Pathways4Life

1. Setting the stage

Pathway information is immensely useful for analyzing and interpreting large-scale omics data[1-4]. Pathway analysis methods ultimately rely on knowledge bases of pathway models. In this proposal, we stress the distinction between *pathway figures*, which are drawn purely for illustration purposes in a graphical file format (e.g., JPG, GIF, PNG) and *pathway models*, which contain standard identifiers and semantics that can be mapped to external resources in a structured file format (e.g., XML, RDF, JSON). Aligned with the broader Open Science goals, this proposal sets out to **unlock the wealth of pathway information currently trapped in countless *pathway figures* by transforming them into *pathway models* amenable to analysis and research**.

Despite the tremendous efforts by all pathway knowledge bases over the past decade, most pathway information is still published solely as static, arbitrarily drawn images—isolated, inert representations of knowledge that cannot readily be reused.

The challenge is to efficiently extract this information and model it as biological knowledge. Computational approaches alone have failed to fully automate this process, which is no surprise given the wide diversity among the images. Likewise, human efforts have fallen short, both at the level of central curation teams and at the level of individual researchers who choose PowerPoint and Illustrator over freely available pathway modeling standards and tools. This challenge is particularly well suited for a computer-assisted, human-crowdsourced solution. We propose to develop the Pathways4Life platform as such an innovative solution. **The Pathways4Life platform combines image processing, text recognition, and cutting-edge pathway modeling software together with scalable infrastructure for content management to facilitate the rapid modeling of pathway images through human crowdsourced tasks.** Each component derives from existing research projects and technologies—the innovation lies in bringing them together to address an ongoing biomedical challenge at an unprecedented rate.

2. Building an innovative platform

For Phase I, we propose two aims. First, we will collect, process, and classify pathway images from the open-access literature. The classification step will allow us to prioritize images based on curation goals (e.g., high-value genes and disease contexts) and feasibility. The second aim will be to develop a platform for presenting these images to human participants with pre-annotated nodes and simple tasks, such as connecting existing nodes and adding new nodes.

Aim 1: Collect, process, and classify pathway images

We will follow in the footsteps of other groups who tackled the problem of parsing and indexing figures from the scientific literature. Yale Image Finder[10] indexed over 1.5 million open-access images, but they parsed only the captions and not the text embedded in the images. Michael Baitaluk, et al. fine-tuned an optical character recognition (OCR) method to parse entire pathways to generate BiologicalNetworks.org[11]. This work will certainly inform our optimization work in Aim 1. Unfortunately, the models from this project were never released in a community standard format, *nor made publically available*. Regardless, even with optimization, the results were limited to 1012 pathways from ~25000 images, of which 87% were considered to be high quality. The only human input in this process was a “like/don’t like” button and comment form. So, **the real innovation we propose is to *not* rely solely on computational OCR, but rather to couple it from the start with a human crowdsourcing, citizen scientist component.** The OCR results are intended to lower the barrier to entry for participants, giving them a handful of recognized nodes to build upon and in the process learn how to add the remaining nodes.

We implemented an efficient process to maximize collection of pathway images from Pubmed Central (PMC) publication figures as part of the preliminary analysis described in the next section. The process starts with a query into the PMC image search feature. Results are returned as HTML, which is parsed to download the full-size figure image files and generate an annotation file for each image. The annotation file will allow us to present critical contextual information during the crowdsourcing stage. It will also be used to index images by author and to support focused keyword searches (e.g., disease-related images). This process will scale to encompass a broad range of queries to collect diverse and high-value pathway image file sets. We plan to collect a set of 16,000 pathway images in the first iteration of Aim 1.

Also as part of the sample analysis, we began to explore text-extraction software. We have already scripted the application of Adobe Acrobat and Tesseract[7] to generate extracted text file sets. These are useful for assessing the results in terms of recognizable gene symbol counts. But these programs also provide positional information for each block of extracted text. We will use this information to generate JSON models of nodes, preserving the annotation and position from the original image. This will allow us to overlay an interactive layer of modeled nodes onto the original pathway figures (see Aim 2). In Phase II it will be worthwhile to improve upon the out-of-the-box performance of these programs, considering the Difficulty Matrix on our sample set of ~4000 images (Fig. 1). Fortunately, the Easy-Easy corner of the matrix (top left) contains a large proportion of pathway figure images, which can be targeted in prototyping our crowdsourcing effort. However, half of the images are in the Easy-Hard quadrant (easy for human, hard for computer; top right), meaning that any improvements in automatic text extraction will result in having more pathways that are ready and amenable for human processing. The lower half of the matrix includes such small percentages that we can simply ignore it in this proposal.

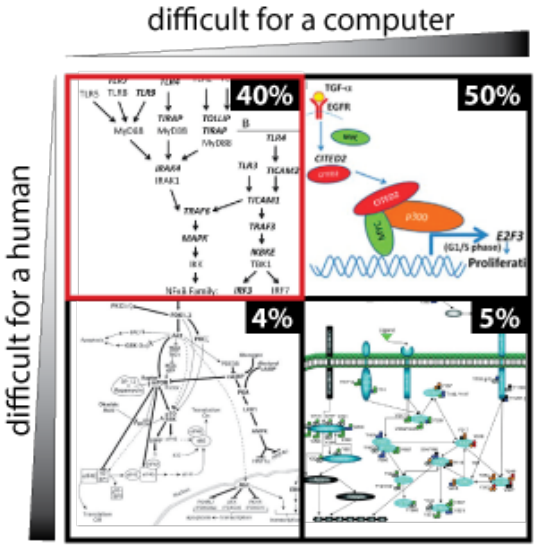

Figure 1. Difficulty Matrix for Human and Computer Parsing of Pathway Images. Examples of images that are easy for humans (top row), hard for humans (bottom row), easy for computers (left column), and hard for computers (right column). The percentages of pathways in each quadrant were estimated by inspecting ~40,000 sample images. Red highlights the “Easy-Easy” corner.

The product of this aim is a new, ever-growing and open access database of pathway images, annotated not only with source information but also with extracted gene symbols, multiple dimensions of functional classifications, and a JSON data overlay ready to be rendered and made interactive in Aim 2.

Beyond scouring html results for published pathway images, the longer-term strategy should be to get closer and closer to the source. The Pathways4Life platform could be extended to allow authors and journal editors to directly populate the figures to be processed by the crowd in sync with the publishing process. Connections with resources like FigShare [12] could help funnel images already tagged as “pathways” into the pipeline as well.

Aim 2: Develop an Interactive Digital Media Platform

We began to explore this approach as an international collaboration in 2007 within the WikiPathways project[5, 16, 17]. We are currently rolling out a JavaScript replacement of the Java Applet, called *pvjs*, which converts our xml pathway models into a JSON model and renders it as SVG. We have laid the foundation for a modularized, model-view-controller architectural pattern that integrates virtual DOM capabilities for fast performance and easy extensibility. This proposal will be the first demonstration of this extensibility, resulting in a novel platform that is entirely independent of WikiPathways and our existing projects.

Backend database and control logic—We will design and host a database to contain image, annotation, and JSON file references. These entries will map to node, and interaction tables in the database. A basic Python/Django web framework will be implemented to form template-based queries and views to serve content

to the customized pvjs tool and to update the database with new contributions. On the server side, we will process and store aggregated data across all sessions to assign confidence scores for each node and interaction. This strategy will allow us to distribute low-CPU, frequent computation at scale with the number of participants while also restricting server-side, moderate-CPU computation to fixed periods that we can adjust according to demand and resources.

We have sufficient infrastructure in place to develop and test the platform. As a modular set of virtualized services, we can deploy them using Amazon Web Services (AWS) to host large-scale beta and production crowdsourcing events as we progress into Phase II.

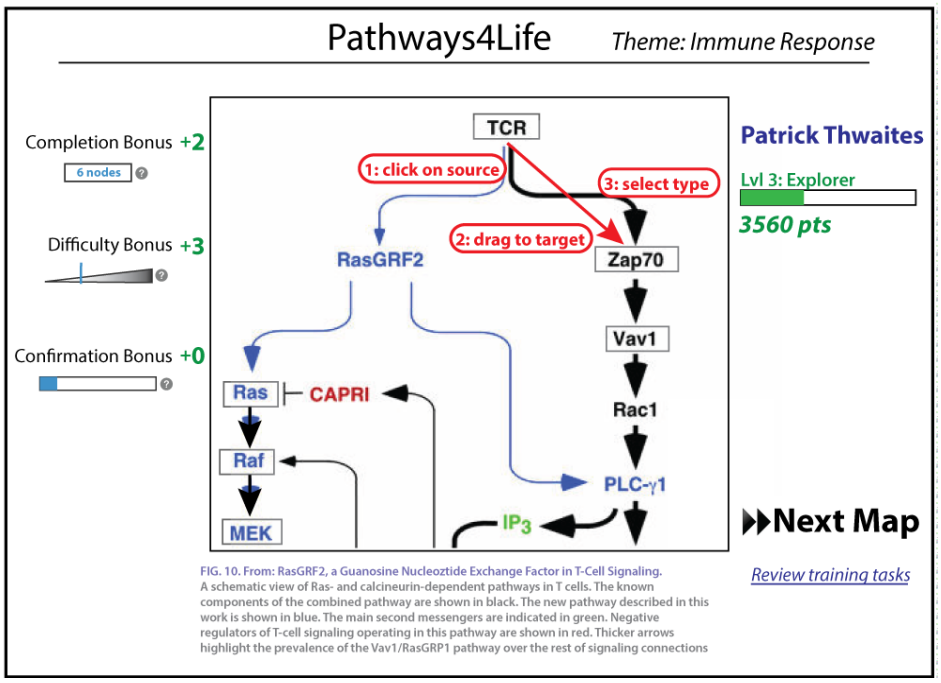

Figure 2. Pathways4Life Interactive Digital Media Interface. A pathway image and SVG-based modeling layer with OCR-identified nodes (boxed) and indications of available tasks. To define an interaction, click on an existing node to anchor the source and then click on a second node to indicate the target. Select from a list of interaction types to complete the task. To define a node, click on the image where the node should be added, type the name or symbol (which triggers an autocomplete pvjs database lookup), and then select the correct identity to complete the task. In this way, complete pathway images can be *traced* and effectively modeled by a series of these two easy-to-learn tasks.

***Customized Pvjs*—**Pvjs will require customization to work as a component of the Pathways4Life platform. The modular architecture of pvjs will readily accommodate customization. The new modules will add support for metadata data to the JSON model and an SVG visual feedback layer to support the task of modeling a pathway image (Fig. 2). The output will be a stream of JSON snippets that represent the individual changes made per task. Each snippet will be associated with a particular pathway image and participant such that we can quickly confirm that a particular snippet is novel or an Nth confirmation of a prior result. When all the snippets on a particular pathway are confirmed, the model will be queued for review. Community review and curation of the results will lead to their dissemination via multiple open-standard formats and communication channels, including but not limited to WikiPathways, Pathway Commons (BioPAX), and linked data (RDF).

3. Advancing open science

A survey of ~4000 published pathway figures highlights the challenges we propose to address. A PubMed Central (PMC) image search using the keyword “signaling pathway” generates over 40,000 results. Visual inspection of the first 5000 results, from publications spanning 2000–2015, revealed that 3985 (79.7%) contain a pathway image; the remainder contained only the word “pathway” in their captions. We then performed OCR using two parallel approaches: Adobe Acrobat Text Recognition (www.adobe.com) and Google’s Tesseract (code.google.com/p/tesseract-ocr). We cross-referenced the extracted text results against all known HGNC human gene symbols[9], including aliases and prior symbols, to assess the potential of these images to inform the curation of human and orthologous pathways. Acrobat and Tesseract each extracted over ~2300 HGNC symbols; ~730 (~32%) contained new information, human genes, and orthologs not captured in any pathway for any species at WikiPathways. Further, these approaches found significantly *different* sets of symbols—each

with its own uniquely trained OCR method—such that the combined results provide greater extraction counts across all categories: 3187 HGNC symbols in total and 1087 (34%) new to WikiPathways (Fig. 3, green).

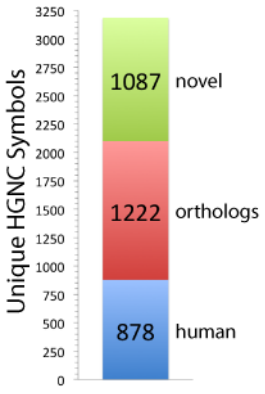

Figure 3. Recognized HGNC Symbols versus WikiPathways Content. Of 3187 recognized HGNC symbols extracted from 3985 pathway images, 878 (28%) matched symbols on human pathways at WikiPathways (blue), 1222 (38%) matched symbols on orthologous pathways at WikiPathways (red); 1087 (34%) are completely novel with respect to WikiPathways (green).

Several caveats to this survey suggest that even more new pathway information can be extracted from published images. (1) Another 38% of extracted symbols (Fig. 3, red) are found only on non-human pathways and thus may still represent novel human pathway content. (2) The OCR methods were used “out-of-the-box” and not trained on pathway images, and the images were not pre-processed. Thus there is a significant opportunity to increase total extraction counts. (3) Since we considered only 5000 of 40,000 results from only a single pathway-related search term, the search result space is much greater. (4) This survey ignored the *interactions* shown in the images. Thus, even the 28% of symbols that overlap with current human pathways (Fig. 3, blue) may provide new interaction content.

It is difficult to estimate the total number of new human gene symbols in the entire corpus of published pathway images. But for perspective, even if we were to only consider the first 3985 images and conservatively estimate the average number of genes per pathway image to be 7, and generously estimate the false positive rate to be as high as 20%, then at a proportion of 34% we would expect ~6100 new genes. **This would almost double the count of unique human genes at WikiPathways today—the equivalent of six years of work at current crowdsourcing rates.** The effort would also greatly expand upon the interactions and biological context for practically all the genes in pathway models.

The scope and timescale of this approach will have a dramatic impact on the rate and limits of new knowledge in the form of pathway models. By scaling up the pathway image collection to 16000 and extrapolating the sample set results (4 x 6100 = 24400), we predict that we will be inside the region of diminishing returns (w.r.t. unique genes because there are only so many), allowing us to prioritize our selection of pathways to crowdsource based on organism, density of novel genes and biological contexts (Aim 1). **This collection will contain more unique human genes than all current pathway archives combined. The modeling of this collection would thus *approach* the goal of having at least one representation of every human gene in a known pathway context.** The impact on the number of interactions modeled from this collection will be even greater, as novel interactions are captured even for non-novel nodes in pathway images. More broadly, the platform has the potential to make a lasting impact on how pathway modeling as well as other knowledge extraction is performed, shifting the focus closer and closer to the source, to capture this information in sync with the act of publication.

4. Disseminating and licensing output

Per our commitment to Open Science, all of this unleashed knowledge will be freely available in multiple formats, including JSON, BioPAX, GPML and RDF, and distributed to a wide range of independent, open access resources, such as WikiPathways[5], Pathway Commons[8], NCBI (www.ncbi.nlm.nih.gov/biosystems) and FigShare[9].

Here are links to our related projects and resources that demonstrate our commitment to open science:

- GitHub repository for this project, including preliminary PMC image extraction results: https://github.com/pathways4life
- WikiPathways, the product of our 9-year international collaboration so far: http://wikipathways.org
- Uniquely open terms of use and CC-BY licensing of WikiPathways content: http://wikipathways.org/index.php/WikiPathways:License_Terms

- 39 open source github repositories started and maintained by our collaboration:
  - https://github.com/wikipathways/
  - https://github.com/pathvisio/
  - https://github.com/bridgedb/
- 11 open access publications by our international collaboration:
  - http://www.ncbi.nlm.nih.gov/pubmed/?term=pico+ar+evelo+c

5. Planning for Phase II

During Phase II, we would refine both the image pre-processing and the modeling toolset based on direct feedback from early users of the Phase I prototype. In terms of OCR, the actively developed, open-source effort, Tesseract, in particular has considerable potential for improvement. We will add a common image pre-processing step that uses ImageMagick (www.imagemagick.org) to increase contrast, adjust orientation, and remove noise (e.g., other graphics) from a copy of the image. We will also use OpenCV[12] to identify and optimize regions of text in the image prior to OCR. Other methods are available to isolate, orient, and filter "objects" that may contain recognizable text[11, 13-15]. In addition, we would then focus on transforming the data from the crowdsourced tasks into confirmed pathway models. We will coordinate with other science crowdsourcing efforts through a common citizen scientist portal we are already exploring with Playmatics (http://playmatics.com). The goal of the portal would be to aggregate and share the activities of individuals across multiple science game and task platforms. Community review and curation of the results will lead to even broader dissemination via multiple open-standard formats and communication channels, including WikiPathways (GPML), Pathway Commons (BioPAX), and linked data (RDF).

References Cited

1. Kanehisa M, Bork P. Bioinformatics in the post-sequence era. Nat Genet. 2003;33 Suppl:305-10. doi: 10.1038/ng1109. PubMed PMID: 12610540.
2. Kelder T, Conklin BR, Evelo CT, Pico AR. Finding the right questions: exploratory pathway analysis to enhance biological discovery in large datasets. PLoS Biol. 2010;8(8). doi: 10.1371/journal.pbio.1000472.
3. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway Analysis: State of the Art. *Frontiers in Physiology*. 2015;6:383. doi:10.3389/fphys.2015.00383.
PubMed PMID: 20824171; PubMed Central PMCID: PMC2930872.
4. Kutmon M, Riutta A, Nunes N, et al. WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research*. 2016;44(Database issue):D488-D494. doi:10.1093/nar/gkv1024.
5. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: pathway editing for the people. PLoS Biol. 2008;6(18651794).
6. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic acids research. 2015;43(Database issue):D1079-85. doi: 10.1093/nar/gku1071. PubMed PMID: 25361968; PubMed Central PMCID: PMC4383909.
7. Smith R. An Overview of the Tesseract OCR Engine. Proc Ninth Int Conference on Document Analysis and Recognition (ICDAR)2007. p. 629-33.
8. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic acids research. 2011;39(Database issue):D685-90. doi: 10.1093/nar/gkq1039. PubMed PMID: 21071392; PubMed Central PMCID: PMC3013659.
9. Singh J. FigShare. *Journal of Pharmacology & Pharmacotherapeutics*. 2011;2(2):138-139. doi:10.4103/0976-500X.81919.
10. Xu S, McCusker J, Krauthammer M. Yale Image Finder (YIF): a new search engine for retrieving biomedical images. Bioinformatics. 2008;24(17):1968-70. doi: 10.1093/bioinformatics/btn340. PubMed PMID: 18614584; PubMed Central PMCID: PMC2732221.
11. Kozhenkov S, Baitaluk M. Mining and integration of pathway diagrams from imaging data. Bioinformatics. 2012;28(5):739-42. doi: 10.1093/bioinformatics/bts018. PubMed PMID: 22267504; PubMed Central PMCID: PMC3289920.

12. Pulli KB, A; Kornyakov, K; Eruhimov, V. Realtime Computer Vision with OpenCV. Queue - Processors. 2012;10(4):40.
13. Rodriguez-Esteban R, Iossifov I. Figure mining for biomedical research. Bioinformatics. 2009;25(16):2082-4. doi: 10.1093/bioinformatics/btp318. PubMed PMID: 19439564.
14. Xu SK, M. Boosting text extraction from biomedical images using text region detection. Biomedical Sciences and Engineering Conference (BSEC); 15-17 March 2011; Knoxville, TN: IEEE; 2011. p. 1-4.
15. Kuhn T, Nagy ML, Luong T, Krauthammer M. Mining images in biomedical publications: Detection and analysis of gel diagrams. J Biomed Semantics. 2014;5(1):10. doi: 10.1186/2041-1480-5-10. PubMed PMID: 24568573; PubMed Central PMCID: PMC4190668.
16. Hu JC, Aramayo R, Bolser D, Conway T, Elsik CG, Gribskov M, et al. The emerging world of wikis. Science. 2008;320(5881):1289-90. doi: 10.1126/science.320.5881.1289b. PubMed PMID: 18535227.
17. Waldrop M. Big data: Wikiomics. Nature. 2008;455(7209):22-5. doi: 10.1038/455022a. PubMed PMID: 18769412.