A Patient-Centric Recommendation Engine for Longitudinal Clinical Trials

"Moving from data-centric to patient-centric paradigm for enabling P4 medicine"

Kenny Yu[1] and Kasthuri Kannan[2]

[1]*Department of Neurosurgery, Memorial Sloan-Kettering Cancer Center*
[2]*Departments of Translational Molecular Pathology, Neurosurgery, UT MD Anderson Cancer Center*

Introduction

Limitations of traditional relational databases for longitudinal patient/patient data tracking

Making sense of large-scale, multi-dimensional datasets to gain pathophysiological insight remains a significant analysis hurdle, and the bar for obtaining actionable information is even higher still. The aggregation, integration, and analysis of datasets collected in the context of a longitudinal study such as the one currently being proposed will require some thought in terms of what computational resources are required, and how data should be organized to accommodate the widest possible range of analyses.

As a simple case of analysis, we can consider associating tumor size between control and experimental drug settings with copy number and methylation changes in specific genes. In traditional analysis settings, these correlative pieces of evidence are made based on prior knowledge, intuition, or computational algorithms. At the same time, the data is organized in a tabular manner in a relational database **without** explicit connections. In the above example, the tumor size, copy number, and methylation data would typically reside in separate tables (see Figure 1) in a database while the correlation is deduced based on some knowledge base, and statistics. This approach typically prioritizes genetics over epigenetics or vice-versa, implicitly or explicitly, without an unbiased approach to generating and evaluating hypotheses regardless of the philosophical position to which one subscribes.

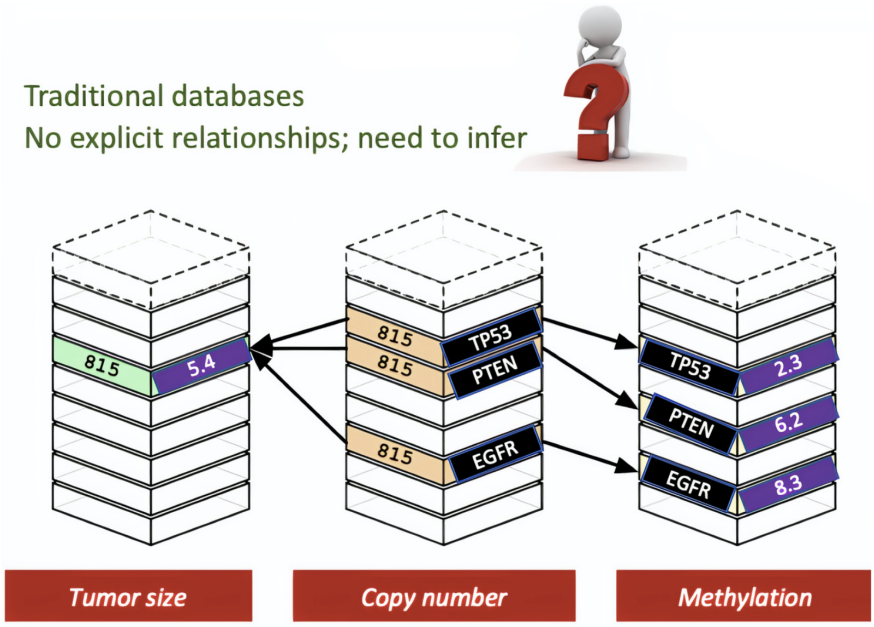

Figure 1: An example of relational database schema containing clinical and genomic data tables

Further, scalability and data integration across multiple data sources is a significant challenge even when constructing static networks from these datasets. Manual examination of these networks (such as gene-gene signaling) to determine the pathways would still result in hypothesis confirmation bias. Moreover, in constructing these static networks, the mechanism to store and query multiple relationships between the data is absent and has to be determined apriori, resulting in an inability to generate and evaluate hypotheses dynamically. Most importantly, biological insights using such static networks are derived by abstracting and isolating molecular data from patients resulting in

interpretations that are data-centric than patient-centric and only enable curiosity-driven than patient-driven medicine. For example, in performing gene set enrichment analysis on patient-derived gene expression data, the molecular signatures and pathways analysis are performed only based on the genes and their expression without including the patients who had those gene expressions altered. By isolating genes from the patients, we lose the contextual information that would otherwise enable personalized medicine.

Also, from a longitudinal clinical trials perspective, a computational framework to temporally track patient and samples outcomes is necessary. These outcomes would constitute any patient response or data interpretation involving clinical/molecular information, radiology/pathology images, sample processing, and quality control information. While tracking patients and their outcomes in a traditional relational database would seem straightforward, such a database design would magnify the above-highlighted shortcomings from a temporal analysis standpoint. Moreover, manually associating data from two points in time for any two patients would implicitly assume the unit of time for tumor progression is the same for those patients, which is an unsupported hypothesis. Furthermore, almost all such temporal analysis thus far forces a linear time model into patient data, an untenable assumption and a significant obstruction to realizing personalized, predictive, preventive, and participatory medicine (P4 medicine). Hence, we need to use non-linear approaches to track patients/samples, which overcomes these limitations. **Graph databases** are non-linear, and relationship (and context) preserving technologies are crucial to developing such a platform.

Graph databases and associated foundational principles

A systems biology paradigm for medicine

Graphs databases are network databases with nodes and edges. Unlike traditional relational databases, in graph databases, relationships matter as much as the data themselves. This feature provides a systems approach to medicine. The underlying principle behind systems philosophy is that "the whole is not equal to the sum of its components," and interactions between the components define the system as much as the components themselves. By explicitly defining the data corresponding to the components and their interactions as relationships, we can construct graph databases that would enable formal frameworks for systems biology in medicine.

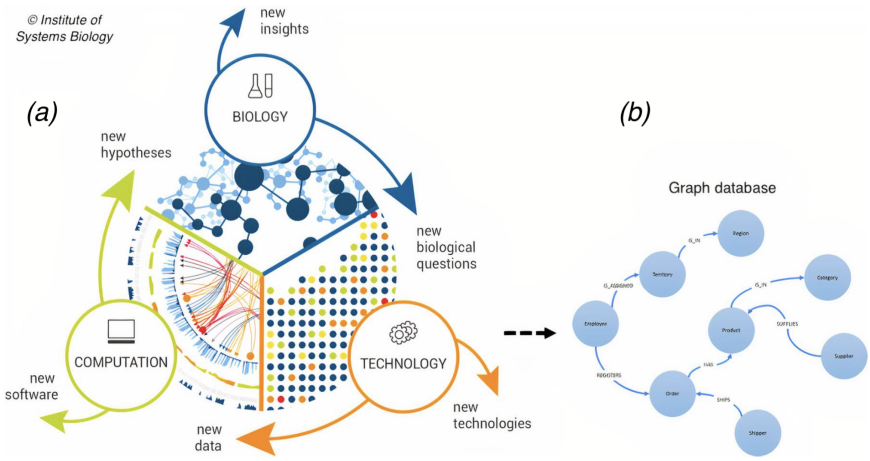

Figure 2: Graph database as a key technology component in systems biology paradigm

Figure 2 (a), highlights one of the general principles of systems biology[1], where multiple high-throughput data from various technologies are integrated to produce a cycle of new hypotheses, insights, and biological questions. However, the underlying integrating framework is usually assumed as an abstract (static) network with defined relationships than as a graph database. Without the graph database as a key technology enabler, as shown in Figure 2, static network integration schemes would still suffer from hypothesis confirmation bias. Hence, modeling the clinical trials and patient data through a graph database is fundamental to

P4 medicine and provides a systems biology framework for patient-driven science and treatment modalities.

From the data perspective, any data comes with a context/relationship. For example, a mutation in a gene is contextual on the patients who undergo that mutation in that gene. While static network analysis could result in an understanding derived from a set of mutations (or genes) abstracted from patients, graph databases would enable including patient information in deriving directed insights. Thus, graph databases provide a compelling framework for tracking and querying patients and their samples by simultaneously storing and modeling data and relationships. We can query the graph database to determine causal pathways using probabilistic graph models such as Bayesian belief networks, Markov random fields, and other statistical probability models. One significant advantage to such a system (as opposed to static networks) is the mutability of the network facilitating hypothesis generation using database queries on the node and edge properties. Although there have been efforts to model patient outcomes through graph-based frameworks, longitudinally integrating patients and their tissues data (imaging or biopsied) has not been attempted.

Relevance to complex adaptive systems: self-organization, emergence, and systems principle

Complex Adaptive Systems (CAS) are dynamic, non-linear systems connected through many connected parts that adapt, constantly refine, and adjust to their environment. A classic example is the swarming behavior of birds, especially starlings, known as "murmurations." Unlike a complicated system such as a manufacturing assembly line where the system's performance is fixed and not capable of autonomous evolution, the hallmarks of CAS are evolvability and emergence. New properties arise from interactions of simpler units resulting in self-organizing behavior, e.g., fractals. Also, new and unexpected interactions between the simpler units can shift the system to a new state with very different properties. This phenomenon is known as emergence. One of the foundational aspects of CAS is that we cannot reliably deduce this self-organizing behavior or emergence in the whole system from knowledge of the properties of the simpler isolated units ("the whole is not the sum of its parts"). Therefore, systems principle is fundamental to CAS that determines the extent to which self-organization and emergence arise in the system.

Life systems and especially cancer are classic examples of CAS. Cancer is a complex ecosystem of tumor and host dynamics such as host immune response, microbiome, genetic and epigenetic modifications in tissue microenvironment that dynamically organizes itself over time, adapting the mechanisms of rapid cell proliferation. For example, in histopathology, there is documentation[2] of a self-organizing principle known as "histostatis," a cell-autonomous, self-organization property of tumor cells that promotes the generation of the characteristic histomorphology. As for the property of emergence that produces new and unexpected patterns, there is a well-documented phenomenon[3] where tens to thousands of chromosomal rearrangements, known as "chromothripsis" occur simultaneously while the integrity of the cell is still maintained. While it has been speculated that chromothripsis as a single devastating onslaught could be the upper limit of what a cell could tolerate[12], this event as a classic example of emergent phenomena is yet to be reported. Therefore, from an empirical standpoint, the cancer ecosystem maintains the phenomena of self-organization and emergence. However, it is difficult to determine the extent to which these phenomena operate because the systems principle exists only as a philosophical basis in CAS without any proper framework. A solid foundation for systems principle in CAS would allow us to understand self-organization and emergent phenomena in cancer systems. Graph databases, as dynamic and mutable data structures, and querying platform fill this void in transforming the systems principle of CAS into a rigorous science where we can systematically study the extent to which these phenomena drives cancer, as the disease longitudinally evolves in patients. Although complex systems such as

the power systems and human immune system have been modeled using the graph architecture[10, 11], graph database as a central CAS engine for tracking disease progression and in particular, cancer, is yet to be proposed.

Network motifs and artificial intelligence

Network motifs are patterns of subgraphs in a larger graph that are over-represented or confer critical functionality to the system under consideration. They are the simple building blocks of the larger network[4]. They have been subject to an extensive study in electrical circuits, ecology, and social networks. In molecular biology, several motifs have been identified[5-7] as typical patterns in different biological networks such as feed-forward loops, bifan network motifs, autoregulation, cascades, and positive/negative feedback loops. While we can identify several of these motifs using static networks, labeled property graph databases would enable determining such motifs dynamically. This dynamic determination is possible because we can use the database query language to combine relationship and node properties and identify patterns through various combinations. An open-source tool *DotMotif*[8] is an example of the case in point in the context of Connectomics, the study of the brain's structural and functional connections between cells.

Graph databases enable the direct application of artificial intelligence (AI) algorithms through Graph Neural Networks (GNNs). GNNs allow learning on graph structures by "embedding" node or edge properties in a high-dimensional space which is then passed into a neural architecture for classification or clustering tasks. Usually, these embeddings are determined using message passing approaches. Using queries on a graph database, we can permute the node or edge properties to enable different message passes to determine optimal embeddings for a given learning task. These different message passes obtained through queries will enable us to obtain AI-driven directed insights from the system. Therefore, a graph database platform will help apply AI algorithms on highly unstructured data through GNNs even though a typical neural architecture such as deep learning operates on structured datasets.

A graph schema for the longitudinal patient and patient data tracking

An example recommendation engine for longitudinal clinical trials

Figure 3 highlights a graph schema for a recommendation engine for longitudinal clinical trials. The schema is designed based on "events" using a property graph model. A node and event are synonymous in our model. An event can have a label of sample collection, clinical, immunological, imaging, or molecular features. Moreover, two events are connected if there are patients who undergo those events. Having patients as edges that connect the events provides a non-linear approach to tracking patients and facilitates data integration across various data sources and favors scalability.

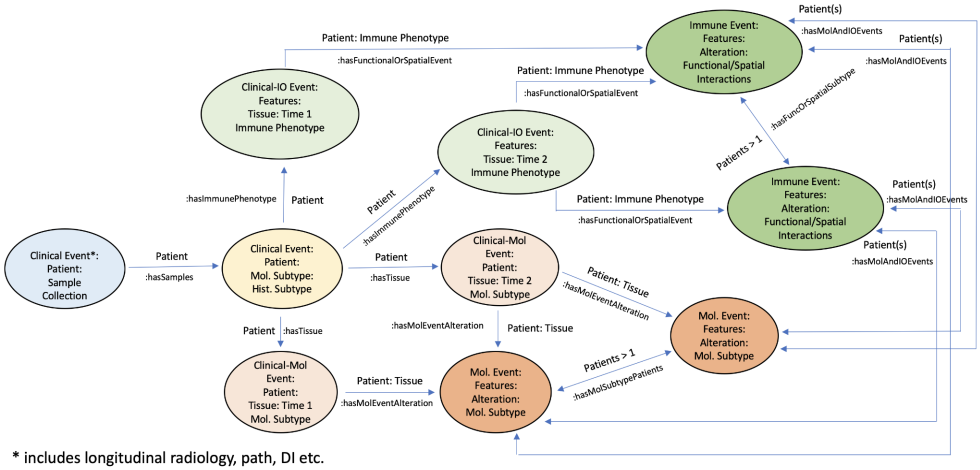

Figure 3: *A patient-centric graph database schema for longitudinal clinical trials with events as nodes and patients as edges*

Further, our event--patients--event database provides a higher level of generalizability for hypothesis generation. For example, relating two events by a patient phenotype, say, "HAS_GBM_G-CIMP," would restrict the hypothesis generation to G-CIMP patients. By associating the events as patients, we can generate hypotheses on population levels rather than disease levels. For instance, we could ask and answer the question "Is PI3K-pathway implicated in human GBMs?" regardless of genetic or epigenetic data sources or a specific patient population, which we cannot, if we restrict the relationship to "HAS_GBM_G-CIMP." Thus, this model provides a novel and compelling paradigm for data integration at various scales ranging from molecular events in a single patient's tumor to integrating imaging, clinical and molecular profile (or genotype) with phenotypic characteristics at the population level longitudinally.

Another advantage in using a graph database over a relational database is in applying graph algorithms. Graph algorithms such as pathfinding, centrality, community detection, and link/node embeddings enable powerful approaches to analyzing connected data because their mathematics is built explicitly on relationships than the data themselves. In particular, since the edges are synonymous with patients in our event--patients--event database, we can correlate clinical data such as overall or progression-free survival times and probabilities with the patient trajectories.

Figure 4 on the left illustrates our graph database with three distinct trajectories the patients could undertake marked by red, black, and blue edges. The corresponding survival times and probabilities for the patients following the trajectories are plotted using the standard Kaplan-Meier curves. As we can see, such correlations between the paths and clinical data would allow us to identify critical molecular/clinical markers or network motifs and will also enable planning wet-lab experiments and clinical trials based on directed biological and clinical insights. The critical point is to note that correlations such as these are enabled **only because the events are associated with each other through patients** rather than patient phenotypes or any other relationships.

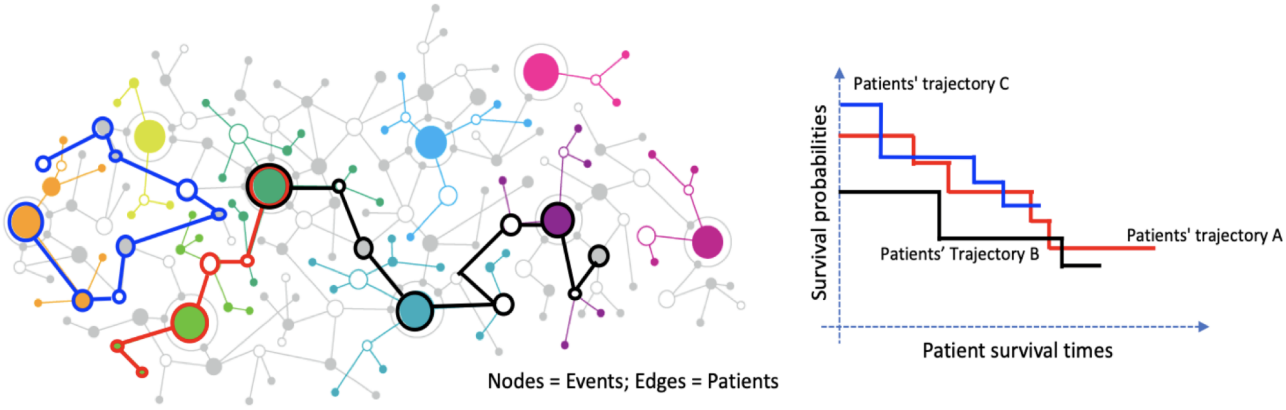

Figure 4: Correlating patient trajectories with their survival times; on left, three patient trajectories (red, black and blue); on right, the corresponding survival times and probabilities for those trajectories

We note that the association between the events/nodes as patients in this schema will prohibit us from answering questions on species levels (for example, is PI3K-pathway implicated in mouse GBMs?). However, the relationships could be modeled as species with human patients and any particular species as properties, allowing us to ask and answer questions on species levels and help us design, say, mouse experiments. Although Balaur et al.[9] suggest the idea of modeling nodes as events, it is for the first time we are proposing modeling edges as patients that would allow us to build a generalized hypothesis-generating engine that we can query to directly correlate with survival outcomes.

In general, by leveraging the longitudinal molecular, clinical, and imaging data, we can build a graph database-based recommendation engine that will serve to address several testable hypotheses for functional, clinical studies and enable iterative employment of the graph strategy in more refined, focused inquiries as the pre-clinical/clinical and molecular, imaging profiling data continues to expand. We, therefore, for the first time, aim to build a molecular and pre-clinical/clinical graph recommendation engine for longitudinal trials, which will serve as a fulcrum for research, clinical decision-making, and oncology, in general. However, we note that our model applies to any other disease equally well.

A graph database and a Cypher query

An example graph database using Glioma Longitudinal AnalySiS (GLASS) consortium data

We have developed a stand-alone version of the database integrating mutations and copy-numbers from the GLASS consortium data. Figure 5 shows the snapshot of the Neo4j database (a graph database) that incorporates recurrence in populations and clonal events in patients. It also integrates various CNV and mutation events and essential clinical information such as disease classification and patient identifier. Using this database, we can ask and answer questions combining CNV, mutations, and patient information to understand mechanisms that lead to glioma initiation and progression. For example, let us say we would want to identify strongly associated molecular events that lead to recurrence in the IDH Codel subtype. This question can be coded in Cypher query language and queried in the database as follows:

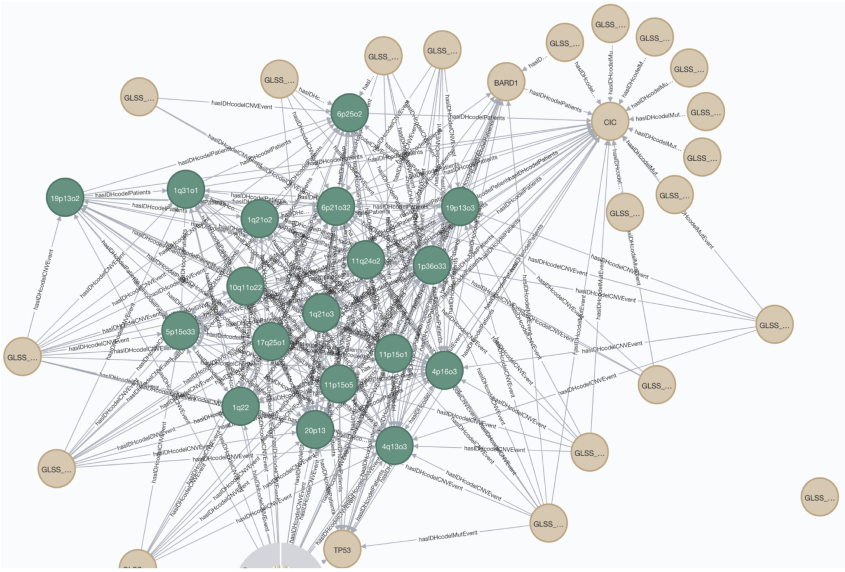

Figure 5: Snapshot of an example graph database for longitudinal glioma data

```
MATCH (m1:Patient:Tissue:IDHcodel) –[r]– (m2:Gene) –-
(m3:Tissue:IDHcodel) WHERE m1.event_type="TP" AND
m3.event_type="R1" OR m3.event_type="TP" AND
m1.event_type="R1" AND (r.node1_node2 >= 0.3) OR
(r.node2_node1 >= 0.3)
RETURN m1, m2 LIMIT 25
```

Figure 6 shows the result of this query that highlights that CIC mutations are centrally involved in recurrences in the IDH Codel subtype. By "strong association," we set the probability of recurrence as more significant than 0.3. Given this framework, we have run queries with a varying probability of recurrence for strong association and identified CIC as a critical molecule in the recurrent population of this subtype.

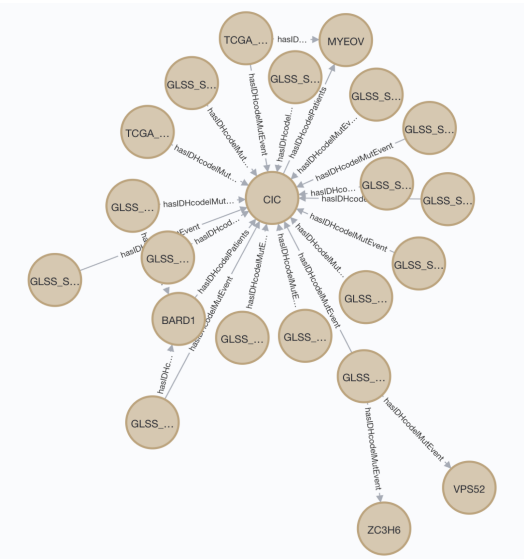

Figure 6: Results of the query identifying main molecular events in IDH Codel

Conclusion

Event-Patients-Event graph database framework is a powerful platform for realizing P4 medicine

In summary, graph databases are dynamic and mutable data structures that enable systems approach to medicine through the graph query language. In particular, the schema we developed where we identify two events through patients who undergo those events provides us with a compelling framework for building a powerful recommendation engine for longitudinal clinical trials. The two main advantages of this framework are (1) enabling generalized hypothesis generation on the population or species levels (when using species instead of patients) and (2) correlating survival times with the different trajectories' patients undergo during disease evolution. Moving from the data-centric paradigm to implementing this patient-centric framework will advance our objective of realizing P4 medicine well within our lifetime.

References

1. Alessandro Villa and Stephen T. Sonis. System biology. Translational Systems Medicine and Oral Disease, Academic Press, 9-16, 2020.
2. Senthil K. Muthuswamy, Self-organization in cancer: Implications for histopathology, cancer cell biology, and metastasis. Cancer Cell, 39(4):443-446, 2021.
3. Maher, Christopher A, and Richard K Wilson. Chromothripsis and human disease: piecing together the shattering process. Cell vol., 148,1-2:29-32, 2012.
4. Wong E, Baur B, Quader S, Huang CH. Biological network motif detection: principles and practice. Brief Bioinform., 13(2):202-215, 2012.
5. Mangan S, Alon U: Structure and function of the feed-forward loop network motif. Proceedings of the National Academy of Sciences, 100(21):11980-11985, 2003.
6. Mangan S, Zaslaver A, Alon U: The Coherent Feedforward Loop Serves as a Sign-sensitive Delay Element in Transcription Networks. Journal of Molecular Biology, 334(2):197-204, 2003.
7. Laurentino Quiroga Moreno, Graphlets and Motifs in Biological Networks. Encyclopedia of Bioinformatics and Computational Biology, Academic Press, 814-820, 2019.
8. Matelsky, J.K., Reilly, E.P., Johnson, E.C. et al. DotMotif: an open-source tool for connectome subgraph isomorphism search and graph queries. Sci Rep 11, 13045, 2021.
9. Balaur I, Saqi M, Barat A, Lysenko A, Mazein A, Rawlings CJ, Ruskin HJ, Auffray C. EpiGeNet: A Graph Database of Interdependencies Between Genetic and Epigenetic Events in Colorectal Cancer. Journal of computational biology: a journal of computational molecular cell biology, 24: 969-980, 2017.
10. Howorth, Gary & Kockar, Ivana. Do We Need a New Architecture for Simulating Power Systems? Conference: 8th International Conference on Simulation and Modeling Methodologies, Technologies and Applications, 190-197, 2018.
11. Deffur A, Wilkinson RJ, Mayosi BM, Mulder NM. ANIMA: Association network integration for multiscale analysis. *Wellcome Open Res*. 3:27, 2018.
12. Maher CA, Wilson RK. Chromothripsis and human disease: piecing together the shattering process. *Cell*. 148(1-2):29-32, 2012.