Published in final edited form as:
Cell. 2014 April 24; 157(3): 534–538. doi:10.1016/j.cell.2014.03.009.

Siri of the Cell: What Biology Could Learn from the iPhone

Anne-Ruxandra Carvunis[1] and **Trey Ideker**[1,*]

[1]Department of Medicine, University of California, San Diego, La Jolla, CA 92093, USA

Abstract

Modern genomics is very efficient at mapping genes and gene networks, but how to transform these maps into predictive models of the cell remains unclear. Recent progress in computer science, embodied by intelligent agents such as Siri, inspires an approach for moving from networks to multiscale models able to predict a range of cellular phenotypes and answer biological questions.

In case you haven't heard, Siri is the virtual personal assistant on Apple's iPhone operating system. Your wish is Siri's command. It understands what you say and can take action on your behalf, such as sending messages, scheduling meetings, placing phone calls, or researching nearby hotels and restaurants. If you speak directly to your mobile device and ask, for instance, "find me a good sushi restaurant for two tonight," the screen will display a list of restaurants from which you can choose to make your dinner reservation.

Although Siri has little to say about cell biology, it is easy to think of biological questions one might ask, if only Siri could answer. For example:

"Patient P has a tumor recurrence with new mutations X and Y—which drugs should I prescribe"?

—Siri as clinical decision support system (Berner, 2007) for diagnosing and treating patients.

"In estradiol-treated SKBR3 cells, which nuclear protein complexes have the greatest change in phosphorylation"?

—Siri as virtual laboratory assistant, suggesting which western blot to do next.

"What is the largest number of genes I can knock out of Mycoplasma for it to remain viable"?

—Siri as synthetic biologist, helping to design the minimal genome.

Here, we discuss why and how, one day, biological questions like these might be answered by a Siri of the cell. We argue that, more than a whimsical analogy, intelligent agents like

*Correspondence: trey@bioeng.ucsd.edu.

Siri inspire new directions for modeling cellular phenotypes and answering biological questions.

A Progression of Cellular Models

Like any model of the world, our view of the cell is inescapably bound by the time and place in which we live. Cells were discovered during the Renaissance directly following the invention of the microscope and were initially depicted as tiny walled rooms analogous to monk's quarters, hence the name "cells" (Hooke, 1665). Later, scientists of the Industrial Age imagined cells as mechanical devices akin to engines, boats, and bridges (Thompson, 1917), leading to the development of the field of biomechanics (Fung, 1993). Other schools have fashioned the cell in a variety of forms, from bags of enzymes (Mathews, 1993), to metabolic channels (Reddy et al., 1977), to feedback circuits (Monod et al., 1963), to complex systems (Kauffman, 1993), to gels (Pollack, 2001), to self-modifying programs in software (Bray, 2009).

A model that has pervaded cell biology for the past 15 years is the so-called "network" view (Figure 1A), which has bloomed in parallel with the emergence of manmade networks such as the Internet and Facebook. This view treats cells as containers for vast networks of "nodes" (genes, gene products, metabolites, or other biomolecules) connected by "links" (physical interactions or functional associations) (Barabási and Oltvai, 2004). Network representations of the cell flow directly from the ability to characterize not only genes and proteins in isolation but also their functional similarities and physical binding partners—a major outcome of transcriptomics and proteomics approaches. Analysis of network information, whether biological or manmade, is an active field leading to algorithms that detect nodes with strategic positions within a network (Barabási and Oltvai, 2004) or that partition tightly interconnected groups of nodes to identify modular structures (Fortunato, 2010).

Why It Is Time to Move beyond Networks to Hierarchies

Although incredibly influential, the network is probably not the ultimate representation of a cell for two reasons. First, network diagrams do not visually resemble the contents of cells. Nowhere in the cell do we observe actual wires running between genes and proteins, unlike for the Internet, which is truly a network of wires among processing units. Rather, the cell involves a multiscale hierarchy of components that is not readily captured by basic network representations. For example, the proteasome has been mapped extensively to identify its key genes and interactions, but the network visualization of these data (Figure 1A) is very different from the proteasome's spatial appearance (Figure 1B). The interactions making up the proteasome factor into a regulatory particle and a core, which, in turn, factor into a base and a lid, and an α and β subunit, respectively (Figure 1C). This hierarchical structure is obscured by the network visualization of pairwise relationships between gene products.

Second, many of the molecular networks published to date are descriptive maps of physical or functional connectivity rather than predictive models. For example, technologies such as yeast two hybrid, protein affinity purification, and chromatin immunoprecipitation are often used to define and draw large networks of protein-protein and protein-DNA interactions

(Chuang et al., 2010), but these static maps do not, by themselves, predict cell behavior. Although the field of systems biology has inferred networks capable of predicting gene function or phenotypic responses (reviewed in Koller and Friedman [2009] and Walhout et al. [2013]), these efforts have tended to focus on a specific class of predictions, i.e., gene expression level or cell growth rate. Assembling a model of the cell that would predict a range of phenotypes, rather than only one type of outcome, requires understanding how cellular phenotypes are interrelated with each other. Here, again, a hierarchy comes into play because cellular organization involves a multiscale hierarchy not only in structure but also in function. For example, the proteasome is a central component of ubiquitin-mediated protein degradation, which, depending on an intricate set of inputs and rules, can result in cellular homeostasis, differentiation, death, and other fates. This multiscale hierarchy of processes is, again, simply not exposed by a standard pairwise network representation.

The most direct representations of data are not always the most desirable for meaningful interpretation of those data. In X-ray crystallography, the most direct representations of X-ray diffraction patterns are 2D images (McPherson, 2009). However, when many such images are integrated and analyzed, exquisite 3D structural models of proteins emerge, which, in turn, enable accurate predictions of protein dynamics and function. Similarly, from many molecular measurements and interaction data sets, the higher-order structure and function of the cell might emerge if only we could figure out how to assemble these images properly.

Capturing Hierarchy with Ontologies

To capture hierarchical organization, a particularly promising direction in computer science has been the development of the ontology, a model that divides its object into a set of fundamental entities and relationships among those entities (Gruber, 1995). Ontologies arise from a branch of philosophy known as metaphysics, which is concerned with the nature of what exists and the categories into which the world’s objects naturally fall. Ontologies build upon and extend network models in two key ways: “entities” refer not only to elemental objects but also to any meaningful grouping of objects, and “relationships” refer not only to direct connections but also to nested structures, such as one entity being a part or type of another. Thus, ontologies explicitly allow for a higher-order organization of knowledge, missing from raw networks. They have been key for building powerful knowledge representation and reasoning systems in many domains (Brachman and Levesque, 2004), including biomedicine (Robinson and Bauer, 2011).

Ontologies have become influential in cell biology through the development of the Gene Ontology (GO) (Ashburner et al., 2000). GO is a major resource of knowledge about genes, gene products, and the hierarchy of cellular components, molecular functions, and biological processes in which they participate. Entities in GO (called GO terms) are hierarchical groupings of other entities. For example, the biological process of “DNA replication elongation” is a type of “DNA strand elongation” and is a part of the more general process of “DNA replication” (Figure 2A). The GO resource is presently very large, with nearly 35,000 GO terms connected by 65,000 hierarchical term-term relations, describing more than 80 different species. The impact of GO is hard to overstate—just try to think of a single

modern “omics” analysis that does not use GO to validate a novel data set or approach or to generate new mechanistic hypotheses. In a sense, GO is the most universal, and universally accepted, model of a cell that we currently have.

One limitation of GO lies in the fact that the ontology structure is constructed by a diverse team of scientists according to their best abilities to curate the published scientific literature. Thus, GO inevitably favors biological entities that have been well studied and misses the large proportion of cell biology that is not yet known or has not yet been curated. As the amount of cell biological literature increases, curating the ontology structure has become a painstaking effort that is proving difficult to scale up (Alterovitz et al., 2010). To address these challenges, we recently investigated whether gene ontologies could be inferred computationally, directly from systematic molecular interaction networks (Dutkowski et al., 2013). In this study, a large fraction of the GO hierarchy was recapitulated de novo directly from network data gathered in budding yeast. For example, the pairwise interaction network for genes and gene products encoding the proteasome (Figure 1A) was transformed to infer the hierarchical structure of proteasomal components to a high degree of accuracy (Figure 1C). In addition, several hundred cellular entities were identified from the data that had not yet been cataloged in GO, pointing to molecular machinery that is novel or has not yet been curated. Data-driven ontologies circumvent some of the problems inherent to human construction in that they provide a systematic view of the cell that is directly reducible to a controlled set of experimental measurements.

An Ontology that Is Dynamic and Predictive

Whether based on expert knowledge or inferred from data, gene ontologies enable representing and reasoning on the structural relationships among biological entities (Myhre et al., 2006; Robinson and Bauer, 2011). However, current gene ontologies are static; they lack any native capacity to capture dynamic biological states or make phenotypic predictions. However, because gene ontologies inherently represent multiscale hierarchy in cellular organization, they provide in theory an ideal substrate for building models that predict a range of cellular responses and phenotypes.

In this respect, Siri provides an excellent example of what a predictive, or “executable,” ontology looks like, supported by recent progress in Artificial Intelligence (AI). At Siri’s core is a series of ontologies containing knowledge that concerns Siri—answers to questions one would usually ask an iPhone (Guzzoni et al., 2006) (Figure 2B). These ontologies cover knowledge on geography and travel, food and recreation, time and scheduling, and so on. For instance, Siri uses an ontology for event planning that treats both meals and movies as types of events, where meals involve a restaurant and a restaurant consists of components such as a name, address, and style of food (Figure 2B). Many artificially intelligent agents other than Siri are also based on ontologies or related structures for knowledge representation and reasoning (Brachman and Levesque, 2004). For instance, IBM’s Watson computer that famously won at *Jeopardy!* in 2011 or Adam the Robot Scientist that successfully characterized new enzymes in yeast (King et al., 2009), both relied profoundly on ontologies for knowledge representation. In many ways, such ontologies are similar in

structure to bio-ontologies such as GO (Figure 2A). Might they also teach us how to develop question-and-answer systems for cell biology (Wren, 2011)?

Unlike gene ontologies, which are essentially descriptive, Siri's ontologies are coupled with dynamic reasoning systems that render them active: "Whereas a conventional ontology is a formal representation of domain knowledge with distinct concepts and relations among concepts, an Active Ontology is a processing formalism where distinct processing elements are arranged according to ontology notions; it is an execution environment" (Guzzoni et al., 2006). These active ontologies not only encode entities and relations, but entities are associated with states and relations are associated with rule sets that perform actions within and among entities. Through a bottom-up execution, input states are incrementally propagated up the hierarchy to impact higher-level entities, whose states are output as the answer to the initial question—the best prediction based on the inputs.

For example, if you ask Siri to "find a good sushi restaurant for two tonight," this query is translated by setting the states of several entities: style is set to "sushi," address to the user's current location, party size to the value "2," and event date to today's date (Figure 2B). These values are propagated through the ontology to generate a list of restaurants, which becomes the state of the event entity. This event result can then be provided to the user or included in further computations. It is precisely because the system can propagate such information, guided by structural and functional relationships between entities, that we consider Siri a muse for cellular modeling.

Toward a Siri of the Cell

How the ontologies within Siri are used to answer questions is very different from how GO is used today in bioinformatics. Typically, GO terms are associated with a set of genes (annotations), but not with dynamic states; the relationships between GO terms are not associated with rule sets that perform actions, at least beyond propagation of gene set annotations. Nonetheless, given the similarity of GO to Siri and other AI agents (Figures 2A and 2B), we propose that it might be possible to assemble an intelligent system for cell biology according to the following general guidelines. Clearly, these guidelines are merely suggestive and will require much research to determine their feasibility and best implementations.

First, the structure of such an ontology could be directly assembled from GO or algorithmically inferred directly from systematic data sets (Figure 2C). Based on our experience in building data-driven ontologies, we suspect that a good first draft of this ontology structure might be achieved by clustering genome-wide data incorporating protein interactions, conditional gene expression, coevolution, and known phenotypic impact of genetic manipulations such as deletions or knockdowns. Given an ontology structure that reflects a hierarchical organization of the cell, a next step would be to associate each entity with a state. A state would naturally correspond to the phenotype or collection of phenotypes that most directly describe the entity, with all phenotypes being directly measurable in the laboratory. Entities near the top of the hierarchy would be associated with whole-cell phenotypes such as growth or differentiation. Lower-level entities would correspond to

phenotypes that are increasingly molecular and concern the action of fewer and fewer genes. High-level phenotypes, low-level phenotypes, and perhaps even genotypes would become interrelated in the continuum of a multiscale hierarchical model of the cell. An important consideration is how to achieve an appropriate mapping between entities and phenotypes. This might involve pruning the ontology of entities for which one cannot conceive of an experimental measurement of state.

Another challenge will be to determine how to dynamically compute the state of each entity based on the states of its neighbors. For example, the state of "DNA replication elongation" would be computed from information about the elongation of both the leading and the lagging strand (Figure 2A), but the underlying mathematical function could take many forms, including logic gates, probabilistic functions, or polynomial or logistic equations. How to determine the specific parameters of these functions, regardless of what form they take, is also unclear. This step could happen by statistical association from many input-output examples using machine learning methods, by including externally generated biological knowledge specific to each entity, or by manual curation from literature.

Siri of the Cell: Hopes and Limitations

Recent applications of AI to the life sciences have already laid some of the groundwork for this vision. In particular, question-answering systems for biology and medicine are already starting to emerge. These include *Inquire Biology*, a project to create an iPad app to enable students to query biology textbooks (Spaulding et al., 2011), or the *LODQA Project* (*Linked Open Data Question Answering*), which attempts to convert natural language questions about disease genes to complex queries directed at databases (Cohen and Kim, 2013). It is quite conceivable that expanding such tools with an active ontology representation of the cell may take us beyond knowledge retrieval and open the door to novel biological predictions.

One way Siri of the cell might be executed is shown in Figure 2C. A first step translates a cell biological question, asked in natural language, to the corresponding input/output relationship implied by the speaker. For example, a question such as "What is the impact on cell growth of knockdown of DNA polymerase epsilon?" might be interpreted to recognize that the input is decreased expression of the gene encoding DNA polymerase epsilon (Polε), and the desired output concerns cell growth. The second step is to map the inputs and outputs to entities in the ontology and to set the state of the input entities accordingly. In our example, Polε is a component of the entity "DNA leading strand elongation"; therefore, the state of this entity would be set to reflect the decreased expression of Polε. In a third step, the ontology is executed to update all of its states, and the phenotypic prediction of interest is read from the state of the output entity(ies). Here, the perturbation of Polε expression would propagate upward in the ontology to impact higher-order entities such as "DNA replication" and "DNA metabolism." Ultimately, the state of the entity corresponding to cell growth would be updated, and a prediction would be made. The validity of this prediction could then be experimentally tested. Beyond this example, one can envision how the suggested implementation of Siri of the cell might accommodate a wide range of queries (Wren, 2011).

The ontology is a very general framework, but whether it can capture the full extent of biological complexity remains to be seen. Certain environmental conditions that do not readily correspond to ontology inputs or outputs may prove difficult to represent. How to combine or extrapolate cell models to the scale of an organism (such as a human patient) is a challenge that must be eventually met. Moreover, because the entities and entity relations of an ontology are typically fixed, such a model is incompatible with evolution, which depends on plasticity in biological structure.

Nonetheless, even if a predictive ontology eventually fails as the ultimate model of a cell, we anticipate that such a system could still perform well at answering general biological questions and predicting cellular phenotypes. This duality between predictive capacity and accurate representation of reality is a major issue for all AI agents. AI has not fully elucidated the human mind, but it certainly has led to many intelligent predictions and agents like Siri. Can such agents ever capture the inner workings of the mind or, for that matter, of a cell? Perhaps. Can they help us find a good sushi restaurant for two tonight? Absolutely.

Acknowledgments

We gratefully acknowledge Janusz Dutkowski, Michael Yu, Michael Kramer, Hannah Carter, Salvatore Loguercio, Benjamin Good, Bing Ren, Andrew Su, Gary Siuzdak, Benjamin Kellman, Heidi Kayser, and our anonymous reviewers for important discussions and comments in preparation for this Essay. This work was supported by the National Institutes of General Medical Sciences grants P41 GM103504 and P50 GM085764.

References

Alterovitz G, Xiang M, Hill DP, Lomax J, Liu J, Cherkassky M, Dreyfuss J, Mungall C, Harris MA, Dolan ME, et al. Ontology engineering. Nat Biotechnol. 2010; 28:128–130. [PubMed: 20139945]

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000; 25:25–29. [PubMed: 10802651]

Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004; 5:101–113. [PubMed: 14735121]

Berner, ES. Clinical Decision Support Systems: Theory and Practice. 2. New York: Springer; 2007.

Brachman, RJ.; Levesque, HJ. Knowledge Representation and Reasoning. Amsterdam: Morgan Kaufmann; 2004.

Bray, D. Wetware: A Computer in Every Living Cell. New Haven: Yale University Press; 2009.

Cohen, KB.; Kim, J-D. Evaluation of sparql query generation from natural language questions. Proceedings of the Joint Workshop on NLP&LOD and SWAIE; 2013. p. 3-7.

Chuang HY, Hofree M, Ideker T. A decade of systems biology. Annu Rev Cell Dev Biol. 2010; 26:721–744. [PubMed: 20604711]

Dutkowski J, Kramer M, Surma MA, Balakrishnan R, Cherry JM, Krogan NJ, Ideker T. A gene ontology inferred from molecular networks. Nat Biotechnol. 2013; 31:38–45. [PubMed: 23242164]

Fortunato S. Community detection in graphs. Phys Rep. 2010; 486:75–174.

Fung, YC. Biomechanics: Mechanical Properties of Living Tissues. 2. New York: Springer-Verlag; 1993.

Gruber TR. Toward principles for the design of ontologies used for knowledge sharing. Int J Hum Comput Stud. 1995; 43:907–928.

Guzzoni, D.; Baur, C.; Cheyer, A. Active: A unified platform for building intelligent web interaction assistants. 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology; 2006. p. 417-420.

Hooke, R. With Observations and Inquiries Thereupon. London: Printed by J. Martyn and J. Allestry; 1665. Micrographia: Or, Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses.

Kauffman, SA. The origins of Order: Self-Organization and Selection in Evolution. New York: Oxford University Press; 1993.

King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, Liakata M, Markham M, Pir P, Soldatova LN, et al. The automation of science. Science. 2009; 324:85–89. [PubMed: 19342587]

Koller, D.; Friedman, N. Probabilistic Graphical Models: Principles and Techniques. Cambridge, MA: MIT Press; 2009.

Mathews CK. The cell-bag of enzymes or network of channels? J Bacteriol. 1993; 175:6377–6381. [PubMed: 8407814]

McPherson, A. Introduction to Macromolecular Crystallography. 2. Hoboken, NJ: Wiley-Blackwell; 2009.

Monod J, Changeux JP, Jacob F. Allosteric proteins and cellular control systems. J Mol Biol. 1963; 6:306–329. [PubMed: 13936070]

Myhre S, Tveit H, Mollestad T, Laegreid A. Additional gene ontology structure for improved biological reasoning. Bioinformatics. 2006; 22:2020–2027. [PubMed: 16787968]

Pollack, GH. Cells, Gels and the Engines of Life: A New, Unifying Approach to Cell Function. Seattle, WA: Ebner & Sons; 2001.

Reddy GP, Singh A, Stafford ME, Mathews CK. Enzyme associations in T4 phage DNA precursor synthesis. Proc Natl Acad Sci USA. 1977; 74:3152–3156. [PubMed: 198773]

Robinson, PN.; Bauer, S. Introduction to Bio-ontologies. Boca Raton: Taylor & Francis; 2011.

Spaulding, A.; Overholtzer, A.; Pacheco, J.; Tien, J.; Chaudhri, VK.; Gunning, D.; Clark, P. Inquire for ipad: A biology textbook that answers questions. In: Biswas, G.; Bull, S.; Kay, J.; Mitrovic, A., editors. Artificial Intelligence in Education. New York: Springer; 2011. p. 627-627.

Thompson, DAW. On Growth and Form. Cambridge: University Press; 1917.

Walhout, AJM.; Vidal, M.; Dekker, J. Handbook of Systems Biology: Concepts and Insights. 1. London: Waltham Academic Press; 2013.

Wren JD. Question answering systems in biology and medicine—the time is now. Bioinformatics. 2011; 27:2025–2026. [PubMed: 21672971]

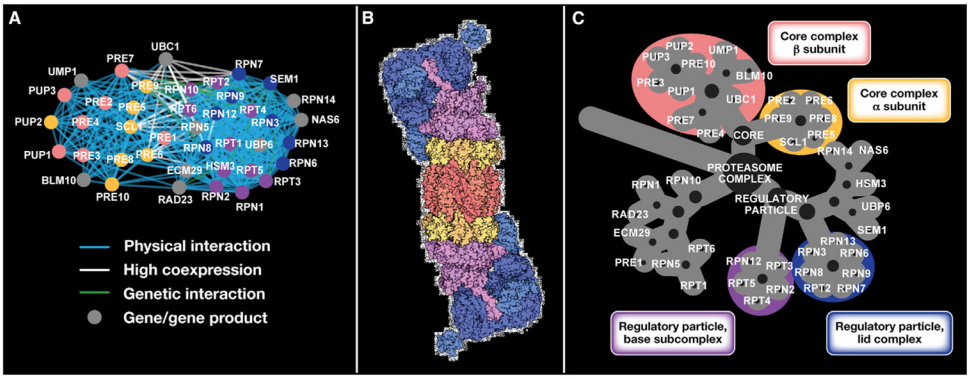

Figure 1. From Networks to Ontologies

(A) Network representation of three types of molecular interactions among genes/gene products that form the proteasome structure, displayed using a force directed layout in Cytoscape (http://www.cytoscape.org).

(B) Cartoon representation of the structure of the proteasome (Protein Data Bank entry 4b4t), created by integrating partial crystallographic structures obtained by analysis of 2.4 million images from electron microscopy.

(C) Hierarchical factorization of the proteasome subcomponents as described by NeXO. Across all panels, colors indicate membership to the core complex β subunit (red), core complex α subunit (orange), regulatory particle lid complex (blue), and regulatory particle base complex (purple) according to the GO (A), the Protein Data Bank (B), and NeXO (C). (A) and (C) adapted with permission from Dutkowski et al. (2013). (B) courtesy of David S. Goodsell and the Research Collaboratory for Structural Bioinformatics Protein Data Bank.

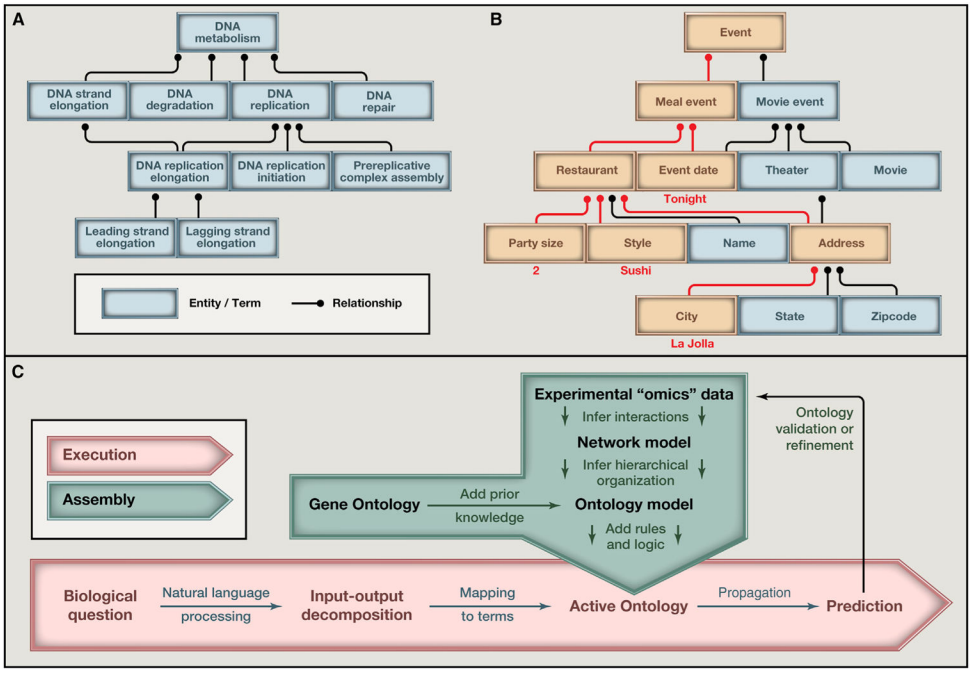

Figure 2. From Ontologies to Active Ontologies

(A) A subset of the Gene Ontology (Ashburner et al., 2000).

(B) A subset of an active ontology for event planning (Guzzoni et al., 2006). Red relationships and entities indicate dynamic computation.

(C) One possible roadmap toward the assembly and execution of Siri of the cell.