Cyc and the Big C: Reading that Produces and Uses Hypotheses about Complex Molecular Biology Mechanisms

Michael Witbrock, Karen Pittman, Jessica Moszkowicz, Andrew Beck, Dave Schneider, Doug Lenat

Cycorp, Inc. 7718 Wood Hollow Drive, Suite 250, Austin, TX 78731
{witbrock,pittman,jmoszko,abeck,daves,lenat}@cyc.com

Abstract

Systems biology, the study of the intricate, ramified, complex and interacting mechanisms underlying life, often proves too complex for unaided human understanding, even by groups of people working together. This difficulty is exacerbated by the high volume of publications in molecular biology. The Big C ('C' for Cyc) is a system designed to (semi-)automatically acquire, integrate, and use complex mechanism models, specifically related to cancer biology, via automated reading and a hyper-detailed refinement process resting on Cyc's logical representations and powerful inference mechanisms. We aim to assist cancer research and treatment by achieving elements of biologist-level reasoning, but with the scale and attention to detail that only computer implementations can provide.

Reading about Complex Mechanisms

While the current state of the art provides shared vocabularies and curated facts about complex mechanisms such as systems biology at enormous scale and with enormous utility, the representations used do not approach the level of detail (or the expressive power) found in a scientific paper. Manual curation of an expressive detailed knowledge base at that level of detail is infeasible, especially given the combined volume of the relevant literature. Moreover, the current state of automatic relationship extraction (Cohen and Hersch, 2005) and inference, within systems biology, does not provide models with the power to answer complex questions about biological pathways' functions and dynamics, or to automatically suggest model extensions and abstractions. In this paper, we present our plan for, and early experience constructing, The Big C ('C' for Cyc), a system that, as part of DARPA's Big Mechanism effort (DARPA, 2014), will enable the construction, maintenance and use of cancer mechanism models of unprecedented complexity via a hyper-detailed reading and refinement process resting on Cyc's uniquely faithful contextual higher-order logic representations, and its uniquely powerful inference mechanisms (Lenat, 1995)(Lenat *et al.,* 2010). We believe formation of such detailed explicit models from textual big data, allowing for automated reasoning, will become an increasingly powerful alternative and adjunct to the currently widespread use of statistical inference to form implicit models. The data scale for The Big C is daunting – PubMed adds more than 160,000 cancer-related papers per year (Corlan, 2004) and the UniProt protein database alone has more than half a million curated entries.

Figure 1 outlines the architecture of The Big C. Scholarly Big Data of two kinds are processed: research publications and structured biomedical data. Once a paper is partly read, Cyc forms biological hypotheses; these hypotheses are subsequently independently verified using Cyc's existing knowledge and loaded data-sources, or by additional system-initiated reading.

We illustrate in this paper examples of how these mechanisms can support biologist-level reasoning, but with the scale and attention to detail only computer implementations can provide. We hope to accelerate cancer research by:

- moving from text search and link graphs towards full question answering;
- holistic analysis across papers, including automated detection of contrary evidence, including for example: detecting claims made in a publication, giving a citation which does not actually make that claim;
- notifying researchers when there is a new finding relevant to some of their own models.

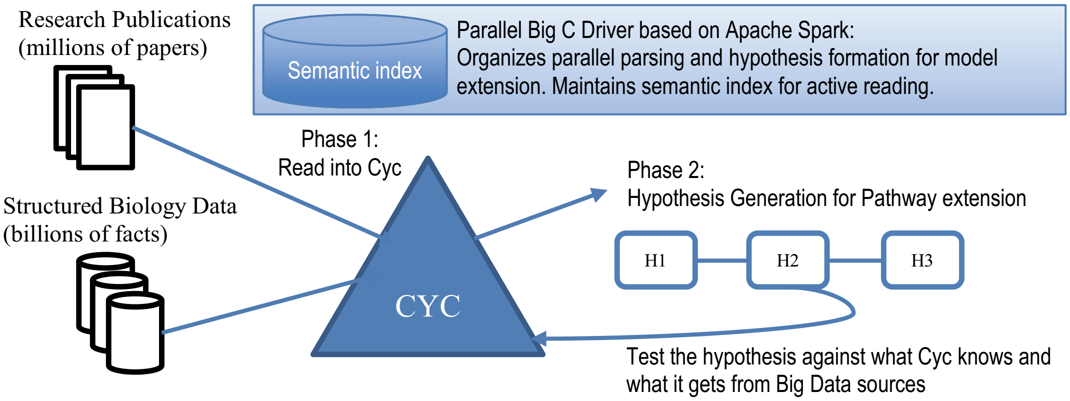
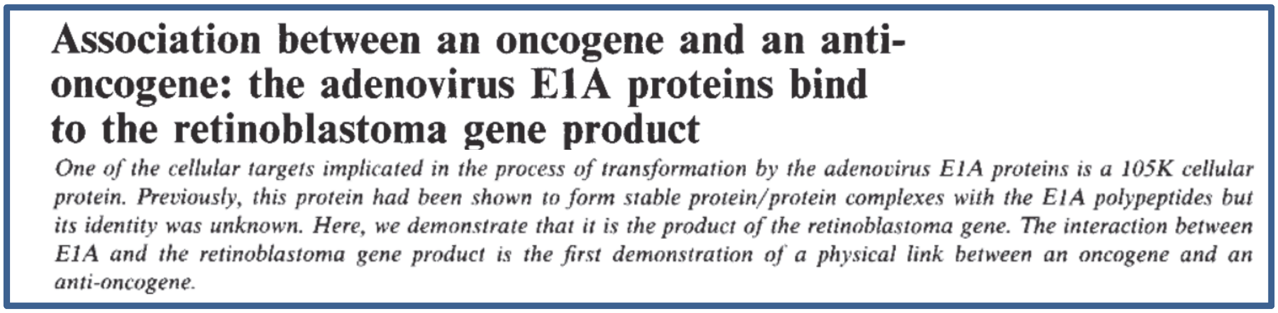

Research Publications
(millions of papers)
Semantic index
Parallel Big C Driver based on Apache Spark:
Organizes parallel parsing and hypothesis formation for model extension. Maintains semantic index for active reading.
Phase 1:
Read into Cyc
Phase 2:
Hypothesis Generation for Pathway extension
Structured Biology Data
(billions of facts)
H1
H2
H3
CYC
Test the hypothesis against what Cyc knows and what it gets from Big Data sources

Figure 1: Basic architecture of The Big C; the hypothesis cycle builds connected models from papers and facts



Association between an oncogene and an anti-oncogene: the adenovirus E1A proteins bind to the retinoblastoma gene product

One of the cellular targets implicated in the process of transformation by the adenovirus E1A proteins is a 105K cellular protein. Previously, this protein had been shown to form stable protein/protein complexes with the E1A polypeptides but its identity was unknown. Here, we demonstrate that it is the product of the retinoblastoma gene. The interaction between E1A and the retinoblastoma gene product is the first demonstration of a physical link between an oncogene and an anti-oncogene.

Figure 2: Big C will read, assemble, and explain the cancer pathway literature, like this example from White et al. (1988).

Because The Big C is a uniform mechanism, straightforward extensions could incorporate drug effects and individual patient (and cancer) genome information, working towards the promise of personalized medicine based on all current research. Moreover, the same modeling techniques can be applied to other fields in biology and beyond.

Use Case: Modeling Tumor Virus Pathways

The Big C system will read material like the 1988 paper of Whyte *et al.*, *"Association between an oncogene and an anti-oncogene: the adenovirus E1A proteins bind to the retinoblastoma gene product"*, the abstract of which is reproduced in Figure 2 (Whyte *et al.*, 1988). This paper was, perhaps, the initial triggering evidence for understanding the tumor virus mechanism in some detail (via co-precipitation of cellular and viral proteins); it demonstrated that an oncoprotein synthesized by cells infected with adenovirus binds directly to pRb, the protein encoded by the Rb gene. Implication of this protein, pRb, in tumor formation, the fact that the oncoprotein binds preferentially to minimally phosphorylated pRb, and the fact that pRb phosphorylation is greater in the late "G1" phase of the cell cycle, when the cell is committed to reproducing, than in the earlier stages of G1, led to the discovery that pRb gates cellular reproduction by sequestering E2F transcription factors. When these E2F factors are released as a result of phosphorylation or by pRB forming a complex with an oncoprotein, they enable transcription of genes leading to late G1 and S-Phase resulting in cellular reproduction. Normal phosphorylation is reversed after this point, but pRb sequestration by the oncogene is not, leading to unrestricted cellular reproduction: the R-point G1 gate has been disabled.

Given a scientific paper, The Big C will assemble the evidence into a model, like that in the previous paragraph, summarize, explain, and answer questions about that model, and form hypotheses, some of them scientifically interesting like the two hypotheses italicized in:

*"*One of the most important – and most curious – discoveries of the 1990s was that virtually all DNA tumor viruses [...] encode proteins that inactivate both Rb and p53. [...] why should these have been singled out for inactivation [...]? ***The answer may be that it is almost impossible for a tumor of epithelial origin to form unless the p53 and Rb tumor-suppressor pathways have been inactivated.*** *We predict that most of the cancers that now appear to be devoid of mutations in these two pathways will eventually be shown to contain them."* (Vogelstein and Kinzler, 2004)

The Big C offers the promise of greatly reducing researcher effort in forming hypotheses. The passage shown in bold above generalizes over the contents of a great many oncogenesis pathways, and then, based on the nature of identified commonalities, gives rise to a hypothesis, which may be supported or refuted by further evidence in the literature or specific research. The current state of the art in biomedical information extraction cannot read relationships of sufficient complexity to represent the pRb model, and the representations used are not inferentially powerful enough to enable either the generalization in bold or the consequent hypothesis in italics[1]. We aim to read to formal representations with expressivity close to English, but supporting the maintenance and elaboration of a network of inferentially productive mechanism models, supporting active hypothesizing and confirmation of extensions, and supporting rich question answering and model description.

Background Knowledge and Curated Fact Sets

A significant fraction of the knowledge that The Big C will assemble into mechanisms will be found in existing formal biological knowledge sources, including the sources comprising the $3.5x10^9$ RDF triples of Bio2RDF (e.g., Go, OBO, iProClass, etc.) (Belleau *et al*., 2008) (Callahan *et al.*, 2013). These can be mapped into the KB using the Cyc SKSI facility, which, after the schemata of databases and SPARQL endpoints have been described using a schema mapping tool, allows their content to be used in inference as if it were represented directly (Masters and Güngördü, 2003). However, the simplicity of the knowledge representation language used in these existing sources limits both their scope for fully representing the content of a scientific paper, and the complexity of the reasoning that can be directly performed with them, even when integrated with a single ontology (Callahan *et al.*, 2013). For this reason, while we will map such sources for use by Cyc, we will also use the Cyc Semantic Construction Grammar (SCG) language-to-logic system to actively read model-salient material into the highly expressive CycL language. SCG coverage will be critically dependent on the use of gene and protein identifiers and types from external curated sources.

Reading with Semantic Construction Grammar

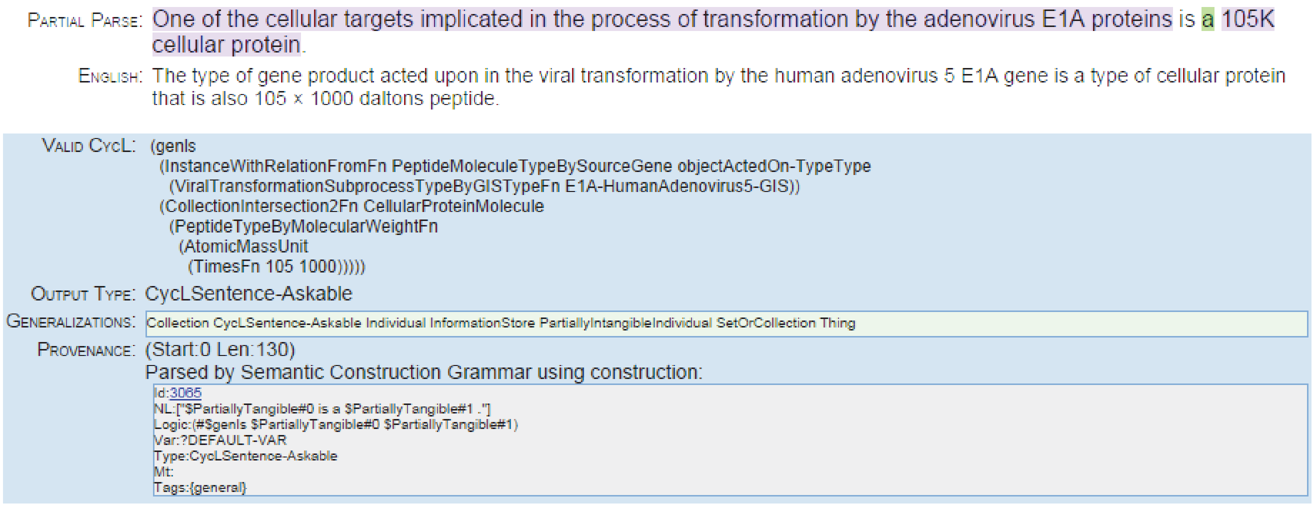
The Cyc SCG parser actively uses existing knowledge when extracting usable semantics expressed in natural language. We enable this by describing SCG constructions: semantically constrained structures in a natural language (like English), for which logically composable logical representations can be produced. The approach, which was inspired both by work in the 1970s by Charles Fillmore that lead to FrameNet (Fillmore, 1976), and by Carnegie Mellon University researchers on Example Based Machine Translation (Brown, 1996), is, in its first implementation, non-syntactic; it relies on recursively finding sequences of semantically understood elements and surface word forms for which a precise semantics can be assigned. In SCG, precise semantics are assigned to sentence elements using the CycL language, which offers a good combination of representational power and vocabulary, and is directly usable in inference.

The SCG system relies on three reasoning processes: lexical matching, taxonomic generalization, and semantic verification. First, it matches terms and phrases in the natural language input to logical symbols, from the Cyc vocabulary[2] and curated data-sources, yielding a sequence of uninterpreted lexical items (mostly closed class "structural" terms) and perhaps several hypothesized semantic interpretations of each recognized term. Unknown words (such as new proteins) can also be assigned semantics by specialized "unknown word" constructions. Each alternative semantic term is then generalized using the Cyc "isa" and "genls" taxonomies. The Cyc KB is very large and detailed, so this generalization may go through tens of levels before terminating at "Thing". Each path through these generalized partial interpretations is then matched against the full set of templates. When a match is found, the template semantic roles are replaced by the logically interpreted terms in the matching language, and the resulting typed logical expression becomes a hypothesized interpretation for the covered section of language. This process continues until no new matches are possible, resulting in one or more logically consistent full interpretations of the input text, or in islands of such interpretations within the text. This process can be applied in parallel across all papers, or on demand, based on hypothesis-driven search in a semantically tagged index. While the resulting interpretations are always syntactically correct logical strings, some of them are meaningless, and can be eliminated by inference; that is, by Cyc proving that either there is no possible world in which the logical statement can be true, or that there cannot be the sort of thing, in the universe, that the sentence describes, or merely that what is described is extremely unlikely to exist.[3] SCG's recursive matching and powerful use of reasoning for verification are significant advances on template techniques previously applied successfully to biomedical IE, *e.g.,* Yu and Agitchen (2003). As well as using local text, the current SCG parser can also instantiate meanings in the logic from anaphoric referents.

[1] For a survey of recent related research in biomedical information extraction, c.f. Pyysalo *et al*. (2013), Preiss (2014).

[2] The Cyc vocabulary contains logical symbols for hundreds of thousands of concepts, and is available under a CC-by license as OpenCyc, meaning that the results of The Big C reading can be reused without hindrance.

[3] These proofs are done by incomplete inference over the whole KB, and may, of course, be incorrect. As a practical matter, though, they are usually correct, and are extremely useful for eliminating incorrect interpretations.



PARTIAL PARSE: One of the cellular targets implicated in the process of transformation by the adenovirus E1A proteins is a 105K cellular protein.

ENGLISH: The type of gene product acted upon in the viral transformation by the human adenovirus 5 E1A gene is a type of cellular protein that is also 105 × 1000 daltons peptide.

VALID CYCL:
```
(genls
  (InstanceWithRelationFromFn PeptideMoleculeTypeBySourceGene objectActedOn-TypeType
    (ViralTransformationSubprocessTypeByGISTypeFn E1A-HumanAdenovirus5-GIS))
  (CollectionIntersection2Fn CellularProteinMolecule
    (PeptideTypeByMolecularWeightFn
      (AtomicMassUnit
        (TimesFn 105 1000)))))
```

OUTPUT TYPE: CycLSentence-Askable

GENERALIZATIONS: Collection CycLSentence-Askable Individual InformationStore PartiallyIntangibleIndividual SetOrCollection Thing

PROVENANCE: (Start:0 Len:130)
Parsed by Semantic Construction Grammar using construction:

```
Id:3065
NL:["$PartiallyTangible#0 is a $PartiallyTangible#1 ."]
Logic:(#$genls $PartiallyTangible#0 $PartiallyTangible#1)
Var:?DEFAULT-VAR
Type:CycLSentence-Askable
Mt:
Tags:{general}
```

Figure 3: SCG applies the construction in "Provenance" to recursively combine other interpretations, completing the sentence.

Figure 3 shows the inferentially productive logical output of the SCG parser for the first sentence in the abstract in Figure 2. Similarly, the italicized portion of (Sent. 2) "Previously, *this protein has been shown to form stable protein/protein complexes with the E1A polypeptides*, but its identity was unknown." is parsed (including a semantic anaphor search) to:

```
(proteinMoleculeTypesFormComplexOfType
  (GeneProductTypeFn E1A-HumanAdenovirus5-GIS)
  (InstanceWithRelationFromFn
    PeptideMoleculeTypeBySourceGene objectActedOn-TypeType
      (ViralTransformationSubprocessTypeByGISTypeFn
         E1A-HumanAdenovirus5-GIS))
```

while (Sent 3.) "Here we demonstrate that *it is the product of the retinoblastoma gene*," is translated to:

```
(coExtensional
  (GeneProductTypeFn RB1-Human-GIS)
  (InstanceWithRelationFromFn PeptideMoleculeTypeBySourceGene
    objectActedOn-TypeType
    (ViralTransformationSubprocessTypeByGISTypeFn
      E1A-HumanAdenovirus5-GIS)))
```

These three sentences read together are sufficient to explain the title of the paper, which supports their correctness for assembly into a cancer mechanism model, as we shall see later. The sentence that starts *"The interaction between E1A and the retinoblastoma gene product…"* illustrates the utility of inference during parsing: although *"the retinoblastoma gene product"* here clearly refers to a protein, due to its interaction with E1A (itself ambiguous between a gene and a protein), on the syntactic face of it, this reference could have, incorrectly, been to (ProductFn RB1-Human-GIS), a version of the genetic information itself offered for sale. This implausible interpretation can only be removed on semantic grounds, either because such products are unlikely (by virtue of any Gene-GIS being an implausible product) or because a product interacting with a gene or a protein is unlikely, if not logically impossible. Similarly, a syntactically tempting partial parse of *"the process of transformation […] is a 105K cellular protein"* is blocked because it can be proved that no process is a protein. The Big C is, in this sense, uniform: reasoning about combining components of interpretations during reading uses the same inference methods as assembly, and those methods, in turn, are used to support explanation.

Major Challenges

Extending and Improving SCG Reading

We envision evaluation-driven extensions to the SCG reader in The Big C, including:

(1) Where available, incorporating biomedical NER and shallow relation extraction to provide soft type and relation hypotheses[4];
(2) Incorporating syntactic as well as semantic features as constraints on some constructions whose purely semantic selectivity is too low;
(3) Incorporating The Big C's assembly model theory (Mt – see below) into reading, so that interpretations track which models support them;
(4) At least partially automating the discovery of semantic constructions by extending the generalization and pattern induction techniques that have been previously applied for shallow fact extraction.

To support hypothesis driven reading, we will build an index based both on document text and on the relatively inexpensive part of the parsing process: lexical tagging of the terms currently represented in The Big C's KB, followed by generalization. This will allow us to find, for

[4] By soft hypotheses we mean hypotheses that can be over-ridden by SCG constraints – a "peptide" NER tag for EA1, for example, could be over-ridden by a construction that required a gene.

example, any reference to HumanAdenovirus5 if an assembly task is seeking knowledge about tumor viruses.

Maintaining Interpretations and Hypotheses

Our aim is to combine curated facts and entities from external sources, detailed knowledge read by SCG, extracted relationships (perhaps with associated probabilities) from other IE techniques, and both supported and unconfirmed hypotheses into a set of connected, compact, and locally consistent theories about mechanisms. Previous work has approached assembly by attachment to a simple ontology (Coulet *et al*., 2011), but the representation language used has limited the detail and inferential power of the resulting models.

Our approach rests on maintaining a set of Cyc microtheories representing consistent possible mechanisms (Mechanism Theories based on Cyc Microtheories, or Mts), and on active extension via hypothesis of those theories. When a newly read fact-candidate, such as those above, becomes available, it will be considered for incorporation into theories to which it is relevant and with which it is consistent: Mts that have "binding sites" for the new assertion. Forward chaining inference based on each addition and the content of the theory may enrich the model directly (for example, confirming a causal link may enable the conclusion that a particular gene gates an entire cellular process), or allow the establishment of a mechanism hypothesis later confirmed by active reading. This notion originates with Swanson's ABC model (Swanson and Smalheiser, 1996): A *influences* B, B *influences* C, *therefore* A *might influence* C; we will greatly extend and specialize such rules for hypothesis formation. Resulting hypotheses will be stored in hypothesis microtheories (Ht) associated with each Mt, and will generally be maintained by Cyc forward inference with truth maintenance.

As an example, a statement that protein A affects the activity of protein B supports a (weak) hypothesis that protein A initiates inactivation by phosphorylation of protein B. However, it certainly does not entail such involvement.

An example hypothesis formation rule, whose conclusion provides possible targets for confirmatory reading, follows: (1) *If a cellular protein X is affected by a process caused in part by protein Y, hypothesize that X is inactivated by hyperphosphorylation in a pathway initiated by Y* (this rule is expressed in logic in Figure 4)

Other hypothesis rules (directly relevant to our reading example) include: (2) *If a cellular protein X forms a stable complex with a viral protein, hypothesize that X is inactivated by phosphorylation within some pathway;* and: (3) *If a cellular protein X is affected by a sub-process of a viral transformation caused by the viral protein Y, hypothesize that X is inactivated by forming a stable protein/protein complex with the viral protein*. Making use of background knowledge about the viral origin of E1A, rule 2's conclusion would raise the hypothesis that pRb is inactivated by phosphorylation. If that hypothetical conclusion is already part of or provable from the Mt, then applicability scores can be raised for the Mt itself, the conclusion, the triggering facts, and the SCG patterns (or external data sources) that produced them. In the case of the Mt, this will raise the likelihood that new facts that share terms with it (e.g. further facts read or retrieved about E1A or pRb) will be tested as possible extensions. Similarly, searching for and successfully performing active reading on material that validates the hypothesis would raise these scores. More simply, non-hypothesizing rules (such as the existing Cyc rule that S-Phase follows G1-Phase) will increasingly often entail what is read with respect to an accurate model; attempts to elaborate an Mt with a read or ingested assertion that is already provable in that Mt can cause similar rescoring. In this way, The Big C can learn to do more effective reading and more effective inference, and can maintain focus on expanding a set of most promising models.

```
(implies
 (and (ist ?MODEL-EXT  (and
   (genls ?CELLP CellularProteinMolecule)
   (genls ?PROTEIN-TYPE ProteinMolecule)
   (someTypePlaysRoleInSituationType ?PROC ?CELLP objectActedOn)
   (someTypePlaysRoleInSituationType ?PROC ?PROTEIN causalActors)))
  (hypothesisContextForModelExtension ?MODEL-EXT ?HYPOTH-MT))
 (ist ?HYPOTH-MT (thereExists ?PATHWAY  (and
  (biochemicalPathwayTypeInitiatedByType ?PATHWAY ?PROTEIN)
  (biochemicalPathwayTypeStepTypes ?PATHWAY
   (HyperphosphorylationOfTypeFn ?CELLP))
  (causes-SitTypeSitType   (HyperphosphorylationOfTypeFn ?CELLP))
   (InactivationOfProteinMoleculeTypeFn ?CELLP))))))
```

Figure 4: Rules like this form hypotheses from read material

While these general molecular biology rules will often have to be manually generated (in consultation with our Mayo Clinic SMEs) and while the inference process may be complex, using as it does the full current state of The Big C's KB, the basic process of maintaining mechanism theories (Mts) could sound straightforward as stated, but it is not. Uncertainty in both the reading process and in the nature of the scientific claims being made means that it will frequently be necessary to frame several competing interpretation hypotheses or predictions. This, combined with the sheer bulk of the cancer pathway literature suggests that a very large number of candidate mechanisms may have to be maintained on the path to finding those that comprehensively and consistently integrate what is known. This model explosion is mitigated in part by Cyc's treatment of theories as first class objects – theories about larger mechanisms can inherit and share claims from theories about sub-mechanisms, and vice versa. Alternative theories then need only explicitly represent their differences, while retaining their full inferential scope. Despite this advantage, we expect that research will be required to mitigate model-assembly cost.

Conclusion

A researcher reading about the viral cancer mechanisms for Adenovirus, HPV, and SV40 can see that they share a common form. From that, it is reasonable to hypothesize, as Vogelstein and Kinzler (2004) did, that all viral cancer pathways share this form. The goal of The Big C is to form such mechanism summaries, and such scientifically meaningful hypotheses. The Big C's use of a mechanism theory (Mt) graph makes this possible. Repeated assertions from these mechanisms will have been moved to a shared generalization Mt during assembly, leaving three sibling graphs with only the differences (the virus, the viral protein, and, in the case of HPV only, the subsequent pRb degradation). This sibling status is important, as is the separation of differences into their own Mts: it makes generalization tractable. Similar to the Semantic Construction Grammar's search in generalization space for a matching semantic pattern, The Big C will generalize predicates and terms in sibling Mts until it finds a common theory or a generalization threshold is reached. We expect that subject matter experts will be able to make use of the English translation (i.e., an automatically generated English version of the formal logical representation within The Big C) of both (1) the general theory, added to the mechanism (Mt) graph, and (2) the generalization pattern (what had to be generalized, and how), represented as a universal hypothesis. As more data is read, the number of these sibling models which logically agree (which differ only in the specific provenance of some of their facts) is likely to increase; the existence of multiple supports with varying provenance, of this type, will be recorded during abstraction, and will increase the score of the resulting model. Similarly, we hope for useful integrative results that could not easily be tracked by people, for example, alerting researchers when a new result is reported which calls into question a whole "ripple" of earlier research which depended on an earlier, now-questionable hypothesis or, worse, what appeared to be a well-known result. Our plan is for The Big C to effectively perform a useful "first pass" over large collections of scientific papers, providing crucial visibility into otherwise disparate literature and information stores. Our hope is that this will result in novel and effective personalized cancer treatments.

Acknowledgements

This effort is sponsored by the U.S. Army Research Office under contract W911NF-14-C-0106. The content herein does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The authors gratefully acknowledge the assistance of Dr. Lawrence Lefkowitz for helpful comments on drafts.

References

Belleau, F. et al., 2008. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform,* October, 41(5):706-716.

Brown, R., 1996. Example-based machine translation in the Pangloss System In *COLING-96: Proceedings of the 16th International Conference on Computational Linguistics.* 169-174. Copenhagen, Denmark.

Callahan, A., Cruz-Toledo, J., Ansell, P. & Dumontier, M., 2013. Bio2RDF Release 2: Improved coverage, interooperability, and provenance of life science linked data. *The Semantic Web: Semantics and Big Data in Lecuter Notes in Computer Science,* Volume 7882:200-212.

Callahan, A., Cruz-Toledo, J. & Dumontier, M., 2013. Ontology-based querying with Bio2RDF's linked open data. *J Biomed Semantics,* 4(S1).

Cohen, A. & Hersch, W. R., 2005. A survey of current work in biomedical text mining. *Briefings in Bioinformatics,* March, 6(1):57-71.

Corlan, A. D. *Medline trend: automated yearly statistics of PubMed results for any query,* 2004. Web resource at URL:http://dan.corlan.net/medline-trend.html. Accessed: 2014-11-30

Coulet, A. et al., 2011. Integration and publication of heterogeneous text-mined relationships on the semantic web. *Journal of Biomedical Semantics,* Volume 2 (Suppl. 2) S10.

DARPA, 2014. *Big Mechanism.* [Accessed October 2014] http://www.darpa.mil/Our_Work/I2O/Programs/Big_Mechanism.aspx

Fillmore, C., 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences, 280:*20–32.

Lenat, D., 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM,* November, 38(11):33-38.

Lenat, D., Witbrock, M., Baxter, D., Blackstone, E., Deaton, C., Schneider, D., Scott, J., & Shepard, B., 2010. Harnessing Cyc to answer clinical researchers' ad hoc queries. *AI Magazine,* 31(3).

Masters, J. & Güngördü, Z., 2003. Semantic knowledge source integration: A progress report. In *Proceedings International Conference Integration of Knowledge Intensive Multi-Agent Systems*, 562-566. Cambridge, Mass.: IEEE.

Preiss, J., 2014. Seeking informativeness in literature based discovery., In *Proceedings of the 2014 Workshop on Biomedical Natural Language Processing*. 112-117. Baltimore, Maryland, ACL

Pyysalo, S., Ohta, T. and Ananiadou, S.. (2013). Overview of the Cancer Genetics (CG) task of BioNLP Shared Task 2013. In: *Proceedings of the BioNLP Shared Task 2013 Workshop*.:58-66, Sofia, Bulgaria, ACL

Swanson, D. & Smalheiser, N., 1996. Undiscovered public knowledge: a ten-year update. *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining:*295-298. Portland, Oregon. AAAI

Vogelstein, B. & Kinzler, K., 2004. Cancer genes and the pathways they control. *Nature Medicine,* August, 10(8):789-799.

Weinberg, R., 2014. *The Biology of Cancer.* 2nd Edition ed. New York: Garland Science.

Whyte, P. et al., 1988. Association between an oncogene and an anti-oncogene: the adenovirus E1A proteins bind to the retinoblastoma gene product. *Nature* 334:124-129.

Witbrock, M., Cynthia Matuszek, C.,Brusseau, A., Kahlert, R., Fraser, C.B. & Lenat,D. 2005. Knowledge begets knowledge: Steps towards assisted knowledge acquisition in Cyc. In *Proceedings of he AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors*:99-105, Palo Alto, CA..

Yu, H. & Agitchen, E., 2003. Extracting synonymous gene and protein terms from biological literature. *Bioinformatics,* 19(Suppl. 1):340-349.