PCquery.jl: Automated Mining, Integration, and Annotation of Biological Signaling Pathways

Matthew Karikomi [1,2,3] Qing Nie [2,3]

[1]Mathematical, Computational, and Systems Biology Program [2]Department of Mathematics, UC Irvine [3]NSF-Simons Center for Multiscale Cell Fate Research

BioPAX is a graph-based representation of knowledge

Pathways in BioPAX [1] are heterogeneous graphs whose edges denote both conceptual (relational) knowledge and dynamics.

Figure 1. BioPAX knowledge schema. Image credit: [1]

Figure 2. Representation of a unique macromolecular state in BioPAX. Image credit: [1]

Pathways can be mined programmatically

Curated pathway databases, such as Reactome [2] and KEGG [3], are aggregated by the Pathway-Commons [6] initiative and ported to BioPAX, allowing PCquery to mine sets of related pathways using structured queries:

```
psParams = Dict(
  "q"=>"notch1",:filter=>[["notch1","nucleus"],["notch1","transcription"]],
  "type"=>"pathway","datasource"=>"reactome")
```

Figure 3. Notch1 Cytosolic Pathway (Reactome HSA-2122948)

Figure 4. Notch1 Nuclear Pathway (Reactome HSA-2122947)

Enumerate nested surface-marker interactions

Figure 5. Expansion of complexes imputes complex-forming interactions and allows GO annotation of reactants to be mined from NextProt [5]

Multi-pathway scope

Model-selection tasks requiring marker IDs from multiple contexts (e.g. cytosolic, nuclear) can be executed within the context of the integrated graph, as shown below for ligand-receptor interactions in the Notch pathway, including the transcriptional targets of Jag1.

Figure 6. Integrated signaling graph for human Notch1.

Package Workflow

Workflow begins with keyword-based search of knowledge graphs from the PathwayCommons [6] online database, which aggregates annotated pathways in BioPAX [1]. Individual pathways are integrated via common molecular identifiers. Once this integrated graph is constructed, PCquery traverses the graph and further labels all vertices with annotations from other online databases, including gene, transcript, and GO terms from NextProt [5], and inter-species gene and transcript orthologs from OrthoDB [4].

Figure 7. A) query BioPAX pathway repository based on organism and keyword search. B) for each annotated pathway (blue,yellow) recursively impute complex-forming reactions (RHS blue arrow) based on existing product annotations (green) and their BioPAX-annotated constituents (purple,blue). C) Integrate augmented pathways (blue, yellow) based on common molecular identifiers (green). D) Mine inter-species pathway markers based on sequence-homology between reference integrated pathway protein gene loci and the gene loci of the target species. E) Enumerate signaling trajectories for integrated pathways. Query gene regulatory interactions downstream of cell-surface interactions based on GO annotation of integrated pathways.

Interspecies ortholog expression

Figure 8. Mapping the gene expression of a single zebrafish neural crest cell to human orthologs in the integrated Notch1 pathway from human

References

[1] Emek Demir, Michael P Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, Peter D'eustachio, Carl Schaefer, Joanne Luciano, et al. The biopax community standard for pathway data sharing. *Nature biotechnology*, 28(9):935–942, 2010.

[2] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1):D649–D655, 2018.

[3] Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, et al. Kegg for linking genomes to life and the environment. *Nucleic acids research*, 36(suppl_1):D480–D484, 2007.

[4] Evgenia V Kriventseva, Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosé Manni, Renata Dias, Felipe A Simão, and Evgeny M Zdobnov. Orthodb v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic acids research*, 47(D1):D807–D811, 2019.

[5] Lydie Lane, Ghislaine Argoud-Puy, Aurore Britan, Isabelle Cusin, Paula D Duek, Olivier Evalet, Alain Gateau, Pascale Gaudet, Anne Gleizes, Alexandre Masselot, et al. nextprot: a knowledge platform for human proteins. *Nucleic acids research*, 40(D1):D76–D83, 2012.

[6] Igor Rodchenkov, Ozgun Babur, Augustin Luna, Bulent Arman Aksoy, Jeffrey V Wong, Dylan Fong, Max Franz, Metin Can Siper, Manfred Cheung, Michael Wrana, et al. Pathway commons 2019 update: integration, analysis and exploration of pathway data. *Nucleic acids research*, 48(D1):D489–D497, 2020.

[7] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[8] O Shaya and D Sprinzak. From Notch signaling to fine-grained patterning: Modeling meets experiments. *Current Opinion in Genetics & Development*, 21(6):732–739, December 2011.

[9] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, April 2005.