
45. Gong S, Yoon G, Jang I, et al. PSIbase: a database of protein structural interactome map (PSIMAP). Bioinformatics 2005;21(10):2541–2543. [PubMed: 15749693]
46. Beuming T, Skrabanek L, Niv MY, Mukherjee P, Weinstein H. PDZBase: a protein–protein interaction database for PDZ-domains. Bioinformatics 2005;21(6):827–828. [PubMed: 15513994]
47. Kim PM, Lu LJ, Xia Y, Gerstein MB. Relating three-dimensional structures to protein networks provides evolutionary insights. Science 2006;314(5807):1938–1941. [PubMed: 17185604]
48. Takamori S, Holt M, Stenius K, et al. Molecular anatomy of a trafficking organelle. Cell 2006;127 (4):831–846. [PubMed: 17110340]
49. Gough NR. Science's signal transduction knowledge environment: the connections maps database. Ann Ny Acad Sci 2002;971:585–587. [PubMed: 12438188]
50. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucl Acids Res 2000;28 (1):27–30. [PubMed: 10592173]
51. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucl Acids Res 1999;27(1):29–34. [PubMed: 9847135]
52. Hackl H, Maurer M, Mlecnik B, et al. GOLD.db: genomics of lipid-associated disorders database. BMC Genomics 2004;5(1):93. [PubMed: 15588328]
53. Diella F, Cameron S, Gemund C, et al. Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. BMC Bioinf 2004;5(1):79.
54. Linding R, Jensen LJ, Ostheimer GJ, et al. Systematic discovery of in vivo phosphorylation networks. Cell 2007;127(7):1415–1426. [PubMed: 17570479]
55. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science 2005;308(5721):523–529. [PubMed: 15845847]
56. Woolf PJ, Prudhomme W, Daheron L, Daley GQ, Lauffenburger DA. Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. Bioinformatics 2005;21(6):741–753. [PubMed: 15479714]
57. Cooper GE, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Mach Learn 1992;9:309–347.
58. Horak CE, Mahajan MC, Luscombe NM, Gerstein M, Weissman SM, Snyder M. GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp –chip analysis. PNAS 2002;99(5): 2924–2929. [PubMed: 11867748]
59. Robertson G, Hirst M, Bainbridge M, et al. Genome-wide profiles of STAT1 Dna association using chromatin immunoprecipitation and massively parallel sequencing. Nat Meth 2007;4(8):651–657.
60. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 2003;423(6937):241–254. [PubMed: 12748633]
61. Harbison CT, Gordon DB, Lee TI, et al. Transcriptional regulatory code of a eukaryotic genome. Nature 2004;431(7004):99–104. [PubMed: 15343339]
62. Tsai HK, Chou MY, Shih CH, Huang GT, Chang TH, Li WH. MYBS: a comprehensive web server for mining transcription factor binding sites in yeast. Nucl Acids Res 2007;35:W221–W226. [PubMed: 17537814]

63. Teixeira MC, Monteiro P, Jain P, et al. The Yeastract database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. Nucl Acids Res 2006;34(suppl 1):D446–D451. [PubMed: 16381908]
64. Christie KR, Weng S, Balakrishnan R, et al. Saccharomyces genome database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. Nucl Acids Res 2004;32(suppl 1):D311–D314. [PubMed: 14681421]
65. Cherry JM, Adler C, Ball C, et al. SGD: Saccharomyces genome database. Nucl Acids Res 1998;26 (1):73–79. [PubMed: 9399804]
66. Zhu J, Zhang MQ. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. Bioinformatics 1999;15(7):607–611. [PubMed: 10487868]
67. Matys V, Kel-Margoulis OV, Fricke E, et al. TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. Nucl Acids Res 2006;34(suppl 1):D108–D110. [PubMed: 16381825]
68. Marinescu V, Kohane I, Riva A. MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. BMC Bioinf 2005;6(1):79.
69. Kel-Margoulis OV, Kel AE, Reuter I, Deineko IV, Wingender E. TRANSCompel(R): a database on composite regulatory elements in eukaryotic genes. Nucl Acids Res 2002;30(1):332–334. [PubMed: 11752329]
70. Kolchanov NA, Podkolodnaya OA, Ananko EA, et al. Transcription regulatory regions database (TRRD): its status in 2000. Nucl Acids Res 2000;28(1):298–301. [PubMed: 10592253]
71. Kolchanov NA, Ananko EA, Podkolodnaya OA, et al. Transcription regulatory regions database (TRRD): its status in 1999. Nucl Acids Res 1999;27(1):303–306. [PubMed: 9847210]
72. Pachkov M, Erb I, Molina N, Van Nimwegen E. SwissRegulon: a database of genome-wide annotations of regulatory sites. Nucl Acids Res 2007;35(suppl 1):D127–D131. [PubMed: 17130146]
73. Krummenacker M, Paley S, Mueller L, Yan T, Karp PD. Querying and computing with BioCyc databases. Bioinformatics 2005;21(16):3454–3455. [PubMed: 15961440]
74. Karp PD, Riley M, Paley SM, Pellegrini-Toole A. The MetaCyc database. Nucl Acids Res 2002;30 (1):59–61. [PubMed: 11752254]
75. Caspi R, Foerster H, Fulcher CA, et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucl Acids Res 2006;34(suppl 1):D511–D516. [PubMed: 16381923]
76. Krieger CJ, Zhang P, Mueller LA, et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucl Acids Res 2004;32(suppl 1):D438–D442. [PubMed: 14681452]
77. Herrgard MJ, Fong SS, Palsson B. Identification of genome-scale metabolic network models using experimentally measured flux profiles. PLos Comput Biol 2006;2(7):e72. [PubMed: 16839195]
78. Edwards JS, Ibarra RU, Palsson BO. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotech 2001;19(2):125–130.
79. Edwards JS, Palsson BO. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. PNAS 2000;97(10):5528–5533. [PubMed: 10805808]
80. FÖrster J, Famili I, Fu P, Palsson B, Nielsen J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. Genome Res 2003;13:244–253. [PubMed: 12566402]
81. Ruvkun G. Molecular biology: glimpses of a tiny Rna world. Science 2001;294(5543):797–799. [PubMed: 11679654]
82. Shalgi R, Lieber D, Oren M, Pilpel Y. Global and local architecture of the mammalian microRNA-transcription factor regulatory network. PLos Comput Biol 2007;3(7):e131. [PubMed: 17630826]
83. Cui Q, Yu Z, Purisima EO, Wang E. Principles of microrna regulation of a human cellular signaling network. Mol Syst Biol 2006;2:46. [PubMed: 16969338]
84. Ma'ayan A, Jenkins SL, Neves S, et al. Formation of regulatory patterns during signal propagation in a mammalian cellular network. Science 2005;309(5737):1078–1083. [PubMed: 16099987]
85. Kersey P, Apweiler R. Linking publication, gene and protein data. Nat Cell Biol 2006;8(11):1183–1189. [PubMed: 17060904]
86. Feldman, R.; Press, JSCU. The text mining handbook. Cambridge University Press; New York: 2006.

87. Hoffmann R, Krallinger M, Andres E, Tamames J, Blaschke C, Valencia A. Text mining for metabolic pathways, signaling cascades, and protein networks. Sci STKE 2005;2005(283):pe21. [PubMed: 15886388]

88. Skusa A, Ruegg A, Kohler J. Extraction of biological interaction networks from scientific literature. Brief Bioinf 2005;6(3):263–276.

89. Roberts PM. Mining literature for systems biology. Brief Bioinf 2006;7(4):399–406.

90. Fundel K, Zimmer R. Gene and protein nomenclature in public databases. BMC Bioinf 2006;7(1): 372.

91. Tsuruoka Y, Tsujii JI. Improving the performance of dictionary-based approaches in protein name recognition. J Biomed Inf 2004;37(6):461–470.

92. Shi L, Campagne F. Building a protein name dictionary from full text: a machine learning term extraction approach. BMC Bioinf 2005;6(1):88.

93. Schuemie MJ, Mons B, Weeber M, Kors JA. Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification. J Biomed Inf 2007;40(3):316–324.

94. Zhou G, Zhang J, Su J, Shen D, Tan C. Recognizing names in biomedical texts: a machine learning approach. Bioinformatics 2004;20(7):1178–1190. [PubMed: 14871877]

95. Zhou GD. Recognizing names in biomedical texts using mutual information independence model and Svm plus sigmoid. Int J Med Inf 2006;75(6):456–467.

96. Cohen AM, Hersh WR, Dubay C, Spackman K. Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. BMC Bioinf 2005;6(1):103.

97. Daraselia N, Yuryev A, Egorov S, Mazo I, Ispolatov I. Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks. BMC Bioinf 2007;8:243.

98. Tao Y, Sam L, Li J, Friedman C, Lussier YA. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. Bioinformatics 2007;23(13):i529–i538. [PubMed: 17646340]

99. Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 2005;21(14):3191–3192. [PubMed: 15860559]

100. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U. ALIBABA: PubMed as a graph. Bioinformatics 2006;22(19):2444–2445. [PubMed: 16870931]

101. Fernandez JM, Hoffmann R, Valencia A. ihop web services. Nucl Acids Res 2007;35(Web Server Issue):W21–W26. [PubMed: 17485473]

102. Hoffmann R, Valencia A. Implementing the ihop concept for navigation of biomedical literature. Bioinformatics 2005;21(suppl 2):ii252–ii258. [PubMed: 16204114]

103. Rzhetsky A, Iossifov I, Koike T, et al. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. J Biomed Inf 2004;37(1):43–53.

104. Yuryev A, Mulyukov Z, Kotelnikova E, et al. Automatic pathway building in biological association networks. BMC Bioinf 2006;7(1):171.

105. Nikitin A, Egorov S, Daraselia N, Mazo I. Pathway studio – the analysis and navigation of molecular networks. Bioinformatics 2003;19(16):2155–2157. [PubMed: 14594725]

106. Rubinstein R, Simon I. Milano – custom annotation of microarray results using automatic literature searches. BMC Bioinf 2005;6(1):12.

107. Vandriel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM. A text-mining analysis of the human phenome. Eur J Hum Genet 2006;14(5):535–542. [PubMed: 16493445]

108. Bajdik CD, Kuo B, Rusaw S, Jones S, Brooks-Wilson A. CGMIM: automated text-mining of Online Mendelian Inheritance in Man (OMIM) to identify genetically-associated cancers and candidate genes. BMC Bioinf 2005;6(1):78.

109. Krallinger M, Valencia A. Text-mining and information-retrieval services for molecular biology. Genome Biol 2005;6(7):224. [PubMed: 15998455]

110. Holford M, Li N, Nadkarni P, Zhao H. VitaPad: visualization tools for the analysis of pathway data. Bioinformatics 2005;21(8):1596–1602. [PubMed: 15564306]

111. Ludemann A, Weicht D, Selbig J, Kopka J. PaVESy: pathway visualization and editing system. Bioinformatics 2004;20(16):2841–2844. [PubMed: 15105280]

112. Hu Z, Ng DM, Yamada T, et al. Visant 3.0: new modules for pathway visualization, editing, prediction and construction. Nucl Acids Res 2007;35(Web Server Issue):W625–W632. [PubMed: 17586824]
113. Hu Z, Mellor J, Wu J, Yamada T, Holloway D. DeLisi C: VisANT: data-integrating visual framework for biological networks and modules. Nucl Acids Res 2005;33(suppl 2):W352–W357. [PubMed: 15980487]
114. Dogrusoz U, Erson EZ, Giral E, et al. PATIKAweb: a web interface for analyzing biological pathways through advanced querying and visualization. Bioinformatics 2006;22(3):374–375. [PubMed: 16287939]
115. Funahashi A, Tanimura N, Morohashi M, Kitano H. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. BIOSILICO 2003;1:159–162.
116. Berger SI, Iyengar R, Ma'ayan A. AVIS: Ajax viewer of interactive signaling networks. Bioinformatics 2007;23(20):2803–2805. [PubMed: 17855420]
117. Meil A, Durand P, Wojcik J. PIMWalker: visualising protein interaction networks using the Hupo Psi molecular interaction format. Appl Bioinf 2005;4(2):137–139.
118. Orchard S, Hermjakob H, Taylor CF, et al. Further steps in standardisation report of the second annual proteomics standards initiative spring workshop. PROTEOMICS 2005;5(14):3552–3555. [PubMed: 16167370]
119. Hucka M, Finney A, Sauro HM, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 2003;19(4):524–531. [PubMed: 12611808]
120. Lloyd CM, Halstead MDB, Nielsen PF. CellML: its future, present and past. Prog Biophys Mol Biol 2004;85(2–3):433–450. [PubMed: 15142756]
121. Stromback L, Lambrix P. Representations of molecular pathways: an evaluation of SBML, Psi Mi and BioPAX. Bioinformatics 2005;21(24):4401–4407. [PubMed: 16234320]
122. W3c Extensible Markup Language (XML). http://www.w3.org/XML/
123. Keating SM, Bornstein BJ, Finney A, Hucka M. SBMLToolbox: an Sbml toolbox for Matlab users. Bioinformatics 2006;22(10):1275–1277. [PubMed: 16574696]
124. Schmidt H, Jirstrand M. Systems biology toolbox for MATLAB: a computational platform for research in systems biology. Bioinformatics 2006;22(4):514–515. [PubMed: 16317076]
125. Klukas C, Schreiber F. Dynamic exploration and editing of Kegg pathway diagrams. Bioinformatics 2007;23(3):344–350. [PubMed: 17142815]
126. Cary MP, Bader GD, Sander C. Pathway information for systems biology. Febs Lett 2005;579(8): 1815–1820. [PubMed: 15763557]
127. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25(1):25–29. [PubMed: 10802651]
128. Consortium TGO. Creating the gene ontology resource: design and implementation. Genome Res 2001;11(8):1425–1433. [PubMed: 11483584]
129. Gene Ontology C. The gene ontology (GO) database and informatics resource. Nucl Acids Res 2004;32(suppl 1):D258–D261. [PubMed: 14681407]
130. Smith B, Ceusters W, Klagges B, et al. Relations in biomedical ontologies. Genome Biol 2005;6(5)
131. Apweiler R, Bairoch A, Wu CH, et al. UniProt: the universal protein knowledgebase. Nucl Acids Res 2004;32(suppl 1):D115–D119. [PubMed: 14681372]
132. The Uniprot C. The universal protein resource (UniProt). Nucl Acids Res 2007;35(suppl 1):D193–D197. [PubMed: 17142230]
133. Sealfon R, Hibbs M, Huttenhower C, Myers C, Troyanskaya O. GOLEM: an interactive graph-based gene-ontology navigation and analysis tool. BMC Bioinf 2006;7(1):443.
134. Garcia O, Saveanu C, Cline M, et al. GOlorize: a Cytoscape plug-in for network visualization with gene ontology-based layout and coloring. Bioinformatics 2007;23(3):394–396. [PubMed: 17127678]
135. Liu H, Hu ZZ, Wu C. DynGO: a tool for visualizing and mining of gene ontology and its associations. BMC Bioinf 2005;6(1):201.

136. Lee J, Katari G, Sachidanandam R. GObar: a gene ontology based analysis and visualization tool for gene sets. BMC Bioinf 2005;6(1):189.
137. Ye J, Fang L, Zheng H, et al. WEGO: a web tool for plotting Go annotations. Nucl Acids Res 2006;34 (suppl 2):W293–W297. [PubMed: 16845012]
138. Zhong S, Storch KF, Lipan O, Kao MC, Weitz CJ, Wong WH. GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in gene ontology space. Appl Bioinf 2004;3(4):261–264.
139. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 2005;21(16):3448–3449. [PubMed: 15972284]
140. Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B. GOToolBox: functional analysis of gene datasets based on gene ontology. Genome Biol 2004;5(12):R101. [PubMed: 15575967]
141. Beissbarth T, Speed TP. GOstat: find statistically overrepresented gene ontologies within a group of genes. Bioinformatics 2004;20(9):1464–1465. [PubMed: 14962934]
142. Boyle EI, Weng S, Gollub J, et al. GO::TermFinder – open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. Bioinformatics 2004;20(18):3710–3715. [PubMed: 15297299]
143. Pasquier C, Girardot F, Jevardat De Fombelle K, Christen R. THEA: ontology-driven analysis of microarray data. Bioinformatics 2004;20(16):2636–2643. [PubMed: 15130932]
144. Busold CH, Winter S, Hauser N, et al. Integration of Go annotations in correspondence analysis: facilitating the interpretation of microarray data. Bioinformatics 2005;21(10):2424–2429. [PubMed: 15746280]
145. Hosack D, Dennis G, Sherman B, Lane H, Lempicki R. Identifying biological themes within lists of genes with EASE. Genome Biol 2003;4(10):R70. [PubMed: 14519205]
146. Nam D, Kim SB, Kim SK, Yang S, Kim SY, Chu IS. ADGO: analysis of differentially expressed gene sets using composite Go annotation. Bioinformatics 2006;22(18):2249–2253. [PubMed: 16837524]
147. Dennis G, Sherman B, Hosack D, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 2003;4(5):P3. [PubMed: 12734009]
148. Goffard N, Weiller G. PathExpress: a web-based tool to identify relevant pathways in gene expression data. Nucl Acids Res 2007;35:W176–W181. [PubMed: 17586825]
149. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005;21(18): 3674–3676. [PubMed: 16081474]
150. Lu LJ, Sboner A, Huang YJ, et al. Comparing classical pathways and modern networks: towards the development of an edge ontology. Trends Biochem Sci 2007;32(7):320–331. [PubMed: 17583513]
151. Racunas SA, Shah NH, Albert I, Fedoroff NV. HyBrow: a prototype system for computer-aided hypothesis evaluation. Bioinformatics 2004;20(suppl 1):i257–i264. [PubMed: 15262807]
152. Tran N, Baral C, Nagaraj VJ, Joshi L. Knowledge-based framework for hypothesis formation in biochemical networks. Bioinformatics 2005;21(suppl 2):ii213–ii219. [PubMed: 16204106]
153. Grady Booch, IJ.; Rumbaugh, J.; Jacobson, I. The unified modeling language user guide. Addison-Wesley; New York: 2005.
154. Webb K, White T. Uml as a cell and biochemistry modeling language. Biosystems 2005;80(3):283–302. [PubMed: 15888343]
155. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol 2007;3:88. [PubMed: 17353930]
156. Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein–protein interaction data. J Comput Biol 2003;10:947–960. [PubMed: 14980019]
157. Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 2003;19(suppl 1):i197–i204. [PubMed: 12855458]
158. Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T. Assessment of prediction accuracy of protein function from protein– protein interaction data. Yeast 2001;18(6):523–531. [PubMed: 11284008]

159. Chua HN, Sung WK, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. Bioinformatics 2006;22(13):1623–1630. [PubMed: 16632496]
160. Von Mering C, Krause R, Snel B, et al. Comparative assessment of large-scale data sets of protein – protein interactions. Nature 2002;417(6887):399–403. [PubMed: 12000970]
161. Chaurasia G, Herzel H, Wanker EE, Futschik ME. Systematic functional assessment of human protein – protein interaction maps. Genome Inf 2006;17(1):36–45.
162. Bandyopadhyay S, Sharan R, Ideker T. Systematic identification of functional orthologs based on protein network comparison. Genome Res 2006;16(3):428–435. [PubMed: 16510899]
163. Liang Z, Xu M, Teng M, Niu L. NetAlign: a web-based tool for comparison of protein interaction networks. Bioinformatics 2006;22(17):2175–2177. [PubMed: 16766562]
164. Liang Z, Xu M, Teng M, Niu L. Comparison of protein interaction networks reveals species conservation and divergence. BMC Bioinf 2006;7(1):457.
165. Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, Ideker T. PathBLAST: a tool for alignment of protein interaction networks. Nucl Acids Res 2004;32(suppl 2):W83–W88. [PubMed: 15215356]
166. Li Z, Zhang S, Wang Y, Zhang XS, Chen L. Alignment of molecular networks by integer quadratic programming. Bioinformatics 2007;23(13):1631–1639. [PubMed: 17468121]
167. Koyuturk M, Kim Y, Topkara U, Subramaniam S, Szpankowski W, Grama A. Pairwise alignment of protein interaction networks. J Comput Biol 2006;13(2):182–199. [PubMed: 16597234]
168. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet 2002;31(1):64–68. [PubMed: 11967538]
169. Milo R, Itzkovitz S, Kashtan N, et al. Superfamilies of evolved and designed networks. Science 2004;303(5663):1538–1542. [PubMed: 15001784]
170. Przulj N, Corneil DG, Jurisica I. Efficient estimation of graphlet frequency distributions in protein – protein interaction networks. Bioinformatics 2006;22(8):974–980. [PubMed: 16452112]
171. Lipshtat A, Purushothaman SP, Iyengar R, Ma'ayan A. Functions of bifans in context of multiple regulatory motifs in signaling networks. Biophys J 2008;94(7):2566–2579. [PubMed: 18178648]
172. Albert I, Albert R. Conserved network motifs allow protein – protein interaction prediction. Bioinformatics 2004;20(18):3346–3352. [PubMed: 15247093]
173. Yu H, Paccanaro A, Trifonov V, Gerstein M. Predicting interactions in protein networks by completing defective cliques. Bioinformatics 2006;22(7):823–829. [PubMed: 16455753]
174. Bu D, Zhao Y, Cai L, et al. Topological structure analysis of the protein –protein interaction network in budding yeast. Nucl Acids Res 2003;31(9):2443–2450. [PubMed: 12711690]
175. Girvan M, Newman ME. Community structure in social and biological networks. PNAS, USA 2002;99(12):7821–7826. [PubMed: 12060727]
176. Newman MEJ. From the cover: modularity and community structure in networks. PNAS 2006;103 (23):8577–8582. [PubMed: 16723398]
177. Rives AW, Galitski T. Modular organization of cellular networks. PNAS 2003;100(3):1128–1133. [PubMed: 12538875]
178. Chen J, Yuan B. Detecting functional modules in the yeast protein–protein interaction network. Bioinformatics 2006;22(18):2283–2290. [PubMed: 16837529]
179. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature 1999;402(6761 Suppl):C47–C52. [PubMed: 10591225]
180. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. PNAS 2003;100(21):12123–12128. [PubMed: 14517352]
181. Newman MEJ. Scientific collaboration networks. Part II. Shortest paths, weighted networks, and centrality. Phys Rev E 2001;64(1):016132.
182. Yoon J, Blumer A, Lee K. An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. Bioinformatics 2006;22(24):3106–3108. [PubMed: 17060356]
183. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. Defining and identifying communities in networks. PNAS USA 2004;101(9):2658–2663. [PubMed: 14981240]

184. Newman ME. A measure of betweenness centrality based on random walks. Social Networks 2005;27(1):39–54.

185. Freeman LC. A set of measures of centrality based on betweenness. Sociometry 1977;40:35–41.

186. Frey BJ, Dueck D. Clustering by passing messages between data points. Science 2007;315(5814): 972–976. [PubMed: 17218491]

187. Sen T, Kloczkowski A, Jernigan R. Functional clustering of yeast proteins from the protein–protein interaction network. BMC Bioinf 2006;7(1):355.

188. Flake GW, Lawrence S, Giles CL, Coetzee FM. Self-organization and identification of web communities. Computer 2002;35(3):66–70.

189. Andreopoulos B, An A, Wang X, Faloutsos M, Schroeder M. Clustering by common friends finds locally significant proteins mediating modules. Bioinformatics 2007;23(9):1124–1131. [PubMed: 17314122]

190. Lubovac Z, Gamalielsson J, Olsson B. Combining functional and topological properties to identify core modules in protein interaction networks. Proteins 2006;64(4):948–959. [PubMed: 16794996]

191. Luo F, Yang Y, Chen CF, Chang R, Zhou J, Scheuermann RH. Modular organization of protein interaction networks. Bioinformatics 2007;23(2):207–214. [PubMed: 17092991]

192. Bader G, Hogue C. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinf 2003;4(1):2.

193. Watts DJ, Strogatz SH. Collective dynamics of small-world networks. Nature 1998;393(6684):440–442. [PubMed: 9623998]

194. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucl Acids Res 2002;30(7):1575–1584. [PubMed: 11917018]

195. Brohee S, Van Helden J. Evaluation of clustering algorithms for protein– protein interaction networks. BMC Bioinf 2006;7(1):488.

196. Scott MS, Perkins T, Bunnell S, Pepin F, Thomas DY, Hallett M. Identifying regulatory subnetworks for a set of genes. Mol Cell Proteomics 2005;4(5):683–692. [PubMed: 15722371]

197. Handorf T, Ebenhoh O. MetaPath online: a web server implementation of the network expansion algorithm. Nucl Acids Res 2007;35(Web Server Issue):W613–W618. [PubMed: 17483511]

198. Lu LJ, Sboner A, Huang YJ, et al. Comparing classical pathways and modern networks: towards the development of an edge ontology. Trends Biochem Sci 2007;32(7):320–331. [PubMed: 17583513]

199. Lage K, Karlberg EO, Storling ZM, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotech 2007;25(3):309–316.

200. Asthana S, King OD, Gibbons FD, Roth FP. Predicting protein complex membership using probabilistic network reliability. Genome Res 2004;14(6):1170–1175. [PubMed: 15140827]

201. Page LB, Sergey Motwani R, Winograd T. The PageRank citation ranking: bringing order to the web. Stanford InfoLab Publication Server. 1998

202. Morrison J, Breitling R, Higham D, Gilbert D. GeneRank: using search engine technology for the analysis of microarray experiments. BMC Bioinf 2005;6(1):233.

203. Berger S, Posner J, Ma'ayan A. Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. BMC Bioinf 2007;8(1):372.

204. Dreyfus SE, Wagner RA. The Steiner problem in graphs. Networks 1972;1(195–207):111.

205. White, AG.; Ma'ayan, A. Connecting seed lists of mammalian proteins using Steiner trees. Acssc 2007. Conf. Record of the 41st Asilomar Conf. Signals, Systems and Computers; 4–7 November 2007; Pacific Grove, CA, USA. p. 155-159.

206. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. Nature 2007;447(7145):661–678. [PubMed: 17554300]

207. The International HapMap C. A haplotype map of the human genome. Nature 2005;437(7063): 1299–1320. [PubMed: 16255080]

208. The International HapMap Project. Nature 2003;426(6968):789–796. [PubMed: 14685227]

209. Frodsham AJ, Higgins JP. Online genetic databases informing human genome epidemiology. BMC Med Res Methodol 2007;7:31. [PubMed: 17610726]

210. Xu J, Li Y. Discovering disease-genes by topological features in human protein – protein interaction network. Bioinformatics 2006;22(22):2800–2805. [PubMed: 16954137]

211. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. PNAS 2007;104(21):8685–8690. [PubMed: 17502601]

212. Franke L, Bakel HV, Fokkens L, De Jong ED, Egmont-Petersen M, Wijmenga C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 2006;78(6):1011–1025. [PubMed: 16685651]

213. Ma'ayan A, Jenkins SL, Goldfarb J, Iyengar R. Network analysis of Fda approved drugs and their targets. Mount Sinai J Med A, J Transl Pers Med 2007;74(1):27–32.

214. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug target network. Nat Biotech 2007;25 (10):1119–1126.

215. Swedlow JR, Lewis SE, Goldberg IG. Modelling data across labs, genomes, space and time. Nat Cell Biol 2006;8(11):1190–1194. [PubMed: 17060903]

216. Bromberg KD, Ma'ayan A, Neves SR, Iyengar R. Design logic of a cannabinoid receptor signaling network that triggers neurite outgrowth. Science 2008;320(5878):903–909. [PubMed: 18487186]

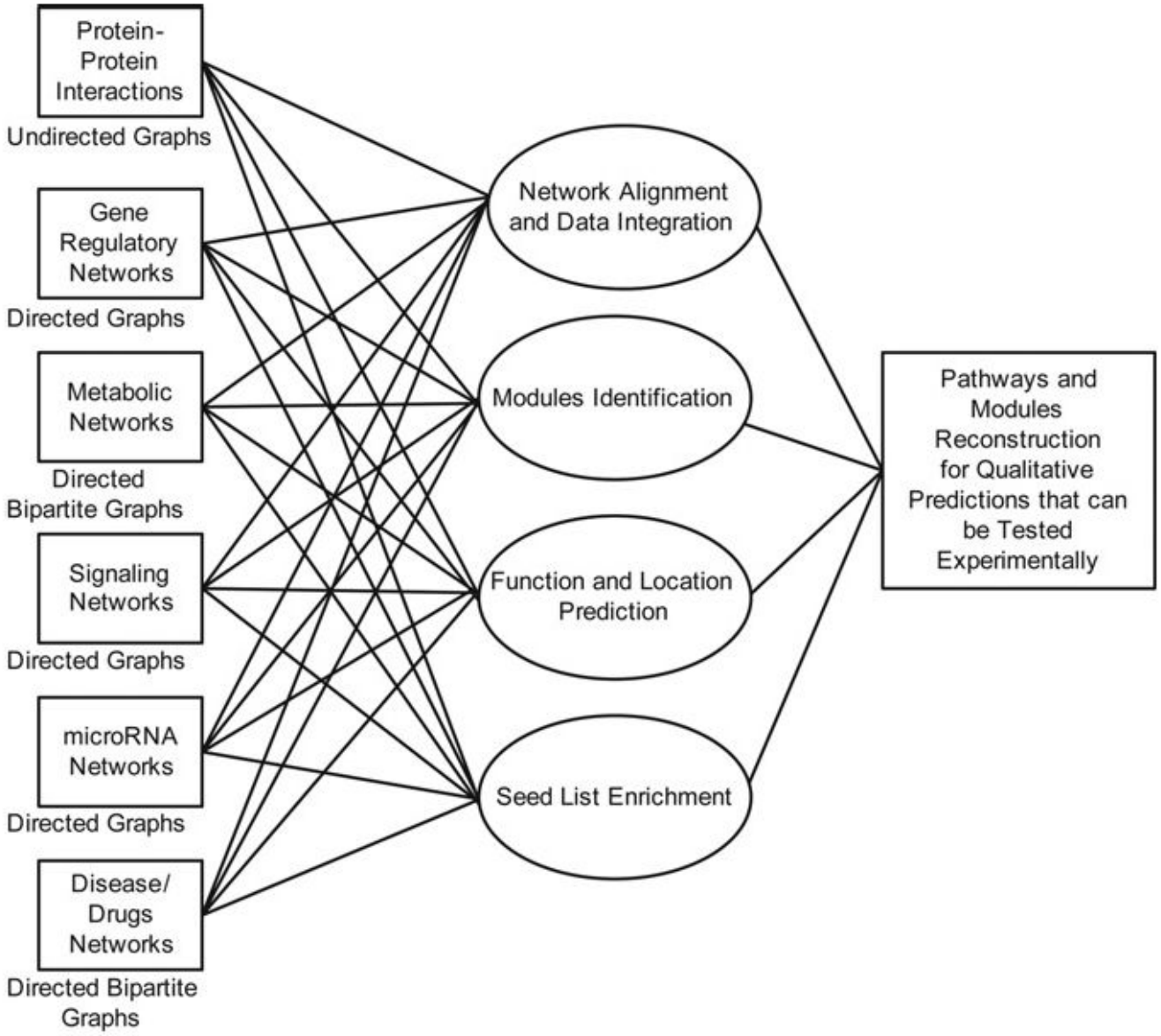

Figure 1. Different intracellular biological networks can be represented by different types of graphs

Since many of the nodes in those networks represent genes or proteins, such networks can be consolidated. Different networks can be aligned and compared with assess overlap and predict missing parts. Clustering methods can identify modules, and neighbourhood analysis can be used to predict function for genes with unassigned GO terms. Seed lists from experiments can be projected onto networks to identify modules that are altered under different experimental conditions. All these analyses can be used to guide experiments by providing means for integrative informative and predictive novel hypotheses

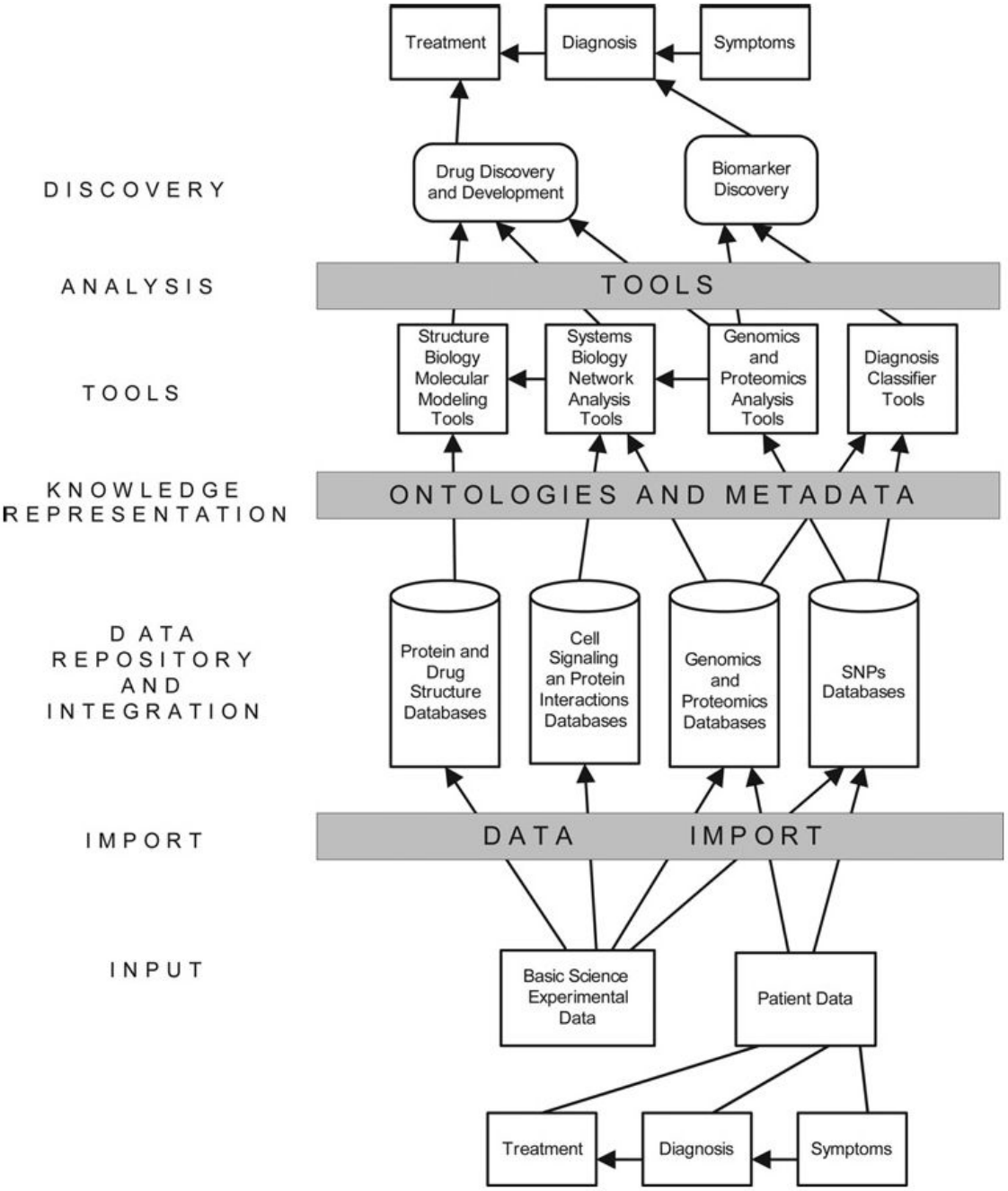

Figure 2. Data from basic scientific experiments and from patients are stored in different silos

To leverage discoveries, combining data from the different silos into knowledge can be accomplished by using ontologies. A layer above ontology development for interoperability is the development of tools that draw information from different silos as well as use output from other analysis tools. Such tools are expected to be used for producing advanced discoveries that would improve diagnosis and treatment