Posted on 20 Jul 2023 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.168984016.69483308/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.

A framework for considering prior information in network-based approaches to –omics data analysis

Julia Somers[1], Madeleine R. Fenner[1], Dharani Thirumalaisamy[1], Garth Kong[1], William Yashar[1], Meric Kinali[2], Kisan Thapa[2], Olga Nikolova[1], Özgün Babur[2], and Emek Demir[1]

[1]Oregon Health & Science University
[2]University of Massachusetts Boston

July 20, 2023

Abstract

For decades, molecular biologists have been uncovering the mechanics of biological systems. Efforts to bring their findings together have led to the development of multiple databases and information systems that capture and present pathway information in a computable network format. Concurrently, the advent of modern omics technologies has empowered researchers to systematically profile cellular processes across different modalities. Numerous algorithms, methodologies, and tools have been developed to use prior knowledge networks in the analysis of omics datasets. Interestingly, it has been repeatedly demonstrated that the source of prior knowledge can greatly impact the results of a given analysis. For these methods to be successful it is paramount that their selection of prior knowledge networks is amenable to the data type and the computational task they aim to accomplish. Here we present a five-level framework that broadly describes network models in terms of their scope, level of detail, and ability to inform causal predictions. To contextualize this framework, we review a handful of network-based omics analysis methods at each level, while also describing the computational tasks they aim to accomplish.

Introduction

Francois Jacob concluded his Nobel lecture in 1965 (awarded for modeling Lac Operon with Monod and Lwoff) with a vision: "We do not know how molecules find each other, recognize each other, and combine to constitute the regulatory network ... What is clear, however, is that the problems to be solved by cellular biology and genetics in the years to come tend increasingly to merge with those in which biochemistry and physical chemistry are involved.". The idea of a network model, where genes and gene products are linked by molecular processes, was present from the very first days of molecular biology. Due to the sheer complexity of biological systems, biologists have traditionally employed reductionist approaches where different fragments of cellular processes are isolated and identified. An implicit goal of this approach has been to assemble a network model in a piecemeal fashion from these reductionist findings, which will eventually be able to explain and predict the behavior of the biological system at large.

More than half a century later, tens of millions such reductionist findings have accumulated in the literature. Multiple databases and information systems have been developed to capture the pathway information accumulated in scientific literature and present it in computable format[1]. Millions of interactions, molecular processes and relationships are curated as networks, including metabolic pathways, signaling pathways, gene regulatory networks, molecular interaction networks, and genetic-interaction networks.

In parallel, our ability to systematically profile cellular processes has grown with the development of modern omics technologies. We now have a range of genomic, transcriptomic, metabolomic, and proteomic techniques at our disposal. We can deeply profile a cellular system in a given context, with an increasing ability to do so spatially and at the level of single cells. These technologies allow us to generate system-scale profiles

Posted on 20 Jul 2023 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.168984016.69483308/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.

without necessarily starting with a specific hypothesis or isolating a specific component—challenging the traditional piecemeal method. In most cases, the data-driven approaches no longer seek explicit biological grounding of their findings—clusters, subtypes and signatures replace mechanisms and pathways. The perceived incompatibility between "hypothesis driven, reductionist" and "data-driven, system-scale" camps led to one of the most polarizing epistemological debates in modern molecular biology[2–4].

Is this truly a fundamental divide—maybe we can have our cake and eat it too? To bridge this gap, we need to computationally combine these prior information fragments with -omics profiles to generate and test mechanistic, falsifiable conjectures at scale. Over the last two decades, thousands of algorithms and methods have been created in the field of network biology to address various sub-problems of this grand challenge, using diverse types of -omic data and prior knowledge. Given multiple data modalities, prior information sources, and tasks, it is often difficult to assess which algorithms are good for which biological questions and how they are related to each other. Here we present a framework to organize these methodologies into broad categories based on their use of prior information and the computational task they target. We also review a few examples from each category. Our goal in this review is to give readers a foundational understanding of the different types of networks, and a mental map to help match their needs with the available tools and algorithms.

Networks are models of biological systems

A biological model is an idealistic construct, *in simulacra* , that allows us to understand, explain and predict biological phenomena. A network, in the current context, is a graph model of a biological system which depicts molecular entities and the interactions between them. Mathematically, a network is a graph G(V,E), where a set of vertices or alternatively nodes (V) map to biological entities, connected by a set of ordered pairs or edges E, which represent the relationships between nodes. Network models vary greatly in their coverage of established biological knowledge, level of detail and interoperability with other networks. A network model can be as simple as the interaction "MDM2 binds to TP53" or can cover a system-level map that encompasses all known cellular processes. In some models, a single node may represent one entity, whereas others may have multiple nodes corresponding to the same entity but representing different states. Similarly, edges may have directionality that indicates the flow of cause and effect from reactant to product in an interaction or be undirected. They may also be signed, meaning they describe the nature of the reaction (ex. activation/inhibition), or unsigned. Some highly complex network models even account for stochiometric ratios and reaction dynamics equations in their construction.

Several domains of biology are modeled by networks, including: (i) M*etabolic pathways* (Figure 1A) are usually characterized by the abstraction of enzymes, substrates, and products. Typically, these reactions involve small molecules, and an enzyme, often a protein, catalyzes the reaction. Inhibitors and activators can also modulate the catalysis event. (ii) *Signaling pathways* (Figure 1B), on the other hand, encompass a range of biochemical reactions, including binding, transportation, and catalysis events involving molecules and complexes. These pathways may describe molecular states such as cellular location, covalent and non-covalent modifications, and sequence fragments. (iii) *Gene regulatory networks* (Figure 1C) involve transcription and translation events, along with their control mechanisms. (iv) *Molecular interactions* (Figure 1D) are typically represented as undirected graphs and cover non-covalent binding events. (v) *Genetic interactions* (Figure 1E) capture relationships between two genes when the observed phenotypic consequence of perturbing both genes is different from what is expected given the phenotypes of each single gene perturbation, such as in the case of epistasis.

Posted on 20 Jul 2023 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.168984016.69483308/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.

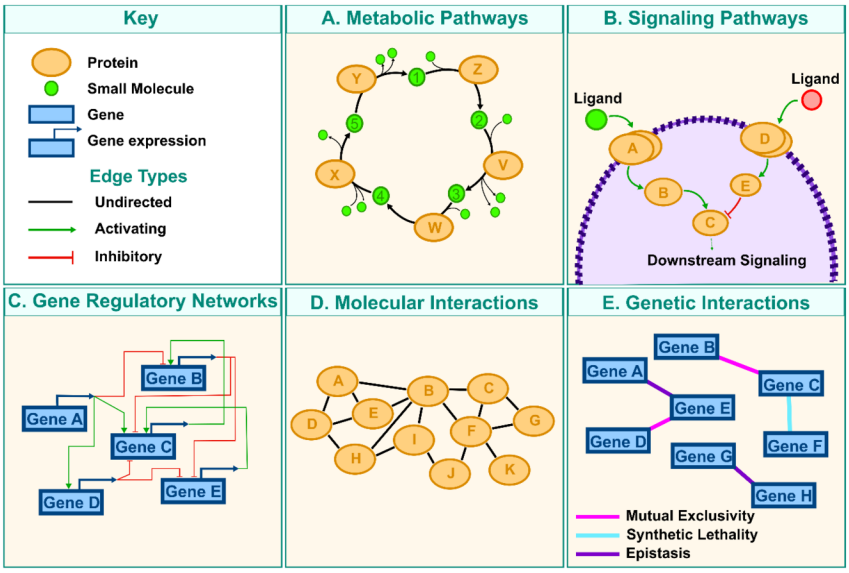

Figure 1. Domains of biological systems described by networksA. Metabolic pathways are typically described as a series of sequential reactions involving small molecules and catalytic enzymes. B. Signaling pathways describe the passage of signaling events, often triggered by a ligand/receptor combination. C. Gene regulatory networks are directed graphs that describe the circuitry of regulatory effects exerted from one gene to another. D. Molecular interaction networks are typically unsigned diagrams that represent uncharacterized interactions between a suite of molecules, often proteins. E. Genetic interactions describe the relationships between genes. Edges here describe the nature of the relationship, as opposed to the mechanisms that are involved in the relationship.

Creating Networks

Network representations of biological systems have been around for decades. Reconstruction of metabolic maps from early biochemical experiments started in the 1950s with Boehringer Mannheim charts. Modern reconstruction efforts like Reactome [5] , SIGNOR[6], KEGG[7], RECON [8], and Disease Maps [9] encompass hundreds of thousands of reactions, curated from scientific publications. Despite this herculean effort, these manually curated databases cannot keep up with the rate of scientific production given the available resources. To support manual curation efforts, multiple natural language processing (NLP) and crowd sourcing approaches to extract computable models from scientific literature have been developed[10], and recent language learning models (LLMs) offer great promise in expanding these efforts[11]. Additionally, in the case where there is very little existing literature about a system, networks can be inferred de novo or by expanding existing models[12]. This approach was particularly popular in the early phases of COVID-19 pandemic, wherein many researchers used network-based approaches grounded in SARS-CoV and MERS-CoV networks to extrapolate the molecular processes governing SARS-CoV-2 biology[13]. When PKNs are incomplete or nonexistent, interactions captured in the data can be used to infer a network structure. For example, in the case of high throughput PPI assays the identified interactions are commonly quantified based on confidence, then filtered using a cut-off score to lessen any noise introduced by the mode of collection. The filter chosen, which may be empirically or statistically informed, can have a significant impact on the rate of false positives and negatives in the resulting network[14]. Finally, some high-throughput modalities such as protein co-IP experiments can be readily expressed as networks without referring to curated sources of prior knowledge. Additional layers such as drug-target relationships can then be mapped to these interaction networks, as was done during COVID-19 to nominate targets for drug repurposing[15].

Networks and Context

Posted on 20 Jul 2023 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.168984016.69483308/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.

The fragments which make up a network often come from different biological contexts—here context is an umbrella term that implies different models, diseases, conditions, observation modalities and perturbations. For example, a group of researchers elucidates the phosphorylation event that drives a signaling cascade using an array of molecular techniques. Another research group identifies an inhibitor of this phosphorylation event. Another identifies a handful of transcription factors which assemble to produce this inhibitor, and so on until a pathway model starts to take shape. An important consideration in the implementation of this pathway is the context from which each of its components arose. If each of these groups were working with cell lines derived from different tissues, treated with different perturbing agents, or grown under different environmental conditions - could their results be stitched together into a common network? How to assemble these fragments properly, and when and which type of context restrictions should be used for a particular problem, are complicated questions with no clearcut answer. It is often necessary to join elements from different contexts to create networks that appropriately match the scope of the high throughput data. For datasets with a narrow scope within well-studied processes, such as a targeted metabolomics assay quantifying components of glucose metabolism, it may be possible to find a manually assembled quantitative model which combines fragments from a consistent context. However, for most -omics applications, which often involve untargeted high-throughput datasets, we often need to use the context-insensitive network and derive the context from the data.

Utility of Networks

Omics profiles offer a molecular snapshot of a biological system under a set of conditions[16]. Molecular structures commonly profiled by omics techniques include the genome (genomics), RNA (transcriptomics), proteins and their post translational modifications (proteomics), metabolites (metabolomics), and the epigenome (epigenomics)[17]. Some modalities can even be profiled at the level of a single cell, giving much deeper resolution. These technologies and the interpretation of their results with network-based methods are a driving force in the field of systems biology. Using networks, we can generate conjectures about the patterns in these highly complex datasets and understand which observed relationships can be explained by existing knowledge, and which relationships point to novel findings.

This *"explainability"* is key for the iterative process of scientific discovery. If an algorithm can make somewhat accurate predictions about the behavior of a system but cannot point to the components that are likely to drive the observed behavior, then the predictions can only be tested phenomenologically and not mechanistically. This is a limiting and inefficient way of analyzing a complex combinatorial system. There is no better example for this claim than commercial drug discovery, which relies on very large phenomenological screens for clinical trials. Despite substantial efficiency gains, between 1950 and 2010 the costs of research and development per approved drug approximately doubled every nine years[18] as we try to tackle increasingly complex diseases.

A related benefit of a grounded, mechanistic inference is the ability to "reason" about the system's response to a previously unknown perturbation such as a new drug combination or a mutation. This is an extension of a biologist's intuition - e.g., inhibiting the inhibitor of a target protein will activate it - but can be done at-scale. Additionally, it enables us to identify the reasons behind the failure of our predictions.

A perhaps less appreciated aspect of network-based approaches is the use of networks as prior information to restrict the search space of statistical algorithms. When evaluating potential network models in the context of their ability to explain or fit a certain data *de novo* , the number of possible network models grows exponentially,$O(2^{n^2})$, as a function of the number of nodes[19]. This leads to substantial problems with model overfitting, multiple hypothesis testing correction and model degeneracy. Multiple hybrid methods were developed that use prior information probabilistically along with *de novo* inference to center the inferred/evaluated models around known biology that can restrict the search space substantially[12].

The combined effect of these advantages is an incremental, iterative discovery process that can be done at-scale. This is crucial, given the rapid evolution of omics technologies and the ever-increasing volume of omics data.

Posted on 20 Jul 2023 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.168984016.69483308/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.

Section 2: A framework for categorizing and classifying network biology approaches.

When conducting network-based omics analysis, the choice of prior knowledge network can impact the results of the analysis[20]. Given the multitude of network databases available, it is useful to have a framework that can guide researchers to make informed decisions. Herein we define three 'tasks' which describe the overarching goal(s) of network-based approaches to omics data analysis. These tasks include network inference, explanation extraction, and phenotype prediction. Additionally, we define a framework for classifying network models into five levels of increasing level of detail: Gene Sets, Interaction Networks, Activity Flow, Process Description, and Quantitative Models (Figure 2). Finally, we review a sampling of network-based approaches at each of these five levels to contextualize the framework. Our classification of networks and the approaches that use them is intentionally broad to provide a high-level organization allowing for nuances in this rapidly evolving area of research.

Computational Tasks

Networks can be combined with -omics data to achieve a wide range of computational tasks. Below we define some broad categories that describe these computational tasks. These categories are not mutually exclusive, as many computational methods have the capacity to perform multiple tasks or hybrids of them. For example, methods which "upscale networks", meaning they output a higher-level network from a lower-level PKN, typically do both network inference and explanation extraction, as they select a small subset of the input PKN that can explain the correlations in the data and then will modify it to infer a new, higher-level network. It is also common to use explanation extraction or network inference task as a precursor to phenotype prediction, especially in clinical applications.

Explanation extraction aims to interpret patterns found within an omics profile and contextualize them using prior information about the system. It addresses hypotheses around system changes, such as differential expression or altered interaction strengths, to elucidate the mechanisms involved[21]. Common examples of explanation extraction tasks include enrichment-analysis and algorithms that produce a relevant subgraph of a larger network. Explanation extraction can also be thought of as emulating the literature search of a molecular biologist to explain the data at hand. As a molecular biologist reads the literature they ask "Is this information fragment compatible with my data? Does it explain it or contradict it? Is this applicable to my experiment's context?". The same questions are interrogated by explanation extraction methods, but in a quantitative manner that scales to high throughput data. Explanation extraction tools generate valuable conjectures that can, for example, guide the selection of subsequent perturbing agents, or recognize parallel mechanisms that unify multiple datasets in a novel way[12,19].

Network inference tasks produce a network model based on the input -omics data. This can be achieved by integrating prior networks or can be done *de novo* . Due to the combinatorial complexity of the model space and the inherent stochasticity of biological systems, inference is always an underdetermined problem and coherence of inferred networks and actual biological reality may be low, independent of the performance of the model. Constraining inference to at least partially conform with known biology can help by "anchoring" inferred networks. Another option is to use a large number of biological models in an ensemble learning strategy to reduce bias.

Some network inference approaches construct an entirely new model while others expand on established networks, in either case, the goal is to generate new mechanistic hypotheses. Upscaling algorithms are a common example of network inference. These approaches infer a higher-level representation (e.g. Activity Flow) from a lower-level prior network (e.g. protein-protein interactions) using -omics profiles. Upscaling can also be used to assign weights, direction, sign and rate constants to edges on a graph.

Phenotype prediction aims to predict how an organism or system responds to disease states and perturbations. These methods may be applied at a cellular level to project signaling events and transformations as well as broad phenomena like cell proliferation and survival, but they can also be extended to a network medicine approach, where predictions are made at a patient level to inform diagnosis, prognosis, or treatment response[22,23].

Posted on 20 Jul 2023 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.168984016.69483308/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.

Effective phenotype prediction is arguably more difficult than the prior two tasks. Phenotype is a function of the whole system that often contains feedback loops and other non-linear response circuitry. It is also inherently multimodal as at minimum, it requires one omic measurement and one phenotype measurement modality –e.g. IC50, GR50 or disease free survival. Each of these factors can be confounding to phenotype prediction tools.

Levels of Prior Knowledge Networks

There are many different ways of representing molecular processes in a graph model. The choice of representation is often dictated by the volume of experimental data informing different parts of the model. As models increase in their level of mechanistic detail, they also decrease in the scope of the biology that they are able to cover. For example, a level 2 PPI network with hundreds of proteins may be constructed from a single co-precipitation assay, while just one relationship in a level 4 or 5 model may synthesize the results of many separate experiments. The 5 levels are expressed visually in Figure 2 and are explained in further detail below.

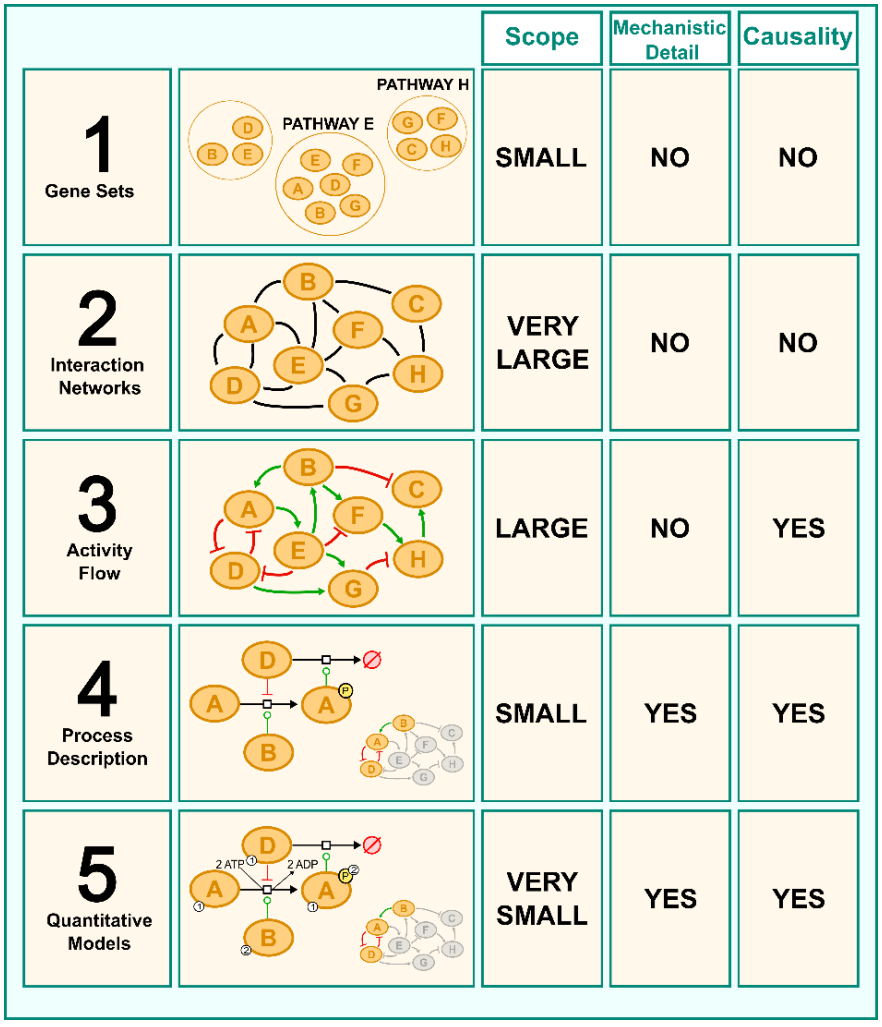

Figure 2. The 5 levels of network models. Scope refers generally to the size of networks and the volume of interactions recorded at that level. Mechanistic detail refers to whether the stepwise processes of a reaction are explicitly given in the network model. Causality refers to whether the network model can be used to

make causal inferences that can be statistically interrogated.

Gene Sets , as the name implies, are curated lists of genes grouped by association with a particular phenotypic outcome, molecular pathway, or cellular event. Gene sets, although not networks per se, are often derived from network representations, such as boundaries of KEGG pathways. Pathway boundaries are *fiat*boundaries[24], induced primarily through human demarcation. For example, despite covering the same biological processes, KEGG pathways contain 4 times more entities on average compared to BioCyc[25] pathways, primarily due to differences in curation guidelines. They also provide substantially different results when these fiat boundaries are used as input for gene set enrichment tasks [26]. Although they encompass well described biological mechanisms, gene sets do not contain mechanistic detail in the form of directed and/or signed edges. Approaches at this level typically perform explanation/extraction. Typically, this involves testing for statistical enrichment of gene sets or their components to propose explanations for observed cellular behavior, e.g. highlighting the most dramatically enriched pathway in a cancer biopsy to determine possible therapeutic targets. This can also be extended to a phenotype prediction task if the gene set describes a particular phenotype, e.g. a gene set composed of markers for epithelial to mesenchymal transition in breast cancer cells.

Interaction Networks represent interactions between biological entities by unsigned, undirected edges. These edges don't contain any cause/effect semantics and therefore can't be used to make causal predictions. These simple interactions can be detected in large quantities by through high-throughput methods, hence there are millions of interactions present in existing data sources, an order of magnitude more than subsequent levels. Additionally, interaction networks are simple to align and integrate with one another, as each entity is typically represented by only one node in the graph. They are commonly used as a starting point in untargeted high throughput assays where quantitative measurements are recorded for many entities and the researcher wants to look broadly at their data without necessarily seeking causal explanations.

Activity Flow networks, like interaction networks, typically contain one node for a given entity, allowing for easy integration of multiple networks so long as naming conventions for entities are consistent. In contrast, activity flow networks add a layer of cause/effect semantics in the form of directed and, sometimes, signed edges. For this reason, activity flow networks can be used for making causal predictions, and while these networks are considerably smaller than level 2, they are expansive enough that they can still be used for interrogating untargeted high-throughput datasets.

Process Description networks illustrate the mechanistic detail of how a reaction occurs. Because these models describe the stepwise events in a reaction, it is not uncommon that one edge could be informed by multiple sources, making them very well grounded in the literature. They are considerably smaller given that most, if not all, of their curation must be done by hand. Unlike prior levels, these diagrams represent the same entity with multiple nodes, corresponding to each of that entity's states through a sequence of events, including covalent modifications, cellular/subcellular locations, and/or complex memberships. This makes the integration of multiple process description networks a considerably more intensive exercise relative to levels 2 and 3.

Quantitative Models were originally derived from canonical chemical equations. These models are like process description networks in that these representations explicitly model the stepwise process of a reaction, but they are expanded to include quantitative factors like concentrations, stoichiometry, and rate constants. An example of a quantitative model would be a metabolic pathway represented as a bipartite graph of substrates, products, catalysts, and reactants. They are often used to describe systems which are very intensively studied and are typically very small compared to the preceding levels, due to the volume of research required to inform their curation.

Some networks and models fall into two consecutive categories. For example, the networks used in PhosphositePlus[27] and CausalPath[28] are represented as activity flow networks, however both describe post-translational modifications, which lends to the mechanistic detail in a process description network. Molecular Interaction Maps (MIMs)[29] are equivalent in semantic detail to process description but retain an activity

Posted on 20 Jul 2023 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.168984016.69483308/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.

Posted on 20 Jul 2023 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.168984016.69483308/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.

flow-like visualization. Finally, some large process description databases curate quantitative values such as enzymatic constants to allow for construction of quantitative models[30].

Classifying Methods within the Framework

To contextualize the above framework, we conducted a limited survey of algorithms and software tools which use networks as prior information in the analysis of omics data and categorized these methods based on the level of network they employ and the computational task(s) they accomplish. Given that hundreds of new algorithms and approaches are published every year, an exhaustive survey is not feasible for the present review. Methods are extremely diverse in their input, operations, and output, but in any case, the overarching goal of these approaches is to produce something that can be perceived and/or interpreted by a human user. We do not include cross-method comparison of features and performance. For each method we give a brief synopsis, discuss key aspects of the method, and finally summarize any real-world applications or validation in a biological system that the authors describe in their manuscript.

Level 1: Gene Sets

ReactomeGSA [31]

ReactomeGSA is an explanation extraction tool for comparative pathway-based gene set analysis. ReactomeGSA defines its gene sets from the pathways curated in the Reactome[5] database, then conducts a comparative gene set analysis at a pathway level to explain and biologically ground the differences between omics datasets, making it a quintessential explanation extraction tool with some phenotype prediction applications.

ReactomeGSA performs a differential expression analysis on a pathway scale for five quantitative omics data types, including microarray intensities, transcriptomics counts (raw or normalized), proteomics (spectral counts or intensity based quantitative data). ReactomeGSA is also capable of analyzing single-cell RNAseq (scRNAseq) datasets by calculating the mean expression for genes in a cluster and using this as 'pseudo-bulk' RNAseq to describe the cluster. For the analysis the user selects an appropriate methodology depending on their datatype and computational capacity. ReactomeGSA currently accommodates three gene set analysis methodologies, PADOG[32], Camera[33], and ssGSEA via GSVA[34]. The results of the analysis are mapped to the complete pathway browser database, where the user can view the pathway-level enrichment scores in the hierarchical 'tree-view' which also descending into individual pathways to view the differential gene expression values mapped to the corresponding genes in each pathway.

To demonstrate the clinical applications of ReactomeGSA the authors conducted a comparative pathway analysis of tumor induced plasmablast-like B-cell (TIPB) signaling across five TCGA cancer cohorts. These included melanoma, breast cancer, ovarian cancer, lung adenocarcinoma, and lung squamous cell carcinoma. The authors compared TIPB-high vs -low in each cohort, in addition to some cross cohort comparisons. They found that pathway-based gene sets describing B-cell receptor signaling and apoptosis were enriched for TIPB-high melanoma and ovarian cancer samples, which they later correlated with improved survival in these groups. When compared to melanoma, lung adenocarcinoma samples with high TIPB retained a unique signaling phenotype. These samples exhibited downregulation of the pathway-based gene sets describing B-cell receptor signaling, NF-kB signaling, p53 associated DNA damage repair, cell cycle, and apoptosis.

Level 2: Interaction Networks

SWAN [35]

SWAN incorporates prior knowledge network into the cutoff selection process for correlation networks. This is a hybrid inference/extraction algorithm that redefines the inference task as defining a cutoff threshold such that the agreement with prior information is maximized.

SWAN works by first constructing a correlation network from the data. The network is then filtered to remove edges that are not statistically significant. The remaining edges are then ranked according to their

Posted on 20 Jul 2023 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.168984016.69483308/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.

strength. Prior interaction networks can be easily integrated with inferred correlation networks. SWAN then selects a cutoff for the network based on prior knowledge. To calculate the correlation, SWAN uses shrinkage partial correlation based on the GeneNet algorithm – although this approach can be generalized to any correlation metric. The overlap is measured using Fisher's exact test p-value, which indicates the agreement between the calculated correlation network and prior knowledge. The optimal cutoff is defined as the point where the overlap is maximal.

SWAN was tested on pan-cancer data of 26 cancer types extracted from The Cancer Genome Atlas (TCGA). The network was able to identify enriched genes (OG) in the elevated pathways and suppressed genes (TSG) within suppressed pathways with a p-value < 0.05. This result was compared with the Gene Set Enrichment Analysis (GSEA) and revealed that SWAN outperformed GSEA. To check if SWAN can study race-specific CNA patterns, ovarian cancer samples from an African American population were collected, and non-Hispanic white patients were used as control. SWAN identified that the cytokine pathway was elevated in the former population which can be mapped to the overall poor prognosis in these patients. Furthermore, SWAN was also able to figure out the effect of the knockdown of metallothionein 2A which led to an increase in formation of uH2AX foci.

GLRP [36]

Graph Layer-wise Relevance Propagation (GLRP) is a novel method that extends the Layer-wise Relevance Propagation (LRP) technique to Graph Convolutional Neural Networks (Graph-CNN). LRP is an existing technique that explains the decisions made by deep learning models. The primary goal of GLRP is to explain the classification results of various omics data and molecular networks which could facilitate the decision-making processes in personalized medicine.

This is a unimodal, hybrid phenotype prediction and explanation/extraction algorithm that aims to ground predicted graphs to known protein-protein interaction networks. GLRP interprets the classification output by leveraging the molecular network and also produces patient-specific subnetworks that can be used to explain clinical outcomes and therapeutic vulnerabilities.

GLRP was trained on gene expression datasets of breast cancer and human umbilical vein endothelial cells (HUVECs). Their predictive performance was evaluated using the 10-fold cross-validation method. In the breast cancer study, GLRP was used to classify patients into metastatic and non-metastatic groups. The results were compared with the classification performance of random forest and glmgraph models as well as weighted gene co-expression network analysis. GLRP outperformed the other models, and the developed patient-specific subnetworks identified meaningful features in breast cancer samples.

Level 3: Activity Flow

CausalPath [28]

CausalPath is an explanation extraction algorithm which uses causal relationships from Pathway Commons[37] as priors to extract a mechanistic explanation for the patterns in proteomics, phospho-proteomics, and transcriptomics datasets. CausalPath produces causal hypotheses about the differences between comparable datasets, for example, biopsies from different conditions or timepoints, or the covariance across a cohort. These explanations are presented as an activity flow sub-network, which can also be expanded as a more detailed process description network. The method mimics a biologist's traditional approach of explaining changes in data using prior knowledge, but does this at the scale of hundreds of thousands of reactions.

CausalPath employs 12 pre-defined patterns that describe causal relationships between biological entities in the network, for example, a kinase phosphorylating another protein implies an expected correlation between the kinase's abundance or activating phosphorylation with the phosphorylation of the target protein). Using these pre-defined patterns, CausalPath assembles an activity flow network showing the causal relationships supported by the proteomic, phosphoproteomic and transcriptomic data.

CausalPath was applied to several publicly available datasets covering a wide range of scenarios and biological

Posted on 20 Jul 2023 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.168984016.69483308/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.

questions. In a set of time-resolved epidermal growth factor (EGF) stimulation experiments, CausalPath detected EGFR activation via downstream signaling of MAPKs, including feedback inhibition on EGFR. From ligand-induced and drug-inhibited cell-line experiments, CausalPath estimated the precision of its predictions. From CPTAC (Clinical Proteomic Tumor Analysis Consortium) protein mass spectrometry datasets for ovarian and breast cancer, CausalPath elucidated general and subtype-specific signaling, as well as regulators of well-known cancer proteins. In RPPA (Reverse Phase Protein Array) experimental datasets of 32 TCGA (Cancer Genome Atlas) cancer studies, CausalPath found a core signaling network that is recurrently identified across many cancer types.

CoPPNet [38]CoPPNet is a phenotype prediction tool which uses level 3 networks to accomplish unsupervised subtyping of cancer. CoPPNet first constructs a functional network of phosphorylation sites based on their co-phosphorylation patterns, and then identifies relevant subnetworks that correlate to subtypes.

The method first constructs a PhosphoSite Functional Association (PSFA) Network that models potential functional relationships between phosphosite pairs. Edges are inferred using information from existing databases: PTMCode is used for functional, structural and evolutionary associations, PhosphositePLUS for kinase-substrate associations and inferring shared-kinase pairs, and BIOGRID PPI for protein-protein interactions. Data from MS-based phospho-proteomics assays is then incorporated using bi-weight mi-correlation to assess co-phosphorylation (Co-P) of phosphosite pairs connected in the PSFA network, resulting in a weighted PSFA network. Finally, subnetworks enriched in highly co-phosphorylated phosphosite pairs are extracted. To achieve this, the weighted PSFA network is searched for subnetworks using a greedy algorithm to maximize Co-P score, resulting in a list of ranked subnetworks referred to as Co-P modules. Modules are then assessed for statistical significance, subtype specificity, predictive ability, and reproducibility.

CoPPNet was applied to two independent breast cancer phospho-proteomic datasets. The phosphorylation patterns of identified Co-P modules were found to strongly correlated with known subtypes (Luminal vs. Basal), and Co-P modules were shown to be reproducible across datasets from different studies.

IntOMICS [39]

IntOMICS is a Bayesian framework that reconstructs gene regulatory networks from integrated multi-omic data including; gene expression, DNA methylation, and copy number variation data as well as prior knowledge from KEGG (regulatory relationships) and target gene-transcription factor associations from ENCODE. This is a network inference algorithm for level 3 representation.

The IntOMICS framework is based on the Werhli and Husmeier (W&H) algorithm[40], which encodes each omics data source into separate energy functions. IntOMICS integrates the omics data by encoding the energy functions into a Gibbs distribution. Effects of multiple upstream controllers are additive. The inverse temperature hyperparameters for each source are tuned by sampling from the posterior distribution with Markov chain Monte Carlo (MCMC). Unlike the original W&H algorithm, IntOMICS uses an adaptive MCMC simulation and Markov blanked resampling to improve the MCMC convergence speed.

For validation and comparison, the authors used IntOMICS to understand the mechanism of chemoresistance using primary colon cancer samples from a randomized Phase III clinical trial. Their goal was to identify downstream mediators of *ABCG2* , which has been shown to contribute to chemoresistance. They compared the network generated from IntOMICS to those from an unaltered implementation of the W&H algorithm as well as two other multi-omic integration frameworks, RACER and KiMONo. IntOMICS nominated more downstream mediators of *ABCG2* , which may be important for chemoresistance in colon cancer and survival.

Level 4: Process Description

ScFEA/FLUXestimator [41]

Single cell flux estimation analysis (scFEA) is a prediction tool that infers metabolic flux from scRNAseq data using hand-curated metabolic pathways from KEGG as well as some hand curated mechanisms as prior

knowledge. In the web-application of scFEA, FLUXestimator, metabolic pathways from Recon3d are also available.

scFEA constructs a reduced network based on the prior network topology, genes with significant non-zero expression, and any preferred sub-network specifications from the user. This reduced network, termed a factor graph, is composed of metabolic modules (variables), representing groups of connected reactions, linked by intermediate metabolites (factors). For estimation, scFEA combines traditional flux-balance analysis with an optimization goal of minimizing influx/outflux imbalances while also incorporating enzyme transcript levels as a proxy for enzyme activity to further constrain the model search space.

scFEA was validated experimentally using matched scRNAseq and targeted metabolomics data collected from cells exposed to hypoxia and/or APEX1 knockdown. The authors observed that the predicted flux changes were consistent with the observed changes in the metabolomics data.

Fast-SL [42]

Fast-SL uses iterative search space reduction for rapid identification of synthetic-lethal gene sets up to an order of four. The overarching goal of this algorithm is to improve the computational efficiency and speed of synthetic lethality prediction from large metabolic networks. Because of its improved computational efficiency, Fast-SL is able to predict higher order synthetic lethal gene sets.

For the deletion of a gene/reaction to be considered lethal, the maximum growth calculated by flux balance analysis (FBA) must be smaller than the specified cutoff (v_{co}), typically 1% of the wild-type growth rate. The algorithm calculates the lethality cutoff v_{co} as 1% of the 'minimum norm', which corresponds to the maximum wild-type growth rate. Beginning with single lethal (first order) reactions, the search space is constrained to all reactions in the system with a nonzero flux in the distribution from the prior step. These reactions are denoted J_{nz}. Reactions in J_{nz} are then exhaustively tested for single-lethality by setting the flux of each individual reaction to zero, calculating the biomass flux, and comparing it to the cutoff, v_{co}. If the biomass flux is less than the cutoff, the reaction is considered lethal and added to the set of single lethal reactions (J_{sl}). Reactions in J_{sl} are then pruned from the search space for double lethal (second order) reactions (J_{db}). When calculating third order lethal reactions, the search space would be further reduced as reactions in J_{db} are removed from J_{nz}. The result is an iteratively pruned search space which becomes smaller with increasing order of lethal gene sets.

Using Fast-SL, the authors successfully identified lethal gene sets up to an order of four in *E. coli, S.Typhimurium, and M. tuberculosis* . They validated these results with an exhaustive search for first, second, and third order lethal gene sets. The authors reported an "exact match" between the number of lethal sets identified in the exhaustive search and those identified by Fast-SL. The authors also compared Fast-SL to another algorithm, SL Finder, which is also intended to reduce the computational intensity of identifying synthetic lethal gene sets. Fast-SL identified 127 novel triplets in *E. coli* which were not found by SL Finder. These novel triplets were predominantly involved in central carbon metabolism and amino acid synthesis.

Level 5: Quantitative Models

INTEGRATE [43]

INTEGRATE is a computational pipeline that integrates metabolomics and transcriptomics data to characterize multi-level metabolic regulation. The pipeline first computes differential reaction expression from transcriptomic data and uses constraint-based modeling to predict if the differential expression of metabolic enzymes directly originates differences in metabolic fluxes. In parallel, the pipeline uses metabolomics to predict how differences in substrate availability translate into differences in metabolic fluxes. It is an upscaling/inference algorithm.

This algorithm uses level 4 stoichiometries as constraints for flux balance analysis, RNA levels as enzyme abundance proxies to predict metabolomic fluxes then compare it with the observed data. The prior in-

Posted on 20 Jul 2023 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.168984016.69483308/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.

Posted on 20 Jul 2023 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.168984016.69483308/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.

formation is a further curated subset of RECON3D metabolomic construction called ERGO2. Once the metabolomic and transcriptomic data is mapped to the network intermediary scores are calculated for Feasible Flux Distributions (based on static analysis), Reaction Activity Scores (based on RNA levels) and Reaction Propensity Score (based on substrate levels). Agreement between these metrics, calculated by Variation Concordance Analysis is the final output and can be used for both explanation/extraction and upscaling.

The pipeline was applied to a set of immortalized normal and cancer breast cell lines. The results showed that the pipeline was able to identify metabolic reactions that are regulated at both the metabolic and gene expression levels. The pipeline was also able to identify metabolic reactions that are differentially regulated in cancer cells compared to normal cells.

SUMMER [44]

SUMMER (Shiny Utility for Metabolomics and Multiomics Exploratory Research) uses reaction rate potentials to perform pathway enrichment analysis on metabolomic data. SUMMER uses level 4 metabolomic networks from the KEGG database[7]. This is a network upscaling method from level 4 to level 5 as a first step of quantitative modeling.

SUMMER uses reaction rate potentials to model the feedback effects between an enzyme, its substrate(s), and its product(s). It also infers the catalytic activity of each enzyme using integrated transcriptomics or proteomics data. The method then uses this integrated model to understand the change in reaction rate potentials between a perturbed condition and a reference condition. The resulting ratio of the resulting reaction rate potentials between a perturbed condition and a reference condition is then bootstrapped to calculate a ranking score between each reaction. Using the rank scores, SUMMER identifies the "hotspot" reactions in the network.

The authors applied SUMMER to re-analyze a metabolomic and transcriptomic dataset generated from a mouse model of accelerated aging and dementia. They wanted to understand the pathways that were altered by a neuroprotective compound. They found that treatment with this compound was associated with an increase in acetyl-CoA activity and an enrichment of TCA cycle activity.

Discussion

Pathway or network analysis is often viewed as a one-size-fits-all approach that can be applied universally to any dataset. However, as our review demonstrates, network analysis encompasses a broad range of approaches with unique data requirements and diverse PKN sources. Any new project or program incorporating network analysis should carefully define the task at hand, explore the available prior information sources, and consider the integration and scalability challenges associated with each resource.

Networks and network-based methods are invaluable tools for the analysis of omics data. It is widely recognized that the selection of prior knowledge network (PKN) can influence the outcome of analysis, therefore selection of an appropriate PKN is key to producing reliable results. With such an enormous suite of network resources available it can become overwhelming to select an appropriate model. To address this challenge, we present a framework for classifying PKNs and network-based methods. This framework characterizes PKNs in terms of their scope, mechanistic detail, and ability to inform causal predictions. We also outline some common computational tasks to describe the aim of network-based analyses. To contextualize the framework, we sampled a handful of published network-based methods and discussed their PKN selection, the tasks they aim to accomplish, their approach to analysis and their real-world applications. While this sampling is not exhaustive, it offers readers a practical glimpse into the application of the framework.

Looking ahead, we anticipate network analysis to gain even greater prominence, shifting towards more detailed approaches for two reasons. First, the rapid advancements in multi-modal, spatial, and single-cell modalities have enabled the measurement of subcellular protein localization changes, post-translational modifications (PTMs), and molecular complexes at a single-cell scale using imaging modalities[45].This wealth of

information primarily resides in level 4 networks and, to a lesser extent, in level 3 networks. Effectively harnessing these rich datasets will necessitate the utilization of more detailed PKNs. Second, recent breakthroughs in large language models[46] have significantly enhanced our ability to extract knowledge from the literature. Combining this capability with crowd-sourcing[47] and human-in-the-loop systems[48] holds the potential to reduce curation costs by two orders of magnitude[47] enabling near-complete curation of the entire biomedical literature on biological molecular processes. The increased completeness of PKNs, along with improved and larger datasets, will unlock extensive application areas for increasingly sophisticated network models.

References

1. Bader, G. D., Cary, M. P. & Sander, C. Pathguide: a Pathway Resource List. *Nucleic Acids Res.* **34** , D504–D506 (2006).

2. Sorger, P. K. A reductionist's systems biology: opinion. *Curr. Opin. Cell Biol.* **17** , 9–11 (2005).

3. Weinberg, R. Point: Hypotheses first. *Nature* **464** , 678–678 (2010).

4. Golub, T. Counterpoint: Data first. *Nature* **464** , 679–679 (2010).

5. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase.*Nucleic Acids Res.* **46** , D649–D655 (2018).

6. Lo Surdo, P. *et al.* SIGNOR 3.0, the SIGnaling network open resource 3.0: 2022 update. *Nucleic Acids Res.* **51** , D631–D637 (2023).

7. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs.*Nucleic Acids Res.* **45** , D353–D361 (2017).

8. Swainston, N. *et al.* Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics* **12** , 109 (2016).

9. Ostaszewski, M. *et al.* COVID19 Disease Map, a computational knowledge repository of virus–host interaction mechanisms. *Mol. Syst. Biol.* **17** , e10387 (2021).

10. Valenzuela-Escárcega, M. A. *et al.* Large-scale automated machine reading discovers new cancer-driving mechanisms. *Database***2018** , bay098 (2018).

11. Conceição, S. I. R. & Couto, F. M. Text Mining for Building Biomedical Networks Using Cancer as a Case Study. *Biomolecules***11** , 1430 (2021).

12. Korkut, A. *et al.* Perturbation biology nominates upstream–downstream drug combinations in RAF inhibitor resistant melanoma cells. *eLife* **4** , e04640 (2015).

13. Messina, F. *et al.* COVID-19: viral–host interactome analyzed by network based-approach model to study pathogenesis of SARS-CoV-2 infection. *J. Transl. Med.* **18** , 233 (2020).

14. Meysman, P. *et al.* Protein complex analysis: From raw protein lists to protein interaction networks. *Mass Spectrom. Rev.***36** , 600–614 (2017).

15. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583** , 459–468 (2020).

16. Yamada, R., Okada, D., Wang, J., Basak, T. & Koyama, S. Interpretation of omics data analyses. *J. Hum. Genet.***66** , 93–102 (2021).

17. Tolani, P., Gupta, S., Yadav, K., Aggarwal, S. & Yadav, A. K. Chapter Four - Big data, integrative omics and network biology. in*Advances in Protein Chemistry and Structural Biology* (eds. Donev, R. & Karabencheva-Christova, T.) vol. 127 127–160 (Academic Press, 2021).

18. Scannell, J. W. & Bosley, J. When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis. *PLOS ONE***11** , e0147215 (2016).

Posted on 20 Jul 2023 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.168984016.69483308/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.

Posted on 20 Jul 2023 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.168984016.69483308/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.

19. Muldoon, J. J., Yu, J. S., Fassia, M.-K. & Bagheri, N. Network inference performance complexity: a consequence of topological, experimental and algorithmic determinants. *Bioinformatics***35** , 3421–3432 (2019).

20. Mubeen, S. *et al.* The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. *Front. Genet.* **10** , (2019).

21. Garrido-Rodriguez, M., Zirngibl, K., Ivanova, O., Lobentanzer, S. & Saez-Rodriguez, J. Integrating knowledge and omics to decipher mechanisms via large-scale models of signaling networks. *Mol. Syst. Biol.* **18** , e11036 (2022).

22. Silverman, E. K. *et al.* Molecular Networks in Network Medicine: Development and Applications. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **12** , e1489 (2020).

23. Ranea, J. A. G., Perkins, J., Chagoyen, M., Díaz-Santiago, E. & Pazos, F. Network-Based Methods for Approaching Human Pathologies from a Phenotypic Point of View. *Genes* **13** , 1081 (2022).

24. Smith, B. & Varzi, A. C. Fiat and Bona Fide Boundaries.*Philos. Phenomenol. Res.* **60** , 401–420 (2000).

25. Karp, P. D. *et al.* The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinform.* **20** , 1085–1093 (2019).

26. Altman, T., Travers, M., Kothari, A., Caspi, R. & Karp, P. D. A systematic comparison of the MetaCyc and KEGG pathway databases.*BMC Bioinformatics* **14** , 112 (2013).

27. Hornbeck, P. V. *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse.*Nucleic Acids Res.* **40** , D261-270 (2012).

28. Babur, Ö. *et al.* Causal interactions from proteomic profiles: Molecular data meet pathway knowledge. *Patterns* **2** , 100257 (2021).

29. Kohn, K. W., Aladjem, M. I., Kim, S., Weinstein, J. N. & Pommier, Y. Depicting combinatorial complexity with the molecular interaction map notation. *Mol. Syst. Biol.* **2** , 51 (2006).

30. Chang, A. *et al.* BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res.***49** , D498–D508 (2021).

31. Griss, J. *et al.* ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis. *Mol. Cell. Proteomics* **19** , 2115–2125 (2020).

32. Tarca, A. L., Draghici, S., Bhatti, G. & Romero, R. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics***13** , 136 (2012).

33. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43** , e47 (2015).

34. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics***14** , 7 (2013).

35. Bowers, R. R. *et al.* SWAN pathway-network identification of common aneuploidy-based oncogenic drivers. *Nucleic Acids Res.***50** , 3673–3692 (2022).

36. Chereda, H. *et al.* Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. *Genome Med.* **13** , 42 (2021).

37. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39** , D685–D690 (2011).

38. Ayati, M., Chance, M. R. & Koyutürk, M. Co-phosphorylation networks reveal subtype-specific signaling modules in breast cancer.*Bioinforma. Oxf. Engl.* **37** , 221–228 (2021).

39. Pačínková, A. & Popovici, V. Using empirical biological knowledge to infer regulatory networks from multi-omics data. *BMC Bioinformatics* **23** , 351 (2022).

40. Werhli, A. V. & Husmeier, D. Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat. Appl. Genet. Mol. Biol.* **6** , Article15 (2007).

41. Alghamdi, N. *et al.* A graph neural network model to estimate cell-wise metabolic flux using single-cell RNA-seq data. *Genome Res.* **31** , 1867–1884 (2021).

42. Pratapa, A., Balachandran, S. & Raman, K. Fast-SL: an efficient algorithm to identify synthetic lethal sets in metabolic networks.*Bioinformatics* **31** , 3299–3305 (2015).

43. Di Filippo, M. *et al.* INTEGRATE: Model-based multi-omics data integration to characterize multi-level metabolic regulation. *PLOS Comput. Biol.* **18** , e1009337 (2022).

44. Huang, L., Currais, A. & Shokhirev, M. N. SUMMER, a shiny utility for metabolomics and multiomics exploratory research. *Metabolomics Off. J. Metabolomic Soc.* **16** , 126 (2020).

45. Rozenblatt-Rosen, O. *et al.* The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. *Cell* **181** , 236–249 (2020).

46. Luo, R. *et al.* BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.***23** , bbac409 (2022).

47. Wong, J. V. *et al.* Author-sourced capture of pathway knowledge in computable form using Biofactoid. *eLife* **10** , e68292.

48. Todorov, P. V., Gyori, B. M., Bachman, J. A. & Sorger, P. K. INDRA-IPM: interactive pathway modeling using natural language with automated assembly. *Bioinforma. Oxf. Engl.* **35** , 4501–4503 (2019).

Posted on 20 Jul 2023 | The copyright holder is the author/funder. All rights reserved. No reuse without permission. | https://doi.org/10.22541/au.168984016.69483308/v1 | This a preprint and has not been peer reviewed. Data may be preliminary.