Collaborative Workspaces for Pathway Curation

Funda Durupinar-Babur[1], Metin Can Siper[2], Ugur Dogrusoz[2], Istemi Bahceci[2], Ozgun Babur[1], Emek Demir[1].
1 Oregon Health and Science University, Computational Biology Program, Portland, OR
2. Bilkent University, Dept. of Comp Engineering, Ankara, Turkey

Abstract— We present a web based visual biocuration workspace, focusing on curating detailed mechanistic pathways. It was designed as a flexible platform where multiple humans, NLP and AI agents can collaborate in real-time on a common model using an event driven API. We will use this platform for exploring disruptive technologies that can scale up biocuration such as NLP, human-computer collaboration, crowd-sourcing, alternative publishing and gamification. As a first step, we are designing a pilot to include an author-curation step into the scientific publishing, where the authors of an article create formal pathway fragments representing their discovery- heavily assisted by computer agents. We envision that this "micro-curation" use-case will create an excellent opportunity to integrate multiple NLP approaches and semi-automated curation.

Keywords—biocuration, pathways,

I. Introduction

Molecular biology studies the molecular components and mechanisms that control a cell's response to stimuli. Traditionally, this was a piecemeal effort primarily conducted through carefully-designed series of experiments that isolate, elucidate and confirm a part of the mechanism under a certain context. A byproduct of this process is knowledge fragmentation – putting these components into a mechanism, often called a pathway, that can describe cellular behavior is extremely challenging. One needs to assemble these pieces across many publications, negotiate differences due to biological context and experimental setup, resolve conflicts and create a coherent model. The majority of this knowledge integration happens informally through review articles within a very limited scope – often focusing around a couple of proteins or genes.

The biology, however, is changing rapidly primarily due to system scale "-omic" profiling and "big data". This, in turn, creates an urgent necessity for large scale, formal models of cellular processes – way beyond the scale of a current review article. This led to a rapid proliferation of pathway databases or curation groups- which currently stands at 600. Together, they have curated hundreds of thousands detailed biochemical reactions and millions of interactions. Yet, this is still only a small amount of the knowledge in the literature (our estimation is somewhere between 1 to 3%) and due to the rapid increase in our knowledge, this gap widens every day. Curation efforts – although extremely valuable-- also tend to be expensive. NIH spent 1.2 billion dollars on supporting data curation in the last decade– and the need is ever increasing.

How can we scale biocuration up by two orders of magnitude? We believe that the solution lies in the intersection of NLP, Artificial Intelligence and crowd-sourcing. It might also require, potentially, a social change in the way we publish and disseminate our results. There are already multiple efforts in these directions – but they are currently very disjoint. In order to enable exploring these directions, we developed a web-based collaborative workspace as a common platform. We also describe a pilot study, as an example downstream application, where authors directly curate formal pathway snippets that they discovered as a part of the publication process.

II. A Collaborative Workspace

The platform is composed of a web-based graphical editor and an application server. Figure 1 shows the platform architecture. The graphical editor is an extension of the SBGNViz framework [1], which is a web application based on Cytoscape.js [4] to visualize BioPAX [2] models represented by SBGN Process Description Notation (SBGN-PD) [3]. Cytoscape.js provides a rich set of graph visualization and manipulation features. SBGNViz adds domain specific glyphs, layouts and tool. Finally, our platform adds collaboration, conflict resolution and a strong link to existing NLP tools.

The application server is based on Node.js – a server-side Javascript environment with an event-driven, asynchronous IO model. This non-blocking structure allows combining agents that operate in different time scales. For example, extracting a relevant piece of information from the literature might take minutes whereas a visualization action can be accomplished in seconds. The server keeps track of the graphical editor's state and updates an underlying shared JSON model. The model can be edited concurrently via an Operational Transformation (OT) library called Share.js. OT provides versioning, concurrency control, conflict resolution in a manner similar to modern distributed code versioning systems and is successfully used in large collaborative editing applications. All operations are made persistent in a MongoDB database.

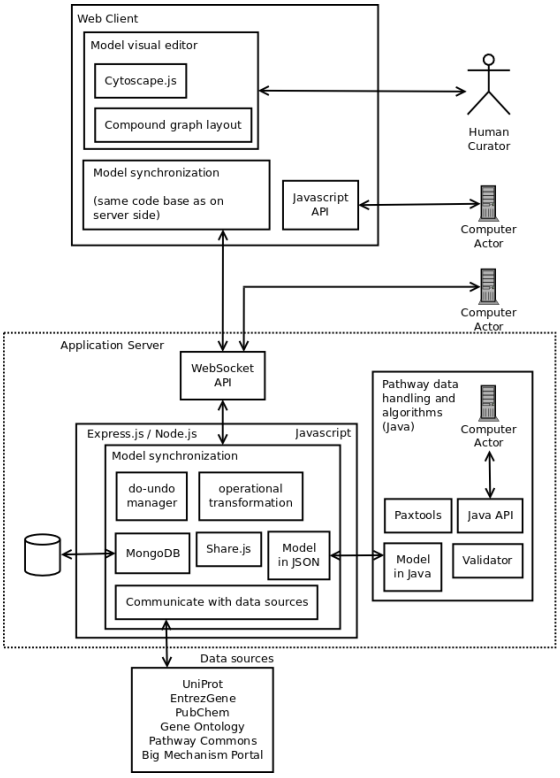

Figure 1. Platform architecture

Computer agents connect to the server through an API that uses socket.io interface. The API provides a two-way communication channel. Therefore, agents can both make changes to the model and get notifications about model updates. The server also provides a chat interface through which human users and computer agents can send text messages or image files. We have implemented the first version of the platform as well as adapters to several NLP systems and Pathway Commons pathway database.

III. MICROCURATION USE-CASE

As an initial use-case we aim to utilize this platform for capturing pathway fragments in scientific manuscripts with the help of authors of the publications. Since authors will have little to no training in curating pathways, maintaining a low-barrier of entry is crucial without sacrificing too much from the level of detail and formality of the representation. We envision to solve this through a heavily assisted, semi-automated process: First we will use text mining agents to automatically create a draft visual diagram of the pathway information in the manuscript. We will visually flag grounding issues as well as potential ways to resolve them. The fragments will also be aligned automatically to the pathways in public databases and users will be given option to use existing curated fragments when possible. The authors will then fix these issues or extend existing information. The resulting artifact will be a visual diagram and a formal BioPAX model, properly grounded and linked to an existing pathway corpus. This model fragment will be published as a supplement to the paper and can then be harnessed easily by curators and algorithms to assemble increasingly large mechanistic models. We envision that such a "micro-curation" step, if successfully embedded to the publication process, can enable rapid and scalable capture of atomic facts about cellular processes. By getting the fragments directly from the scientists who discovered them we hope to reduce communication noise, ambiguity and interpretation errors inherent to our current scientific communication and curation process.

IV. A GENERAL TOOL FOR BIOCURATION

In parallel to this pilot study, we also plan to actively use this platform to select most successful NLP and AI tools for pathway curation. The open real-time nature of the platform makes it extremely straightforward to swap, compare and combine different NLP approaches. This platform can also be used to extract non-interaction/pathway related information such as diseases, drug targets and mutations and combine these with pathway models. We can also explore how to extract supporting information such as experimental evidence or biological context. The web-based system enables collaborative use-cases and it can be used to collect large human curation data that can be used to build a gold standard corpus. We believe that the platform can become a major tool for biocuration and NLP communities.

ACKNOWLEDGMENT

This work was funded by the DARPA Big Mechanism program under ARO contract W911NF-14-1-0395.

REFERENCES

[1] M. Sari, I. Bahceci, U. Dogrusoz, S.O. Sumer, B.A. Aksoy, O. Babur, E. Demir, "SBGNViz: a tool for visualization and complexity management of SBGN process description maps", PLoS ONE, 10(6), e0128985, 2015.

[2] Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, et al. The BioPAX community standard for pathway data sharing. Nature Biotechnology. 2010;28(9):935–942. doi: 10.1038/nbt.1666. pmid:20829833

[3] Novère NL, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, et al. The systems biology graphical notation. Nature Biotechnology. 2009;27(8):735–741. doi: 10.1038/nbt.1558. pmid:19668183

[4] Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, et al. Integration of biological networks and gene expression data using Cytoscape. Nature Protocols. 2007;2(10):2366–2382. doi: 10.1038/nprot.2007.324. pmid:179479