In *Advances in Network Science*, A. Wierzbicki, U. Brandes, F. Schweitzer, edts, *Proc. Int. Conf. NetSci-X 2016,* Wroclaw, Poland, Jan 11-13, 2016, Lecture Notes in Computer Science, vol. 5964, 2016, Berlin, pp. 178-185.

Synergy Landscapes: A Multilayer Network for Collaboration in Biological Research

Konstantin Kuzmin[1], Christopher Gaiteri[2], and Boleslaw K. Szymanski[1]

[1] Network Science and Engineering Center,
Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[2] Rush University Medical Center, Rush University, Chicago, IL 60612, USA
kuzmik@rpi.edu, gaiteri@gmail.com, szymab@rpi.edu

Abstract. Physical interactions among molecules, cells, and tissues influence research in biology. While conferences and departments are created to study these interactions, previous attempts to understand the large-scale organization of science have only focused on social relationships among scientists. Here, we combine the structure of molecular interaction networks with other science networks, such as coauthorship networks, for a more complete representation of the interests and relationships that determine the direction and impact of research. This multilayer network that we call *Synergy Landscapes* will allow us to identify broad patterns of scientific research, and in particular factors that predict innovative and high-impact research. Synergy Landscapes also will dynamically track research trends in a customized framework that informs scientists of research on molecules which are relevant to their core research areas. This will facilitate collaborations that would otherwise be difficult to produce and which mirror the natural organization of biological systems.

Keywords: Multilayer networks, Collaboration networks, Molecular networks

1 Introduction

Biologists frequently have a deep understanding of the experimental and disease relevance of specific molecules. In contrast to the historical emphasis on developing highly specific knowledge, omics technologies, which can measure several thousands molecular features simultaneously, have increased the breadth of knowledge about the molecular interactions that carry out biological functions. It is challenging to simultaneously perform detailed research on a core topic of interest and also understand the relevance of hundreds of molecules that are connected to this core via molecular networks. Collaboration enables combining expertise among researchers and conducting experiments that are both highly detailed and reflect the new information in omics data. Yet, finding relevant researchers to create synergistic effects is challenging when a single molecule may be relevant to many biological processes, each of which has its own complexity and nomenclature. Furthermore, the omics technologies that assess these interactions are evolving and growing, creating complex molecular networks that link researcher interests, but are rarely used in guiding researchers to beneficial collaborations.

We introduce a novel multilayer network approach to fostering innovation and collaboration among bio-medical researchers. The initial components of our multilayer networks are: (i) collaboration networks of bio-medical coauthors, (ii) networks of molecules interacting in bio-processes and papers describing them, and (iii) networks of bio-processes involved in different diseases.

The Synergy Landscapes project aims to combine those networks to establish new collaborative links between researchers, molecules, and diseases. This will enable, for example, identifying researchers that may collaborate in previously unknown ways to address complex diseases and tremendously impacting medical innovation and efficient disease research. The total effect will be synergistic, beyond a simple sum of the components.

2 Related Work

The idea of combining several different but related datasets into a single multilayer network is widely used in complex systems. The applications are mostly found in sociology and social information systems. A comprehensive review by Boccaletti et al. [1] contains a detailed description of the properties and structural and dynamic organization of networks that represent different relationships as layers. Such networks have shown utility in economics, technical systems, ecology, biology and psychology. We include molecular interaction networks as a novel layer in Synergy Landscapes. These networks originate from many experimental sources and model organisms. In many omics analyses it is now standard to project results into these networks structures, to identify the overall functional role of the results or additional related molecules. Many free and commercial online tools are available for this purpose (e.g., [3] and [4]). At the same time, methodologically related studies of coauthorship and human social networks have emphasized the relevance of network structure in determining patterns of collaboration [5]. Despite similar goals of understanding the scientific relevance of groups in molecular and social networks, the two approaches have never been fused to combine molecules and people in an integrated network space as shown in Fig. 1.

3 Synergy Landscape Concept and Use

Here, we introduce a unified solution to the dual problem of diverse causes of complex diseases and barriers in scientific collaboration. Our idea is that if molecules A and B interact, the researchers who study molecule A could benefit from interacting with those who study molecule B. The combined effect is achieved by connecting researchers and resources through the structure of molecular interactions. Fig. 2 shows how a synergy network combines relationships between molecules, ideas, and people.

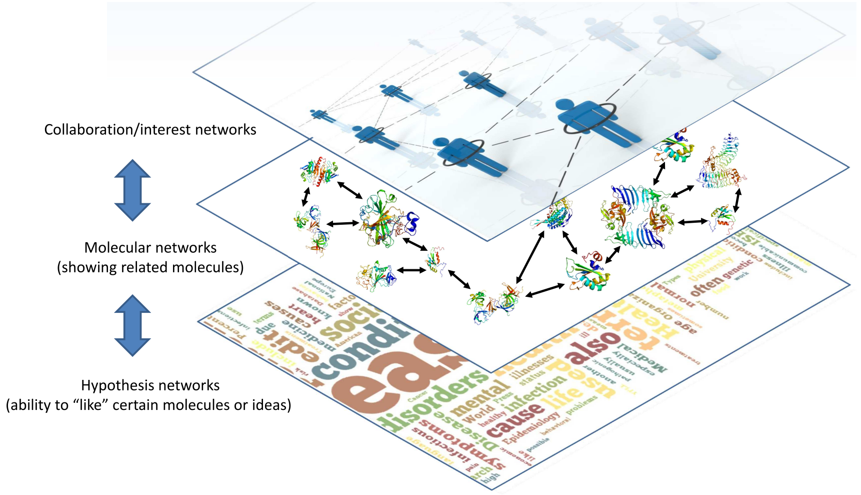
Collaboration/interest networks
Molecular networks
(showing related molecules)
Hypothesis networks
(ability to “like” certain molecules or ideas)

Fig. 1. Multilayer synergy network illustrating types of networks merged to form the basis of Synergy Landscapes. Network is prepopulated with molecular interests of specific scientists based on published papers in Scopus. Molecular interactions are determined from multiple sources. Specific molecular networks that are most relevant to the field of study of particular researchers can be selected as the basis for calculating researcher–researcher distances.

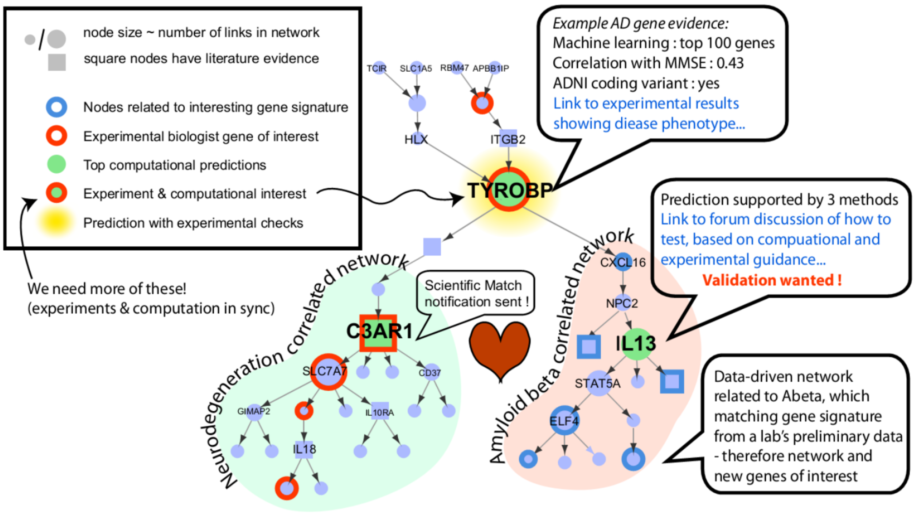
node size ~ number of links in network
square nodes have literature evidence
Nodes related to interesting gene signature
Experimental biologist gene of interest
Top computational predictions
Experiment & computational interest
Prediction with experimental checks
We need more of these!
(experiments & computation in sync)
Example AD gene evidence:
Machine learning : top 100 genes
Correlation with MMSE : 0.43
ADNI coding variant : yes
Link to experimental results
showing diease phenotype...
Prediction supported by 3 methods
Link to forum discussion of how to
test, based on compuational and
experimental guidance...
Validation wanted !
Data-driven network
related to Abeta, which
matching gene signature
from a lab’s preliminary data
- therefore network and
new genes of interest
Scientific Match
notification sent !
TCIR
SLC1A5
RBM47
APBB1IP
HLX
ITGB2
TYROBP
Neurodegeneration correlated network
C3AR1
SLC7A7
CD37
GIMAP2
IL10RA
IL18
Amyloid beta correlated network
CXCL16
NPC2
IL13
STAT5A
ELF4

Fig. 2. An example of how molecular entities are annotated with human interests in Synergy Landscapes.

The first use we consider is designed around the ways a typical scientist can utilize Synergy Landscapes to increase funding and publications. The ultimate objective is to suggest customized, optimal research directions and collaborators for users. The Synergy Landscapes multilayer network will be accessible as a website searchable for molecules or people based on a user supplied list of query terms. The output will resemble a personalized newsfeed based on the specific interests of the user, as shown in Fig. 3.

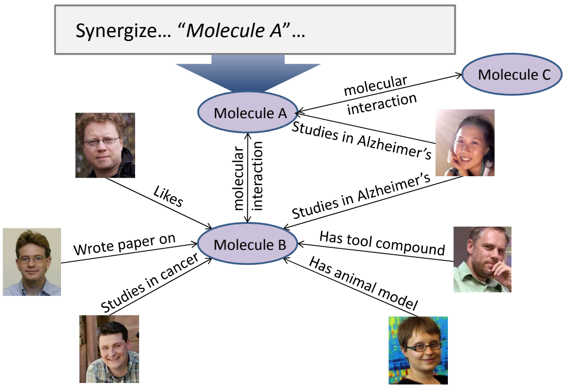

Fig. 3. Example of search functionality on Synergy Landscapes. The graphical output emphasizes underlying molecular networks.

The search engine will traverse edges in all three component networks in paths that enable discovery of "neighboring" scientists, ideas, and resources. Multiple molecular networks can be used to individually or collectively compute researcher–researcher distances and to predict research synergy. As new omics resources become available they will facilitate customized research landscapes and updated distances between researchers. For instance, researchers who primarily utilize *drosophila* will find molecular interactions in that system most relevant to guiding them to collaborators by selecting a *drosophila*-based molecular network and then surveying the landscape around them. Adding new interaction knowledge to Synergy Landscapes acts like a molecular wormhole — bringing some researchers who were previously distant into close contact.

To detect less obvious potential collaborators who could contribute to highly innovative research we will classify adjacent researchers into those who are within the user's community (defined by co-authorship clusters or location) vs. those who link to other research communities. The latter, which can be highlighted in the user interface (UI), may be ideal partners for interdisciplinary projects.

The Synergy Landscapes will also provide data to thoroughly study patterns of innovation and significance in research, and then to facilitate high-impact findings. Thanks to the availability of date-stamped and cross-referenced publications, it is possible to track the origin of influential trends in terms of how they are positioned in the molecular and coauthorship networks. This enables predicting topics and pairs or groups of researchers who are likely to collaboratively produce valuable findings. When such a matching is predicted, users can receive notifications whenever a "nearby" publication appears meaning that potential collaboration can result in a high-impact paper.

Institutions can utilize the hybrid network of researchers and molecules to improve efficiency and to organize collaborations across thousands of researchers. Synergy Landscapes creates an expansive definition of the molecules and humans that are relevant to a particular topic. In practice, by identifying their core molecules of interest, conference organizers can identify a radius of related researchers, even when those researchers do not formally belong to the field nor study molecules that are traditionally associated with the field. In this way conference organizers can recruit a diverse yet appropriate set of conference presenters. Using Synergy Landscapes, the participants will be able to meet other people who are likely to collaborate on future projects.

Another use of Synergy Landscapes is ranking job applicants based on their average distance in the molecular landscapes from all researchers currently on the team. Similarly, the connectivity of potential hires to two teams can be calculated in Synergy Landscapes. This provides a quantitative measure of the likelihood of future collaboration patterns that fulfill team objectives.

4 Architecture

Synergy Landscapes is designed to follow a multi-tier architecture model in order to separate presentation (UI), application processing, and data manipulation from each other. Moreover, each layer communicates with other layers using well-defined standardized protocols. Therefore, the internal implementation of each layer can be changed without affecting any other layers or requiring any changes in other parts of the system. Such an approach provides excellent scalability and enables easy expansion through the modular structure of its components.

The Synergy Landscapes architecture is discussed in the context of a typical expected query. One example is searching for authors working on molecule m who also worked on diseases d_i and d_j and another is finding diseases that were studied by researchers who considered molecule m in their publications. The architecture should also enable more complex queries. For example, one can start with some molecule m_i and find all diseases with which m_i has been associated in past publications. Then it would be possible to find if some other molecules m_j and m_k have ever been studied with those diseases and if so what authors and publications were involved. Finally, it can be determined if a pair of molecules m_i and m_j is associated with different diseases and who were the experts who described those reactions in their publications. The basic molecular network should be extensible — for instance by introducing additional layers associated with medications and their relationships with molecules and diseases. Similarly, the architecture should support user-selected subsets of networks that reflect the relationships most relevant to their research interests. The architecture described below supports these expected queries in a scalable extensible framework.

Synergy Landscapes obtains its data from the Scopus database. According to the study by Falagas et al. [2] which compares PubMed, Scopus, Web of Science, and Google Scholar, Scopus offers about 20% more citation coverage than Web of Science and more consistent search results than Google Scholar. In addition, Scopus provides a convenient API in a form of RESTful services which can be queried by user code.

User queries are executed against the multilayer network which is gradually built layer by layer. First, a network of molecules is created. The initial list of names and aliases of molecules is processed into a network where each unique molecule is represented as a node, and edges correspond to relations between molecules. Although some edge information can be inferred from the initial list of molecules, the major part of the connectivity information corresponds to the relations which are not described by the initial data and which we would like to discover through our synergistic process. Following the same procedure that was used to create the network of molecules, additional layers can be added (e.g., based on the list of diseases) to enrich our multilayer network and provide greater flexibility in our ability to generate subsequent layers. For instance, we might consider not only publications related to certain molecules but also those which mention specific diseases.

Publication data are used as the source for the second group of layers in our multilayer network. The source publication data are extracted from an existing source or sources based on a list of search terms which are already available from layers of molecules, diseases, etc. As a result, several network layers can be generated from this data.

First, the publication layer of the network is generated with nodes representing publications and edges connecting publications which are related in a certain way (e.g., which are dedicated to the same molecule or disease). At this point there are no edges in this layer as relations are to be determined after processing subsequent layers and discovering associations between different parts of the network. Similarly to the publications layer, a layer of grants is also created. Since the publication layer is created from the molecule layer and the disease layer, the publication–molecule and publication–disease cross-layers are easy to build.

For instance, in a publication–molecule cross-layer, an edge connects a certain molecule to the publications which are known to refer to this molecule. Likewise, a publication–disease layer links diseases with publications dedicated to them. Such cross-layers represent layers consisting entirely of edges. Moreover, instead of connecting nodes of a single underlying node set, the edges in cross-layers go “vertically” across any two different layers, effectively “stitching” them together. Therefore, cross-layers are fundamental entities in the multilayer network since they facilitate "vertical" connectivity between layers and allow network analysis tools to traverse the whole stack of layers rather than being trapped in any single one of them. In a wider context, a cross-layer with two corresponding node sets can be regarded as a separate bipartite network linking two different entities (publications and molecules, authors and diseases, etc.)

An additional dimension to the publication layer is created by utilizing the publication–topic association. This layer is another example of a layer using the same node set as some other layer or layers (in this case, it is the set of publications) but constructs a completely different edge set. In the topic layer, two publications are connected if they share the same subject area classification as declared by the authors and recorded in the underlying publication database. For instance, Scopus provides a codified subject area classification; therefore, publications classification is consistent throughout the database. Each publication can be marked as related to zero or more subject areas. The weight on an edge is determined by the *overlap coefficient* according to the number of common subject area classifications that two publications share. Given the publication node set P, the overlapping coefficient (also known as the Szymkiewicz–Simpson coefficient[6]) is defined by (1) as follows:

$$w(p_i, p_j) = \frac{|SA_{p_i} \cap SA_{p_j}|}{min(|SA_{p_i}|, |SA_{p_j}|)} \tag{1}$$

where $p_i \in P$ and $p_j \in P$ are two publication nodes, and SA_{p_i} and SA_{p_j} are the sets of subject area classifications of the corresponding publications. A layer which links publications with index terms (index terms layer) is created following the same approach as for the topic layer.

Then, since each publication or grant also lists authors, the author and collaboration networks are naturally created from the same data used for the publication layer. For this layer, nodes represent authors and weighted edges connect authors who have collaborated on at least one publication or grant proposal. A publication–author cross-layer is also created. It consists of unweighted undirected edges linking publications with their authors.

Finally, a citations dataset is the third network layer extracted from the publication data. A citation layer contains only directed edges which connect nodes from the underlying publications. An edge from publication i to publication j is added to the layer if publication i cites publication j. Thus, the layer represents a "cites" relationship.

Using nodes from the publication layer, the author layer, the citation layer, and the publication–author cross-layer, an author citation layer is created. Although derived, this layer provides a convenient way of establishing links between different authors who cite the work of others. The author citation layer is comprised of edges only, using nodes from the author layer as its nodes. There is an edge from person i to another person j if and only if author i has ever cited any publication which was authored by j. Given the sets of nodes of authors A and publications P, the weight of an edge in the author citation layer is determined by the fraction of the number of times author i cited author j in their publications to the total number of times author i cited other authors' work, as given by (2):

$$w(a_i, a_j) = \frac{|\{p \in \mathcal{PC}_{a_i} | a_j \in \mathcal{A}_p\}|}{\sum_{p \in \mathcal{PC}_{a_i}} |\mathcal{A}_p|} \quad (2)$$

where $a_i \in A$ and $a_j \in A$ are two author nodes, $\mathcal{PC}_{a_i} \subseteq P$ is the set of all publications cited by a_i, $\mathcal{A}_p \subseteq A$ is the set of authors of publication p.

Once all the layers have been created, we can start querying our multilayer network to provide useful information about authors, molecules, diseases, and publications. The result of each query can be saved as a set which, in turn, can be used for subsequent queries. Thus, Synergy Landscapes can provide meaningful answers to complicated questions by combining the data from network layers with additional filtering and grouping capabilities of reusable queries.

5 Conclusions and Future Directions

Molecular networks play an increasing role in disease research. They are also central to Synergy Landscapes which uses them to facilitate scientific results, accelerate research, and foster interdisciplinary collaborations. These capabilities are directly useful to scientists but also essential for understanding consistent social and molecular features of the most innovative and cited scientific projects. Because Synergy Landscapes is the first hybrid human–molecular network, it opens the doors to improved higher-level management and distribution of scientific resources. For instance, granting institutions may use Synergy Landscapes to study the impact of their funds and their distribution across the community structure around their topics of interest. In this way, not only can scientists respond to incentives, but the incentives themselves can be created to achieve certain objectives in light of the current distribution of scientific interest and resources.

References

1. Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C.I., Gómez-Gardeñes, J., Romance, M., Sendina-Nadal, I., Wang, Z., Zanin, M.: The structure and dynamics of multilayer networks. Physics Reports 544(1), 1–122 (2014)
2. Falagas, M.E., Pitsouni, E.I., Malietzis, G.A., Pappas, G.: Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. The FASEB journal 22(2), 338–342 (2008)
3. Krämer, A., Green, J., Pollard, J., Tugendreich, S.: Causal analysis approaches in ingenuity pathway analysis (ipa). Bioinformatics p. btt703 (2013)
4. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., Morris, Q., et al.: Genemania: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol 9(Suppl 1), S4 (2008)
5. Newman, M.E.: Coauthorship networks and patterns of scientific collaboration. Proceedings of the national academy of sciences 101(suppl 1), 5200–5205 (2004)
6. Simpson, G.G.: Notes on the measurement of faunal resemblance. American Journal of Science 258(2), 300–311 (1960)