Driving on Cellular Pathway #66

Luis A. Nunes Amaral

Dept. of Chemical and Biological Engineering and Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208 USA

Abstract. The interconnectedness of gene regulation, protein interaction, and metabolic networks is responsible for the remarkable efficiency and adaptability of biological systems, as well as the extraordinary challenges facing researchers trying to understand them. The torrents of new biological data generated daily should lead to overcoming the challenges to understanding biological processes. However, our understanding of these systems has not grown proportionally to the amount of data generated. This disparity arises from the fact that the behavior of a biological system is not a linear superposition of the behaviors of its components. Higher-level structures within organisms can be maintained precisely because of the complex network of nonlinear interactions among lower-level components. As a result, scientists increasingly recognize that in order to advance our ability to understand and purposefully manipulate biomedical systems, we must take a systems level approach. However, it is not yet clear *what* systems level approach is optimal. I contend that we will only be able to make sense of systems-level information if we can develop methods that enable us to extract the small set of information that is relevant at the scale of interest.

Keywords: gene regulation, complex network, biomedical systems.
PACS: 87.16.Yc

How Will Big Pictures Emerge from A Sea of Biological Data?

This question, posed in this form in a recent issue of *Science* [1], is arguably one of the most important and challenging questions we now face as biomedical scientists. Torrents of new biological data are being generated daily, but our understanding of biological systems has not grown proportionally. Consider the protein-protein interaction networks of fruit flies [2] and humans [3]displayed in the respective articles. Two facts become immediately apparent when looking at these figures. First, the networks look quite different. In [2], the network looks quite planar while in it appears organized in bands. Moreover, there is not standard set of symbols. Second, one would be hard-pressed to identify the proteins of system-wide importance or even different pathways/modules.

It is not surprising that our understanding is playing catch-up to the data. Our brains have evolved to handle in a meaningful manner only a handful of different pieces of information. Indeed, the reductionist approaches that have dominated science for the last several centuries relied precisely on reducing the number of dependent variables, with the idea that understanding would be gained by studying one component at a time. Unfortunately, it has become clear that reductionist approaches are not going to enable us to solve many of the most important biomedical questions. Understanding a single neuron is not going to enable us to understand the *emergence* of consciousness. The behavior of the entire system is not a linear combination of the behaviors of its components. In fact, the system has emergent properties that result precisely from the complex nonlinear interactions among the components.

CP922, *Noise and Fluctuations, 19th International Conference*, edited by M. Tacano, Y. Yamamoto, and M. Nakao

Contrast these networks with the maps one can find at any gas station and their ability to convey complex geopolitical information. In the continental US, there are approximately 20,000 localities (villages, towns, and cities) connected by millions of roadways. Most of us have never heard about the vast majority of those localities. However, we can easily locate even rather unremarkable towns. The reason is that maps have a remarkable property. They present information in a scalable manner. That is, *even as the amount of information increases, the representation is able to extract the information that is relevant at a given scale of observation.*

If we are to study biomedical systems at the system level, and still be able to make sense of the information we have, we must develop appropriate "cartographic" methods. The ultimate goal of my research is to make deciphering complex biological networks as easy as it is to find the best route between Evanston and Boston.

Cellular Processes as Networks

Cellular processes are typically comprised of a large number of potentially heterogeneous components. These components are connected through a web of interactions that defines a graph or network. The study and representation of networks has a long history, dating back to the 1700s and Euler's work of the Koenigsberg bridges' problem.

More recently, special attention was given to random networks [4]. Random networks form the "maximally disordered" end of a spectrum of possible network topologies. At the opposite of the spectrum of possible network topologies, one has fully ordered, finite dimension lattices.

Whereas the analysis and representation of random and ordered networks is straightforward, significant challenges exist when considering other types of networks. Significantly, the networks we find in cellular processes and in other real-world systems are neither random nor ordered [5]. The significance of these more general classes of networks was first demonstrated for social systems by Stanley Milgram [6,7], but much important work has since been conducted by other social scientists such as Granovetter, White, Freeman et al [8].

Recently, the characterization and modeling of complex real-world networks has gained incredible impetus. This interest has result in significant advances spearheaded by, among others, Watts [9], Amaral [10,11,12], Barabasi [13,14], Newman [15], Vespignani [15,17] and Alon [18].

Presently, complex networks are analyzed from two main perspectives. A popular perspective, which has been advocated by Barabasi [14], focus on obtaining a global, *but averaged*, picture of a complex network. Reference quantities include the distribution of number of connections (i.e., the degree) of each node, the average minimum path length, and the average degree of cliquishness of the nodes. Unfortunately, these average global quantities are only informative and adequate when one of two strict conditions is fulfilled: (i) the network lacks a modular structure, or (ii) the network has a modular structure but (ii.a) all modules were formed according to the same mechanisms, (ii.b) all modules have similar properties, and (ii.c) the interface between modules is statistically similar to the bulk of the modules, except for the density of links. If neither of these two conditions is fulfilled, then any theory proposed to explain, for example, a scale-free degree distribution must take into account the modular structure of the network. To my knowledge, no real-world network fulfills either of the two conditions above, implying that *global properties are unlikely to provide insight into the mechanisms responsible for the formation, growth, and function of these networks.*

An alternative perspective, suggested by Alon [18], approaches the characterization of

complex networks from the bottom-up. Specifically, when following this approach one attempts to identify local patterns, i.e., small subgraphs, that are present significantly more (or less) than one would expect from chance alone. This approach, which requires techniques rather more sophisticated than the global approach, has the great advantage of not requiring any unrealistic assumptions about the homogeneity of the network. Unfortunately, one shortcoming of this approach is that it cannot be used to accurately identify large-scale patterns, and thus it cannot provide a true understanding of the *global* organization of the network.

The Cartographic Approach

My research group has introduced and is following a third perspective that aims to do for the representation of complex networks what cartography did for the representation of geopolitical information. Specifically, we developed new algorithms [19,20] that accurately extract the most significant information, at a given scale, from complex networks. These algorithms wed concepts and techniques—such as optimization by simulated annealing, energy landscapes, scaling and universality, and structural equivalence—that originated in several different disciplines, including chemistry, computer science, statistical physics, and social network analysis.

The cartographic approach is based upon two assumptions. The first assumption is that the nodes in a network can be grouped into modules. The modules are analogous, in the geographic picture, to regions or neighborhoods and enable a coarse-grained and, thus simplified, description of the network. Guimera and Amaral's algorithm, which builds directly on Newman's [15] algorithm for grouping nodes into modules, rests on the expectation that nodes are more tightly connected to other nodes in the same module than to nodes outside of their module, just as one would expect two researchers from the same department to have a greater likelihood of collaborating than researchers from different departments.

The second assumption at the core of the cartographic approach is that one can classify the nodes comprising a network into a small number of *system-independent* "universal roles." Our algorithm for classifying nodes into roles rests on the expectation that the nodes in a network are connected according to the *role* they fulfill. For example, most large universities have a president who is in direct contact with high-level administrators, the members of the board of trustees, and a few high-profile faculty members, but not with the typical assistant professor or the typical student. Importantly, this fact holds irrespective of the particular university one considers. We thus define the role of a node according to (i) how many connections it has within its module and (ii) what fraction of its connections are with nodes in outside modules. We defined four main types of roles: *hub connectors*, which have many connections to both other nodes in their module and to nodes in other modules; *provincial hubs*, which have many connections but only to nodes inside their module; *satellite connectors*, which have few connections but act as bridges between modules; and *peripheral nodes*, which have few connections and mostly to nodes inside their module.

To demonstrate the power of this novel cartographic perspective, we studied the overall organization of the cellular metabolisms of twelve organisms—four archae, four bacteria and four eukaryotes [19]. We found that the metabolic networks of each of these organisms comprised about 20 modules that correlate very strongly with previously identified metabolic pathways [21,22]. Additionally, we found that about 90% of the metabolites in these organisms are classified as peripheral nodes, i.e., metabolites that participate in a few reactions, mostly within a single pathway. This result indicates a very weak signal-to-noise ratio—the important metabolites are a small fraction of all metabolites, and thus very difficult to identify using standard representations.

The most striking result revealed by our analysis, however, is that metabolites classified as satellite connectors, those which participate in a small number of reactions (i.e., have few connections) but have a significant fraction of their connections to metabolites outside their module, *are significantly more conserved across species* than provincial hub metabolites (which participate in a much larger number of reactions). This is a remarkable result, comparable to finding a needle in a haystack: Out of the hundreds of metabolites with a small number of connections, our cartographic representation identifies the 5-10 metabolites that were conserved (and thus are presumably critically important) in the metabolic networks of organisms that diverged more than one billion years ago.

Probing the Hierarchical Structure of Biological Processes

Biological systems have a hierarchical organization (entire organisms are comprised of organs, which are comprised of tissues, which are comprised of cells, and so on…). At present there are no methods for the identification of the hierarchical organization of nodes in a network that fulfills two necessary requirements: (i) accuracy for many types of networks, and (ii) ability to identify the different levels in the hierarchy as well as the number of modules and their composition at each level. The first condition may appear as trivial, but we make it explicit to exclude algorithms that only work for a particular network or family of networks.

The second condition is more restrictive, as it excludes methods whose output is subject to interpretation. Specifically, a method does not fulfill the second condition if it organizes nodes into a tree structure, but it is up to the researcher to find a "sensible" criterion to establish which are the different levels in that tree. An implication of the previous two requirements is that any method for the identification of node organization must have a null output for networks, such as random graphs, which do not have an internal structure.

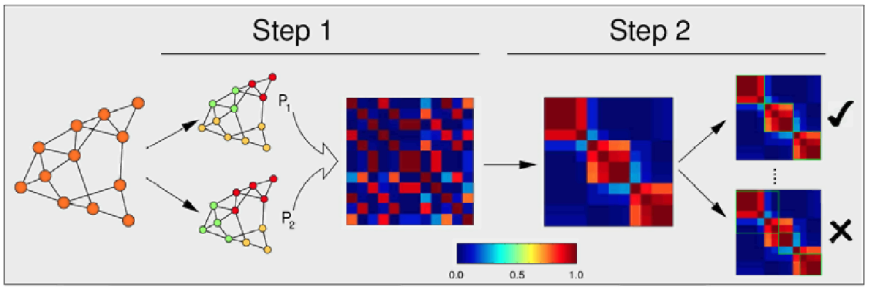

FIGURE 1: Illustration of method for extracting hierarchical organization of large complex networks.

To my knowledge, there is no procedure that enables one to simultaneously assess whether a network is organized in a hierarchical fashion and to identify the different levels in the hierarchy in an unsupervised way. Indeed, many methods, such as hierarchical clustering, yield a tree even for networks with no internal structure. My research group has developed a new method that is able to determine the hierarchical structure of complex networks of arbitrary type (see figure below). The first step in our method is to measure the affinity between two nodes in the network. We define the affinity between two nodes as the probability that these nodes are placed in the same module when locally maximizing a "modularity function" (Sales-Pardo et al 2007). The affinity matrix thus obtained enables us to determine if the network has a *statistically significant* modular structure.

If the network is modular, our method then proceeds to step 2, which consists in ordering the affinity matrix so as to obtain the most block diagonal structure possible. The matrix has a

block diagonal structure if there are square boxes with high-values of affinity along the diagonal of the matrix. Finally, we fit a block diagonal model to the ordered matrix. The boxes yield the top-scale modules in the network. We then iterate the above procedure for the sub-networks defined by the nodes in each module.

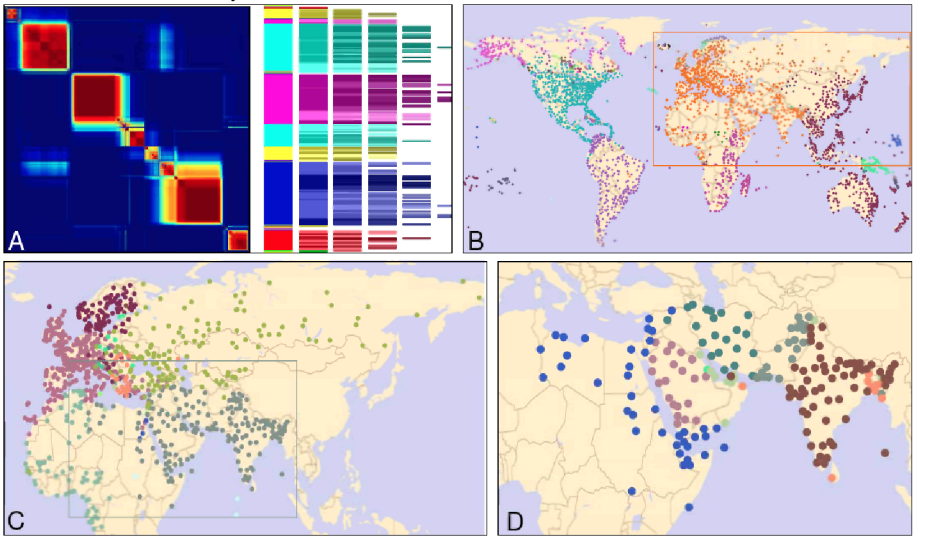

FIGURE 2: Hierarchical organization of worldwide air transportation network. (A) The block-diagonal organization of the affinity matrix clearly reveals the modular structure of the network. Note the box-within-box structure of the second from left large block, hinting at the hierarchical structure. The color bars on the right represent the different modules and sub-modules. (B) Airports in the network. Different color indicated different top-level modules. The "Old World" module is shown in orange. (C) Second-level organization of the "Old World" module. Different colors indicate different sub-modules. (D) Third-level organization of the "Middle East" sub-module. Different colors indicate different sub-sub-modules.

We have applied our method to both model networks and real world networks. In figure 2, I show the results obtained for the worldwide air transportation network. Panel A shows the affinity matrix for the entire network. It is clear that the matrix has a block diagonal structure (red-yellow squares along diagonal). Each box corresponds to a top-level module. On panel B, the circles indicate the location of the airports, and colors indicate module membership at the top level. Note how the system is broken down into "sensible" modules. Panels C and D show the partitions of modules at the second and third level of the hierarchy, respectively.

Future Directions and Concluding Remarks

The problem of developing optimal cartographic methods for the study and representation of complex networks is far from solved. Indeed, our methods currently have a number of limitations that need to be addressed before they can be applied more widely. In the following, I will briefly discuss some of the challenges ahead.

Multipartite networks —Complex systems are, by nature, heterogeneous. For example, enzymes catalyze reactions among metabolites. Current network analysis methods, including our own algorithms, were not designed to deal in a natural manner with networks comprising different types of nodes (i.e., multipartite networks).

Computational load — The algorithms we have developed for the identification of modules

are computationally very demanding. For example, our algorithm for the identification of modules requires a computation time that increases with the square of the number of nodes in the network. Thus, going from a network with 1000 nodes to one with 5000 nodes results in a twenty-five fold increase of the computation time.

Network representation — At present there are no standard, universally-accepted, software packages and no standard, universally-accepted, methods for representing biological information in a *scalable way*. This is clearly essential if we want researchers work on different problems to be able to communicate their results to a broader audience.

I believe that in order to move forward our understanding of biomedical systems, we must take a systems level approach. However, we can only make sense, and make use of, systems level information if we are able to develop methods that enable us to extract the small set of information that is significant at a given scale of observation. A scalable cartographic representation of a complex biological reality will enable us to purposefully design or re-engineer biological systems for therapeutic purposes. I imagine a time in which designing a molecular-level therapeutic approach will be similar to planning a driving route between two distant cities.

References

1. Pennisi E. (2005) *Science* **309**, 94.
2. Giot L. et al (2003) *Science* **302**, 1727.
3. Rual J.-F. et al (2005) *Nature* **437**, 1173.
4. Bollobas B. (1985) *Random Graphs* (London, Academic Press).
5. Amaral L. A. N. & J. M. Ottino (2004) *Eur. Phys. J. B* **38**, 147.
6. Milgram S. (1967) *Psychol. Today* **1**, 61.
7. Travers J. & S. Milgram (1969) *Sociometry* **32**, 425.
8. Wasserman S. & K. Faust (1994) *Social Network Analysis* (Cambridge University Press, Cambridge, UK).
9. Watts D. J. & S. H. Strogatz (1998) *Nature* **393**, 440.
10. Barthelemy M. & L. A. N. Amaral (1999) *Phys. Rev. Lett.* **82**, 3180.
11. Amaral L. A. N., A. Scala, M. Barthelemy & H. E. Stanley (2000) *Proc. Nat. Acad. Sci. USA* **97**, 11149.
12. Mossa S., M. Barthelemy, H. E. Stanley & L. A. N. Amaral (2002) *Phys. Rev. Lett.* **88**, art. No. 138701.
13. Barabasi A.-L. & R. Albert (1999) *Science* **286**, 509.
14. Albert R. & A.-L. Barabasi (2002) *Rev. Mod. Phys.* **74**, 47.
15. Newman M. E. J. (2003) *SIAM Rev.* **45**, 167.
16. Pastor-Satorras R. & A. Vespignani (2001) *Phys. Rev. Lett.* **86**, 3200.
17. Pastor-Satorras R. & A. Vespignani (2001). *Phys. Rev. Lett.* **87**, art. No. 258701.
18. Milo R. et al (2002) *Science* **298**, 824.
19. Guimera R. & L. A. N. Amaral (2005) *Nature* **433**, 895.
20. Guimera R. & L. A. N. Amaral (2005). *J. Stat. Mech. Theor. Exp.* **2**, art. No. P02001.
21. Lee S. Y. & E. T. Papoutsakis, eds. (1999) *Metabolic Engineering* (Marcel Decker).
22. Schuster S., D. A. Fell & T. Dandekar (2000) *Nature Biotechnol.* **18**, 326.
23. Guimera R., S. Mossa, A. Turtschi, and L. A. N. Amaral (2005) *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7794.