WikiPathways: community-based curation of biological pathways

Alexander R. Pico[1], Thomas Kelder[2], Martijn P. van Iersel[2], Kristina Hanspers[1], Bruce R. Conklin[1] and Chris Evelo[2]

[1]*Gladstone Institute of Cardiovascular Disease, 1650 Owens Street, San Francisco, CA 94158 USA. apico@gladstone.ucsf.edu, khanspers@gladstone.ucsf.edu, bconklin@gladstone.ucsf.edu*

[2]*Department of Bioinformatics (BiGCaT) , Maastricht University, Universiteitssingel 50, 6229ER Maastricht, the Netherlands. thomas.kelder@bigcat.unimaas.nl, martijn.vaniersel@bigcat.unimaas.nl, chris.evelo@bigcat.unimaas.nl*

1 INTRODUCTION

Assembling biological pathways from information in scientific literature and biological databases is a challenging task. Building a pathway requires domain knowledge from specialists in various biological research areas. Furthermore, research is generating new biological knowledge continuously, making pathway curation an ongoing and dynamic process. To facilitate this process we developed WikiPathways[1], a wiki where users can curate pathways, using an easy to use pathway drawing tool. WikiPathways currently has over 1200 registered users and contains more than 1300 pathways for various organisms, spanning human, mouse, zebrafish, fruit fly, worm, yeast, plants and bacteria. WikiPathways is entering a phase of rapid growth, incorporating new pathway resources and engaging diverse communities with unique pathway needs.

In this presentation we will provide an overview of the background, current state, and future of WikiPathways. We will highlight several approaches we take to improve community-based curation with WikiPathways. And we will provide context for this project within the larger shifts toward data sharing, data curation and new models for publication.

2 METHODS

WikiPathways is built on top of the Mediawiki platform, implemented as a collection of extensions written in PHP as well as critical Java applet components embedded into the wiki interface. The applets control the entry of specific data fields and the pathway diagram itself (see PathVisio[3]). Pathway information is represented in a custom GPML format, which is stored in the native Mediawiki SQL database and rendered as images and text on each pathway page. Pathway objects representing genes, proteins, metabolites and other pathways are associated with database objects in embedded Derby databases.

3 RESULTS

WikiPathways started out as an experiment to see if we could improve upon existing pathway curation efforts, utilize community participation, and ultimately enhance the cycle of new data – new pathways – new ideas – new data. It is just one of many new communication tools in biology that aims to give researchers greater access, utilization and control over their collective data.

Three years on, the WikiPathways project has reached a number of milestones. Yet the project is still in its infancy, as is the entire field, as cultural shifts slowly expand the role of researchers to include greater data utilization and curation.

MILESTONES

Online. WikiPathways went online in March of 2007, just a few months after its original conception. Born of necessity, WikiPathways quickly replaced our internal system of pathway collection, curation and distribution. We had been curating a collection of pathway diagrams through the GenMAPP project since 2000 and the progress was frustratingly slow. Thousands of GenMAPP users would download our free software and make use of the various analysis and visualization features. Some would even generate novel pathway content in their field of expertise. In order to collect these additions and enhancements, we would have to track publications and then individually contact authors. At other times, we would directly contact labs with apparent expertise to contribute or review pathway content. The pathway files were emailed back and forth as platform-dependent .mapp files. This inefficient system resulted in the slow accumulation of an ever-outdated archive of pathway diagrams.

With WikiPathways, we were able to streamline the sharing, collection and updating of pathway content. All of the sudden it was a trivial act to work collaboratively on a pathway and then immediately include the new product in the release collection. Within the first year, the quality and quantity of pathways had increased more than in the prior 3 or 4 years combined.

This is an important point to stress, especially for individuals or organizations thinking about starting a community-based system: *it is critical that the new technology be immediately useful to your existing, internal group.* It will take a long while (if ever) for a community of contributors to adopt your new tool. If it's not immediately useful to *you* it will not likely survive long enough to attract a critical mass. Ideally, the new technology replaces an existing system that your existing group agrees is in dire need of an overhaul.

First unknown user. The unofficial "birthday" of WikiPathways is January of 2008 when we logged the first edit by a user no one in the former curation group knew. This was the first piece of evidence that this new communication tool could not only enhance the work of our group, but also effectively expand our group. We were officially crowdsourcing. Within a year, the size of our "crowd" rapidly grew from 1 to dozens to hundreds.

This is an important test for a community-based tool: can it handle larger and larger numbers of contributors? This is a test of not only the technology (is it stable and robust?), but also of the community itself (is it open and cooperative). Depending on the culture of a given field, community curation might just not work. Members of the community need sufficient mutual trust to work in the pre-competitive environment of collaborative data sharing and annotation.

Blessing by father of wiki. The concept and first implementation of a wiki was by Ward Cunningham in 1994. In addition to the wiki, Ward is considered a pioneer of both design patterns and Extreme Programming. These ideas share a common theme of *using minimal or transparent technology to enhance group development efforts*. Whether it is communicating design principles with colleagues or collaboratively writing code or collaboratively maintaining a web site, the real power of these technologies is that they facilitate collaboration and synergy by subtly directing each individual's effort. When used properly, the individual is more productive as a result of the facilitated collaboration and gains a deeper sense of teamwork and community.

So, we were greatly encouraged in our efforts with WikiPathways when we were contacted by Ward Cunningham:

"You may be interested, When I created the first wiki 13 years ago I had in the back of my mind the creation of a community of practicing software professionals (my discipline) that could engage in a "simplified scholarship" inspired by our best scientific disciplines. Imagine my satisfaction when my internet techniques are found attractive by the communities that most inspired me."

This is really an endorsement for all biological wikis. Our communities may still have a ways to go before they develop and adopt optimal practices for rigorous and rapid cycles of discovery, but this is an encouraging perspective from an expert in community practices that we have a solid (even inspiring) foundation in the discipline of biological sciences.

Quality and quantity. All metrics for quality and quantity of pathway content at WikiPathways are on the rise. In this presentation, I will provide the latest numbers and graphs regarding visitors, contributors, various levels of annotation, and utilization. One of the most interesting metrics is the growing proportion of contributors as our total number of registered users increases. This is not necessarily expected. For example, with our first 200 registered users there were approximately 20 (or 10%) that were actual contributors, those that had edited a pathway. Today, with over 1200 users there are approximately 260 (or 22%) contributors. Why would the percentage increase with numbers and/or time?

One possibility is that this trend is indicative of a slow cultural shift where data producers and data consumers are becoming data curators. In our particular field of pathway bioinformatics, many researchers are already practicing data consumers, meaning they routinely mine and analyze data from public databases and online resources to enhance or contrast their own data. Once you enter this practice, you quickly become aware of the dire need for better data format standards and annotation standards. Thus, you have a personal motivation to participate in the curation of the data yourself, to improve it for yourself as well as for others.

Other signs of our content quality include recent partnerships with model organism groups, other content curators, and granting agencies to provide pathway content and tools to their communities. We are working with Reactome to host their content at WikiPathways to collect broader feedback to incorporate back into their centralized curation

effort. We formed a partnership with the California Institute for Regenerative Medicine (CIRM) to provide relevant pathway content to represent and connect their various stem cell research grantees. These funded labs can, in turn, access WikiPathways through a custom portal to generate and curate pathway content directly. We recently added a new species database and initial set of pathways for the study of Tuberculosis. The curation and utilization of this content is being driven by a collaboration involving academic and industry groups spanning the US at three different sites. This is a first: real-time collaborative annotation and research of pathway information. It highlights unique strengths of a wiki-based tool. Finally, we recently collaborated with Gene Wiki and various wikipedians overseeing molecular and cellular biology content at Wikipedia. We developed and refined an interactive pathway illustration tool that highlights the role played by various genes, proteins and metabolites in critical biological pathways, such as Statin Metabolism and the TCA Cycle. You can click on neighboring genes in the pathway image to go to their dedicated wiki page. This representation of WikiPathways content helps to fulfill our outreach mission to use pathway diagrams in education.

WIKIPATHWAYS TODAY

The current set of tools and resources making up WikiPathways strive to maintain **ease of use.** Curation tools should lower the threshold for new users to start contributing. These tools facilitate pathway editing, establishing gene linkouts, citing literature, pathway tagging (e.g., by quality and ontology), tracking changes and threaded discussions.

Contributors receive **credit** for their work, for example by prominently listing authors for each pathway, ranking contributions and actively distributing pathway content through a CC license. These are examples of micro-attribution and open access that are transforming scientific publishing.

The **utility** of pathways as a research tool is an important incentive for our users. Therefore, we actively work on integrating new biological databases and analysis tools such as Cytoscape. In addition, all data can be accessed through our web service[2], workflow tools such as Taverna, and various export and integration options.

Communities with specific interests (e.g. stem cell research) can organize themselves via portals on WikiPathways. A portal is a customizable entry page centered around a subset of pathways and contributors within an established community.

The tutorial session for WikiPathways will detail the major features of the site. We will present several code examples to show how the web service can be used by bioinformaticians (e.g. perform enrichment analysis on WikiPathways in R) and web developers (e.g. how to link to relevant pathways on WikiPathways). We will demonstrate working examples of WikiPathways integration at Wikipedia/GeneWiki, SNPLogic, Taverna, Cytoscape and NCBI, showing you how it works and how to make it work for you.

WIKIPATHWAYS TOMORROW

We have an active and growing open source development community around PathVisio and WikiPathways. In addition to ongoing efforts to engage existing research communities interested in pathway content, we will be developing methods to integrate with more data formats and pipelines to improve exchange and curation efforts. We will be developing new extensions tailored for curators and for researchers, to provide relevant perspectives and tools to interact with the content. And, as always, there are important improvements to the user interface in general that are needed to enhance the user experience and optimally translate intention into action.

BIG IDEAS

WikiPathways is just one player in a much larger effort currently unfolding. While our specific goal is to transform “pretty” pictures of pathways into effective research tools, we are also participating in larger movements spanning all biological domains. These include shifts toward open access and open source, greater data sharing trends, and new models for publication and attribution. We are at an interesting intersection of a few early trends. Researchers are just beginning to get comfortable with full data utilization, meaning playing the roles of data producer, data consumer, and data curator. The first is the traditional and more familiar role. More and more, researchers are becoming fluent consumers of databases and computational tools. But the last role, as curator, is still very new. Solidifying this role will require cultural changes involving new metrics for participation in curation efforts and online communities, new standards for citation and publication, or rather microcitation and micropublication, and new standards for the data itself (to optimize human interaction with digital content and keep track of attribution, etc). Through the WikiPathways project we are eagerly playing our small part in these movements and welcome

collaborative efforts across wiki and other new communication technologies.

REFERENCES

1. A.R. Pico, T. Kelder, M.P. van Iersel, K. Hanspers, B.R. Conklin and C. Evelo. 2008. WikiPathways: Pathway Editing for the People , PLoS Biology, Vol. 6, No. 7. (1 July 2008), e184.
2. T. Kelder, A.R. Pico, K. Hanspers, M.P. van Iersel, C. Evelo and B.R. Conklin. 2009. Mining Biological Pathways Using WikiPathways Web Services. PLoS ONE. 2009;4(7).
3. M.P. van Iersel, T. Kelder, A.R. Pico, K. Hanspers, S. Coort, B.R. Conklin, C.Evelo. 2008. Presenting and exploring biological pathways with PathVisio. BMC Bioinformatics. 9: 1. 09.