Reconciling Function and Structure in Scientific Models

Scott Friedman, Mark Burstein, David McDonald
SIFT, Minneapolis, MN, USA
{friedman, burstein, dmcdonald}@sift.net

James Pustejovsky, Peter Anick
Brandeis University, Waltham, MA, USA.
{jamesp, panick}@cs.brandeis.edu

Rusty Bobrow
Bobrow Computational Intelligence, LLC
rjbobrow@gmail.com

Brent Cochran
Tufts University School of Medicine, Boston, MA, USA
brent.cochran@tufts.edu

Abstract

Despite our increasing understanding of the structure and dynamics of scientific domains, functional knowledge and functional language— such as referring to a central purpose or function of a molecule— permeate scientific articles. Cognitive systems that collaborate with scientists must therefore represent functional knowledge to support machine reading and explanation. This paper describes our progress on automatically inferring and representing functional knowledge in R3 (*Reading, Reasoning, and Reporting*). R3 automatically reads biology articles from PubMed Central, using a massive domain model from Pathway Commons (www.pathwaycommons.org/) as background knowledge. R3 now relates functional language to its background structural model and explains functional knowledge, which is the central contribution of this paper. We motivate the representation of functional knowledge in the biology domain— which many existing ontologies omit— using examples from PubMed articles. We then describe how R3 automatically adds functional knowledge to its model by parsing textual summaries of biological processes and extracting semantics. We then describe how R3 builds event structures and compositional models with functional knowledge, and we illustrate how R3 uses its functional knowledge to diagram protein activity from the information it learned from reading.

Introduction

The concepts and factors we use to model scientific domains for our intelligent systems are often incommensurable with the concepts and factors we use to communicate scientific findings to our human peers. This is for a good reason, since intelligent systems and humans often serve complementary roles in the scientific process: machines engage in parallel search and discovery over vast structural models and networks of entities, while people frequently learn and communicate salient forms of entities with functional or intentional language. For instance, biologists often describe proteins and other natural kinds with functional contextual descriptors such as "active" and "inactive," and they compactly refer to the "activity" of an entity as its central function within a complex system.

Biologists often use artifactual mental models— such as *molecular switches*— to describe and reason about proteins. The molecular switch metaphor explicitly describes natural kinds (i.e., proteins) as *artifacts* (i.e., on/off switches), rather than just describing the behavior or capability of the natural kinds. For instance, this sentence from Akinleye et al. (2013) describes the proteins of the Ras family as molecular switches that are *inactive* (i.e., functionally *off*) when bound to GDP and *active* (i.e., *on*) when bound to GTP:

> "H-Ras, K-Ras, and N-Ras function as molecular switches when an inactive Ras-GDP is converted into an active Ras-GTP."

This relates a structural change (i.e., GDP/GTP binding) to a contextual function: when bound to GTP, Ras is able to perform its agreed-upon function (as opposed to many other reactions that Ras engages in) within a specific cell signaling pathway.

Intelligent systems that learn by reading must bridge this structural-functional gap: given only the *structural* knowledge from an ontology describing a complex system, a system can not resolve references to "active" or "inactive" entities that collectively "contribute to" some macro behavior, nor can it resolve references to the "activity" or functional capabilities of an entity. Our *Reading, Reasoning, and Reporting* (R3) system, developed as part of DARPA's Big Mechanism program (Cohen, 2015), reads articles in molecular biology to extend and revise its structural and functional models of biological mechanisms (Friedman et al., submitted, McDonald et al., 2016).

Our recent extensions to R3 aim to *automatically* bridge the structural-functional gap. This involves extending traditional compositional modeling semantics (Falkenhainer and Forbus, 1991) to support event structure (Pustejovsky, 1991b) and *telic qualia* (i.e., functional descriptions) from Generative Lexicon (GL) theory (Pustejovsky, 1991a).

Extending the modeling semantics gives R3 the representational capabilities, but it also needs the *content* to construct and populate these models. Fortunately, the model we are extending is annotated with English summaries of the individual reactions involved, written by human experts. R3 is thus able to extend the domain model to include critical functional information by automatically reading these textual summaries embedded within the model. R3 extracts the functional semantics from the summaries and automatically extends its model by adding functional characterizations of the reactions. It automatically identifies (1) events that comprise the entity's function (2) structural preconditions for

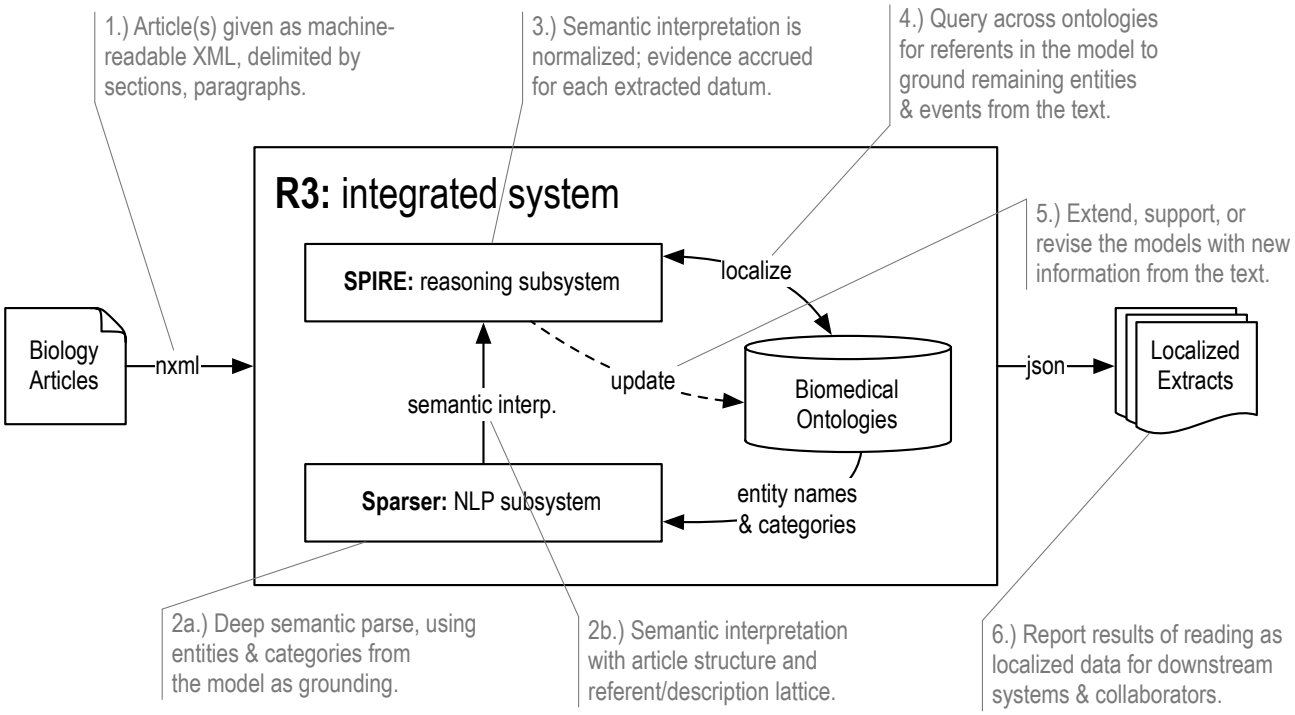

Figure 1: The R3 architecture, and the flow of information by which R3 reads articles, updates its mechanism models, and publishes extracted knowledge for human and machine collaborators.

functional status, and (3) precursor events that enable functional status.

R3's resulting models relate the structure of entities to the functional status of entities, as well as the events that comprise their function. This allows R3 to marshal functional knowledge into context when reading about structure, and also, to read-between-the-lines and ground functional references.

We begin by describing how R3 extracts and recognizes biological events and interactions from text. We then discuss R3's use of traditional qualitative modeling semantics with GL semantics in order to describe constituent entities, event structure, and functional knowledge. We discuss R3's approach to automatically extending its domain model with functional knowledge, and we present R3's automatically-generated results, including event structures and compositional models that accurately describe the well-known functions of the corresponding entities. We review related knowledge representation work for expressing the function of systems and devices, and distinguish our linguistically-motivated knowledge representation. We close with a discussion of the implications and future work for R3.

Deep Parsing in Molecular Biology

Biomedical research articles are written to be read by other professional biologists who are presumed to have the requisite technical background. Building a system such as R3 that can read-with-a-model in the biomedical domain poses key research challenges for the general task and the specific domain:

- Texts frequently use the same word to mean different types of objects in the model. *"RAS"* can refer to a protein, a gene, or a larger multi-protein complex, within a single article.
- Texts may describe things at different levels of abstraction than the model. For example, authors frequently talk about the the *function* of events while a purely mechanistic model may only describe the biochemical reactions taking place.
- One process or event may be part of many other processes or events in the domain model.

To address these challenges, R3 integrates deep semantic parsing, ontology mapping, and reasoning about structure, function, and mechanism-level causality, as shown in Figure 1.

Deep parsing allows R3 to extract precise semantics and determine entity types from local lexical context. R3 uses the SPARSER (McDonald, 1996) rule-based, type-driven semantic parser to read the texts. SPARSER'S rules succeed only if the types of the constituents to be composed satisfy the type constraints (value restrictions) specified by the rule. SPARSER compiles a semantic grammar from a semantic model of the information to be analyzed and a specification of all the ways each of the concepts can be realized in the language of the genre, such as biomedical research articles. This ensures that everything SPARSER is able to parse it can

Datum	Model Fragment	Generative Lexicon
Category of the resulting instance.	Type	Formal role
Instances & types of constituents/parts.	Participants	Argument structure
Existence criteria over participants.	Constraints	Constitutive role
Applicability of the model to the class of problem.	Modeling assumptions	N/A
Criteria for asserting behavior.	Conditions	Telic role(s) LHS
Behavior of the object/relation/event.	Consequence	Telic role(s) RHS
Events and state changes over explicit time.	Encapsulated histories*	Event structure
Function or purpose of the object/relation.	N/A	Telic role(s)

Figure 2: High-level mapping of qualitative model fragment semantics against Generative Lexicon semantics. [*] Encapsulated histories (Forbus, 1984) are defined *outside* of model fragments, but like GL event structure, encapsulated histories make explicit reference to time and state.

model, and that every rule in the compiled grammar has an interpretation.

R3 performs ontology-mapping with the SPIRE reasoner to map from SPARSER'S ontology into the BioPAX ontology of the domain model. This allows R3 to perform structure-mapping between semantic parses and the domain model in order to transfer knowledge from the parse into the model. This is crucial for learning by reading, as we discuss below.

Modeling with Generative Lexicon

Generative Lexicon (GL) theory assumes that word meaning is structured according to four generative *qualia roles* that describe how people understand entities and events and relations in the world. In GL, *lexemes* are words, word roots or phrases and their variants. They refer to entities, events, or relations, together with their associated semantics represented by *argument structure* and *qualia*, and organized by *habitats*.

There are four qualia roles of any lexeme:

- *Formal*: the basic ontological category.
- *Constitutive*: the relationships among constituent parts.
- *Telic*: its purpose or function.
- *Agentive*: how it came into being.

In addition to qualia roles, GL *argument structure* describes the constituents (i.e., the primitive entities, other lexemes, or sets thereof) that jointly participate in the lexeme, and their role within the lexeme. The GL *habitat* of an entity is a partial minimal model that enhances its qualia structure (Pustejovsky, 2013), and describes the *event structure* (i.e., events and sub-events) with reference to time. We provide example lexemes below to illustrate all of these GL concepts in the biology domain.

GL theory semantics supports the basic model fragment semantics for compositional modeling (Falkenhainer and Forbus, 1991, Rickel and Porter, 1997) as shown in Figure 2. The Figure 2 mapping has gaps: GL does not have an analog for meta-level modeling assumptions, so it cannot natively specify problem-level information such as applicable levels of granularity. We overload the GL telic role to describe both behavior (i.e., model fragment conditions and consequences) and function (i.e., the teleology or purpose of the lexeme). These gaps exist for good reason, since the two representations aim to solve different problems, but as shown in Figure 2, composition is a central capability of both representations.

Our lexeme representation violates the *no-function-in-structure* principle (de Kleer and Brown, 1981), which states that the rules for specifying the behavior of any constituent part of a system can in no way refer— even implicitly— to how the overall system functions. These lexeme-based models violate the no-function-in-structure principle by explicitly representing the function of proteins (through telic roles and habitats) within the larger cell signaling pathway.

GL theory distinguishes between *artifacts* and *natural kinds* via the telic role: artifacts have a telic role to express their function, whereas natural kinds have no inherent function, and therefore have no telic role. Biology articles do not adhere to this distinction: biologists frequently refer to a protein's *"function"* or *"activity,"* which effectively ascribes a purpose to a natural kind. We therefore model proteins as *artifactual* types with telic roles.

Artifactual models of natural kinds allow us to represent the widely-used artifactual "molecular switch" model of proteins within a larger pathway. Two such molecular switch lexemes induced by R3 are shown below, one for RAS, and one for MAPK. Both are summarized from the original R3 output for simplicity.

The RAS and MAPK molecular switches are defaulted to *off*, but as described in their telic roles, when events such as GTP-binding or phosphorylation occur, the molecules enter an *on* state. In the RAS lexeme, the *on* state enables RAS to function as a reactant in the activation of RAF, as a catalyst in the activations of MAP2K and MAPK, and as a catalyst in its own deactivation.

$$\left[\begin{array}{l}\textbf{H-RAS(x)} \\ \text{QUALIA} = \left[\begin{array}{l}\text{FORMAL} = \textbf{Protein(x):}id = P01111 \\ \text{TELIC} = \mathcal{C} \rightarrow [\textbf{bind}(x, \textbf{GTP})]On(x)\end{array}\right] \\ \text{HABITAT} = \left[\begin{array}{l}\text{ON(X)} = reactant(x, act(\textbf{RAF})) \\ \text{ON(X)} = reg_{+}(x, act(\textbf{MAP2K})) \\ \text{ON(X)} = reg_{+}(x, act(\textbf{MAPK})) \\ \text{ON(X)} = reg_{+}(x, deact(x))\end{array}\right]\end{array}\right]$$

In the MAPK lexeme, activation occurs when the MAPK1 dimer is phosphorylated at Threonine and Tyrosine sites (not represented below), which enables its function of (1) positively regulating the deactivations of RAF and MAP2K and (2) translocating from the cytosol to the nucleoplasm.

$$\left[\begin{array}{l}\textbf{MAPK1\_dimer\_cytosol(x)} \\ \text{ARGSTR} = \left[\begin{array}{l}\text{SELF} = x\textbf{:Complex} \\ \text{COMPONENT} = y\textbf{:Protein:}id = P28482 \\ \text{COMPONENT} = z\textbf{:Protein:}id = P28482\end{array}\right] \\ \text{QUALIA} = \left[\begin{array}{l}\text{FORMAL} = \textbf{Complex(x)} \\ \text{CONST} = [comp(x, y), comp(x, z)] \\ \text{TELIC} = \mathcal{C} \rightarrow [\textbf{phos}(\{y, z\})]On(x)\end{array}\right] \\ \text{HABITAT} = \left[\begin{array}{l}\text{OFF(X)} = loc(x, \textbf{cytosol}) \\ \text{ON(X)} = reg_{+}(x, deact(\textbf{MAP2K})) \\ \text{ON(X)} = reg_{+}(x, deact(\textbf{RAF})) \\ \text{ON(X)} = reactant(x, deact(x)) \\ \text{ON(X)} = [\textbf{move}(x)]loc(x, \textbf{nucleoplasm})\end{array}\right]\end{array}\right]$$

These lexemes bridge the gap between structural status of entities (e.g., the phosphorylation status, molecular subcomponents, and molecule bindings) and the salient functional capabilities of entities within a larger system (e.g., their ability to translocate and activate other entities). With these lexemes available to its reading operations, R3 can utilize textual references to *"active RAS"* and *"RAS function"* to marshal important background knowledge about the structure of RAS when in the functionally active state, and the events that constitute RAS' function when active, respectively.

The events described in the habitats of the above autogenerated lexemes comprise small fractions of the RAS and MAPK events in R3's initial domain model; these are the events that R3 ascribed only to the active forms (i.e., *on* state) of these proteins.

As we describe below, R3 has no *a priori* knowledge of what constitutes the "active" functional form(s) (i.e., *on* state) of any single molecule, since its initial BioPAX model contains no information about "active" forms. Furthermore, activation is associated with different molecular configurations for different proteins, so active states cannot generally be inferred directly from chemical makeup. Consequently, R3 reads texts to extend its domain model with this functional knowledge, using human characterizations of activity when specifically attributed to different proteins in the model. We next describe how R3 learns by reading, and we present results of R3 using its learned functional knowledge to generate diagrams of protein function.

Learning-by-Reading Experiment

Here we describe R3's approach to automatically mining a large, existing biology model to generate lexical knowledge and event models that can be used to represent and reason about protein function. R3 parses textual data embedded within the model, including experts' summaries and reaction/molecule display names, into semantic interpretations. R3 extends the model with novel knowledge from these interpretations about the *function* of the described proteins and complexes, since the model is initially describes only structures, locations, and reactions. R3 propagates and analyzes these functional labels— such as whether a protein is active or inactive— in order to characterize the protein's function with the following dimensions:

- Structural conditions (e.g., phosphorylation status), location conditions (i.e., where within the cell), and binding conditions (e.g., in complex with another molecule) for protein activity.
- Event precursors that enable or disable the above conditions. These are the protein's activation and deactivation events, respectively.
- Events that depend on the active form of the protein and *not* the inactive form, such as reactions where the active protein is a catalyst or a scaffold. These events— which often activate or deactivate other proteins— comprise the protein's function within the signaling pathway.

We continue with a description of the dataset and a summary of R3's reading operations, whereby R3 extends its model with functional knowledge to describe active and inactive forms of proteins. We then present the results of its functional knowledge mining and lexical KB population using these labels.

Dataset

For this experiment, we used the entire "Signaling by EGFR" subset of the open-source, peer-curated Reactome pathway database.[1] Reactome pathway models describe reactions, reactants (i.e., complexes, proteins, and other molecules), catalysis and regulation relations, and protein modifications (e.g., phosphorylation, ubiquitination). These are downloadable as BioPAX (Demir et al., 2010), an RDF/OWL-based standard for describing pathways and the molecules and reactions that comprise them. The "Signaling by EGFR" Reactome subset contains 128 biochemical reactions and 911 molecules (i.e., proteins, complexes, small molecules, and other physical objects).

At the time of writing, BioPAX (Level 3) lacks categories and relations to describe the functional activity of a given protein with respect to a pathway, e.g., when it is active or inactive *within the context of the pathway*. We extended BioPAX to add categories and relations to represent these functional forms.

Machine Reading

R3 parsed summaries and display-names (i.e., descriptive labels) of 17 reactions that refer to molecules as "active," "inactive," "stimulated," or "activated," or that refer to the "activation" of a protein or describe how a protein "activates" another. Consider this example:

[1]The BioPAX OWL files are downloadable via the pathway browser: http://www.reactome.org/PathwayBrowser/

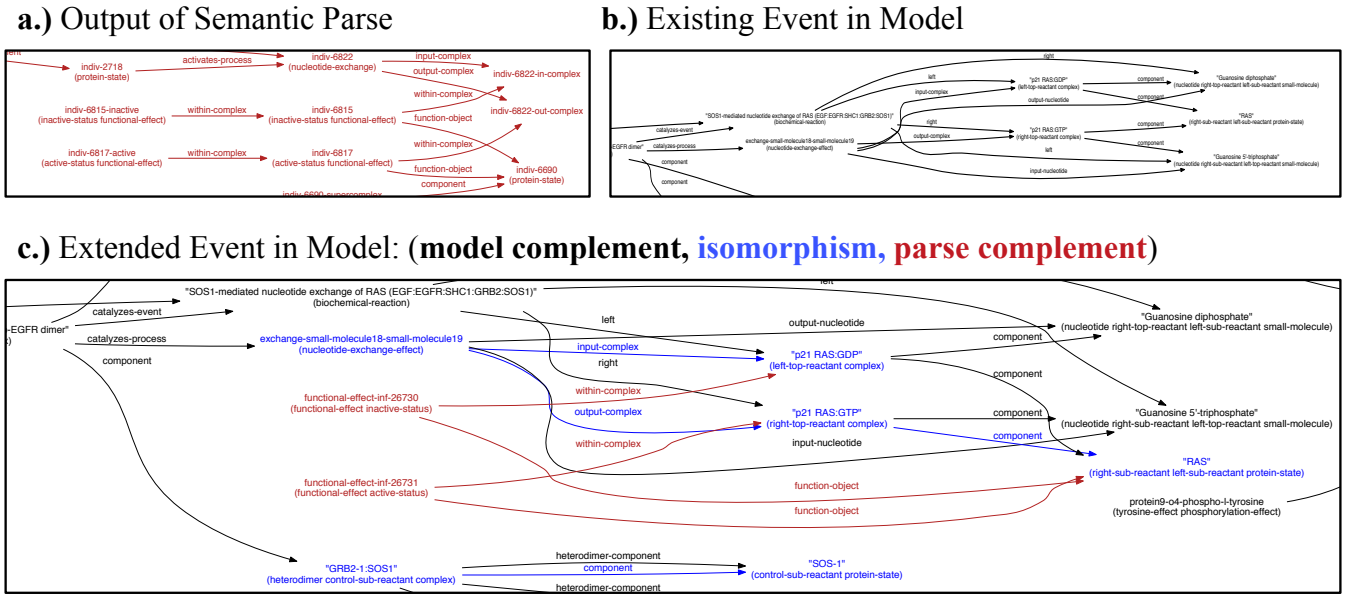

Figure 3: R3 extends its model with functional knowledge by reading: (a) its semantic parser interprets an expert's summary into a semantic graph; (b) it builds the semantic graph of the corresponding reaction; and finally (c) it computes the maximum common subgraph between the two, projecting the complement of the interpretation semantics into the model.

> "SOS1 is the guanine nucleotide exchange factor (GEF) for RAS. SOS1 activates RAS nucleotide exchange from the inactive form (bound to GDP) to an active form (bound to GTP)."

R3 parses this summary to produce a semantic interpretation graph, a portion of which is displayed in Figure 3(a). It then uses the BioPAX semantic graph of the reaction, shown in Figure 3(b), to match against the interpretation graph, using a structure-mapping algorithm (Falkenhainer, Forbus, and Gentner, 1989) that supports additional constraints for identicality-matching. This produces the mapping shown in Figure 3(c). Here we illustrate that some entities and relations of the biology model (shown in blue) are isomorphic to the entities and relations of the semantic interpretation, despite using different symbols to describe entities. Provided the isomorphic subgraph, R3 projects a portion of the interpretation into the model: when the protein RAS is bound to GTP (to form the complex "p21 RAS:GTP"), it has an *active-status*, and when it is bound to GDP (to form the complex "p21 RAS:GDP") it has an *inactive-status*. R3 propagates this inference to all relevant super-complexes and reactions that contain these forms of RAS. R3 performs these steps for every summary and display-name it reads.

Populating Lexemes & Event Models

After identifying all of the active and inactive forms of proteins referenced in the summaries of the "Signaling by EGFR" subset of Reactome, R3 analyzes each active protein to generate lexemes.[2] This involves identifying structural preconditions, location preconditions, and molecular binding preconditions for *"active"* status, as shown in the RAS and MAPK lexemes listed above. In total, R3 generated 15 lexemes to describe active variants of RAS (3 lexemes), RAF (4 lexemes), MAP2K (3 lexemes), and MAPK (4 lexemes). There are multiple lexemes of each protein, since within the EGFR signaling subset of Reactome, *"active RAS"* refers to GTP-bound HRAS, KRAS, and NRAS of the RAS family. Similarly, *"active MAP2K"* can refer to a phosphorylated homodimer of MAP2K1, a phosphorylated homodimer of MAP2K2, or a phosphorylated MAP2K1/MAP2K2 heterodimer.

After identifying active components and building lexemes, R3 uses the functional knowledge in its lexemes to build an event graph of protein activation and protein function, as shown in Figure 4.

The event structure describes the active and inactive forms of molecules across cellular locations, as well as the biochemical-reactions ("R" nodes) that activate, deactivate, and translocate them. The triangular arrowheads indicate input and output reactants to the reactions, and the circular arrowheads indicate direct regulatory relationships— such as catalysis— between entities and reactions.

This automatically-generated event structure comprises a very small subset of R3's original BioPAX model, and it closely resembles the well-known RAS-RAF-MAP2K-MAPK activation cascade.

[2]Some proteins, such as SOS-1, are not referenced as being "active" in any textual Reactome summary, despite having an "active" form in the wider literature, so R3 does not have functional knowledge about these proteins.

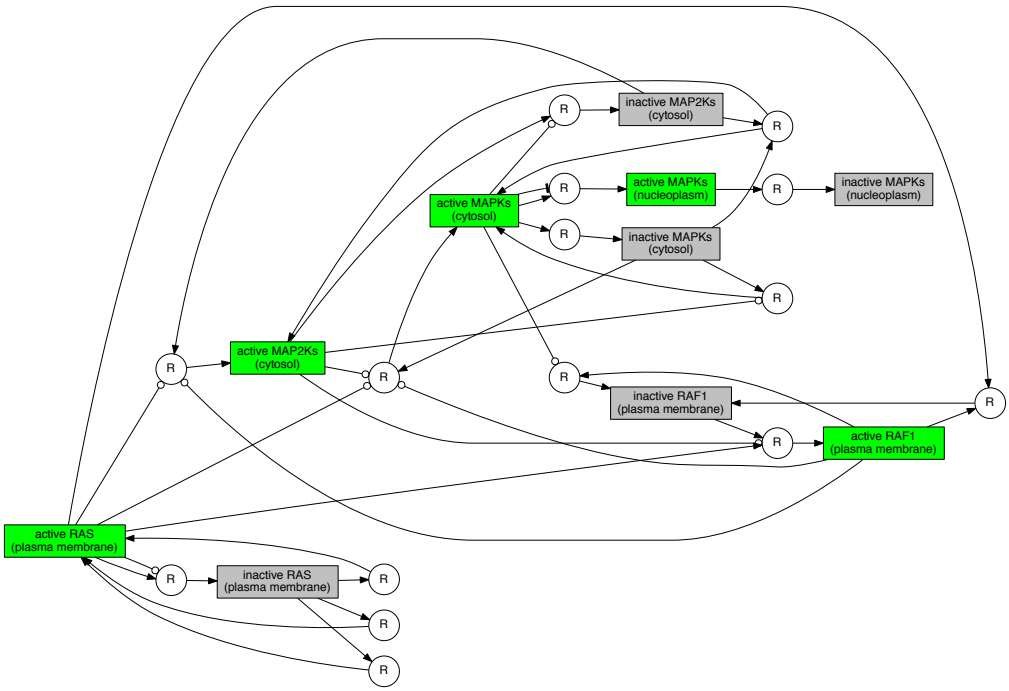

Figure 4: R3 automatically infers and displays the interrelated event structures of RAS, RAF1, MAP2K (MEK), and MAPK (ERK), including activations, deactivations, and translocations of the entities.

Related Work

Other work in philosphy and biology has discussed biological function, dysfunction, and malfunction (Krohs and Kroes, 2009) and identified different notions of function, including activity, role, biological advantage, and effect (Wouters, 2003). This work on R3 evokes the notion of function as biological activity, which we achieve by modeling natural kinds as artifacts.

Other research has used compositional modeling methods in biological domains (e.g., Mallavarapu et al., 2009, Rickel and Porter, 1997) to represent structure and behavior. These compositional modeling methods are useful for reasoning with the output of R3, which automatically produces scientific model components from text.

In engineering, functional knowledge has been used in simulation, diagnosis, design, and other tasks for decades (e.g., Chandrasekaran, 1994, Freeman and Newell, 1971, Goel, 2013). Knowledge-based systems use *functional representation* (FR) languages for describing the function of systems or components and the structural and causal processes that achieve the function. Previous approaches generally focus on encoding the *intent* of devices— in a top-down FR that complements traditional qualitative modeling— and then annotating causal transitions to explicate how the device achieves the intent (Chandrasekaran, 1994). Conceptual design systems also utilize functional knowledge— such as structure-behavior-function (SBF) models— by taking FR specifications and producing a structural specification that achieves the desired function (Goel, Rugaber, and Vattam, 2009, Goel, 2013).

Like other FRs, R3's lexeme-based approach explicitly describes the relation of structure to function; however, since R3 is a learning-by-reading system, its functional representation uses linguistically— and cognitively— motivated representations (Pustejovsky, 1991a,b). This will allow R3 to retrieve lexemes at parse-time and marshal background knowledge, e.g., about structure and function, into its semantic interpretation.

Conclusion & Future Work

This paper outlined our approach to reconciling structural and functional knowledge in the biology domain. We described extensions to traditional qualitative modeling semantics by inforporating telic roles and habitats from GL theory to encode functional knowledge. This enabled us to model the popular "molecular switch" artifactual perspective of natural kinds in our R3 system. We presented results of R3 learning functional knowledge by reading in order to to populate functional models, and then we demonstrated that R3 can use its functional knowledge to display protein activity (i.e., the function of a protein when the molecular switch is *"on"*).

This paper does not demonstrate R3's using its learned functional knowledge *while it reads*; this is a primary focus of our present work on R3. For instance, if R3 encounters a mention of "active Ras-GTP" in an article (e.g., Akinleye et al., 2013), it will use the lexemes induced in this work to marshal relevant functions of active Ras from the lexeme's habitat, including the phosphorylation— and activation— of Raf. When it reads the next sentence of the same article,

stating that "Ras recruits and activates Raf kinases," the corroborating background knowledge about the event and processes is already present in R3's reasoning context. These lexemes thereby support R3's overall objective of reading-with-a-model.

Acknowledgments

This work was supported in part by the DARPA "Big Mechanism" program, BAA 14-14, under ARO contract W911NF-14-C-0109. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the position or policy of the Government.

References

Akinleye, A.; Furqan, M.; Mukhi, N.; Ravella, P.; and Liu, D. 2013. Mek and the inhibitors: from bench to bedside. *J Hematol Oncol* 6(1):27.

Chandrasekaran, B. 1994. Functional representation: a brief historical perspective. *Applied Artificial Intelligence an International Journal* 8(2):173–197.

Cohen, P. R. 2015. Darpa's big mechanism program. *Physical Biology* 12.

de Kleer, J., and Brown, J. S. 1981. Mental models of physical mechanisms and their acquisition. *Cognitive skills and their acquisition* 285–309.

Demir, E.; Cary, M. P.; Paley, S.; Fukuda, K.; Lemer, C.; Vastrik, I.; Wu, G.; D'Eustachio, P.; Schaefer, C.; Luciano, J.; et al. 2010. The biopax community standard for pathway data sharing. *Nature biotechnology* 28(9):935–942.

Falkenhainer, B., and Forbus, K. D. 1991. Compositional modeling: finding the right model for the job. *Artificial intelligence* 51(1):95–143.

Falkenhainer, B.; Forbus, K. D.; and Gentner, D. 1989. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence* 41(1):1 – 63.

Forbus, K. D. 1984. Qualitative process theory. *Artificial intelligence* 24(1):85–168.

Freeman, P., and Newell, A. 1971. A model for functional reasoning in design. In *Proc. 2nd Int. Joint Conf. Artificial Intelligence*, 621–633.

Friedman, S.; Burstein, M.; McDonald, D.; Paullada, A.; Plotnick, A.; Bobrow, R.; Cochran, B.; Pustejovsky, J.; and Anick, P. submitted. Learning by reading: Extending & localizing against a model. In *Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems*.

Goel, A. K.; Rugaber, S.; and Vattam, S. 2009. Structure, behavior, and function of complex systems: The structure, behavior, and function modeling language. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 23(01):23–35.

Goel, A. K. 2013. A 30-year case study and 15 principles: implications of an artificial intelligence methodology for functional modeling. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 27(03):203–215.

Krohs, U., and Kroes, P. 2009. *Functions in biological and artificial worlds: comparative philosophical perspectives*. Mit Press.

Mallavarapu, A.; Thomson, M.; Ullian, B.; and Gunawardena, J. 2009. Programming with models: modularity and abstraction provide powerful capabilities for systems biology. *Journal of The Royal Society Interface* 6(32):257–270.

McDonald, D.; Friedman, S.; Paullada, A.; Bobrow, R.; and Burstein, M. 2016. Extending biology models with deep nlp over scientific articles. In *Proceedings of the AAAI-16 Workshop on Knowledge Extraction from Text*.

McDonald, D. D. 1996. The interplay of syntactic and semantic node labels in partial parsing. In Bunt, H., and Tomita, M., eds., *Recent Advances in Parsing Technology*. Kluwer Academic Publishers. 295323.

Pustejovsky, J. 1991a. The generative lexicon. *Computational linguistics* 17(4):409–441.

Pustejovsky, J. 1991b. The syntax of event structure. *Cognition* 1(41):47–81.

Pustejovsky, J. 2013. Dynamic event structure and habitat theory. *Proceedings of GL2013* 1–20.

Rickel, J., and Porter, B. 1997. Automated modeling of complex systems to answer prediction questions. *Artificial Intelligence* 93(1):201–260.

Wouters, A. G. 2003. Four notions of biological function. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 34(4):633–668.