PHENOSCAPE

Advisory Board Teleconference
Dec 4, 2012

BACKGROUND

Principal Investigators:

- Paula Mabee (University of South Dakota)
- Todd Vision (University of North Carolina at Chapel Hill)

Co-PIs and Co-Is:

- David Blackburn (California Academy of Sciences)
- Judith Blake (Mouse Genome Informatics, Jackson Laboratories)
- Hong Cui (University of Arizona)
- Hilmar Lapp (National Evolutionary Synthesis Center)
- Paul Sereno (University of Chicago)
- Monte Westerfield (ZFIN, University of Oregon)
- Aaron Zorn (Xenbase, Cincinnati Children's Hospital Medical Center)

Other Senior personnel:

- Jim Balhoff (Software developer)
- Wasila Dahdul (Senior Curator)
- Nizar Ibrahim (postdoc, University of Chicago)
- Peter Midford (Taxonomy Ontology Curator)

The Phenoscape project is funded by a collaborative NSF award to U South Dakota and U North Carolina (grant numbers DBI-1062404 and DBI-1062542) for a total of $3.3M for 48 months (July 1, 2011 to June 30, 2015).

The goal of Phenoscape is to generate hypotheses about candidate genes for evolutionary novelties by semantically integrating data on phenotypic variation among species and the phenotypic effects of genetic variation in model organisms. More broadly, Phenoscape is exploring opportunities for computer-assisted knowledge discovery in the field of evolutionary developmental biology (evo-devo). In the first round of NSF-funding for Phenoscape (2007-2011), we developed a proof-of-concept that culminated in a working knowledgebase (KB, at http://kb.phenoscape.org) with over 500K phenotype assertions. The evolutionary data covered the clade of ostariophyan fishes and the genetic data was that available for zebrafish from ZFIN. In the course of the project, the team developed several useful resources, including multiple ontologies (including a multispecies anatomical ontology and a taxonomic ontology), curation software for semantic annotation of evolutionary phenotypes, and contributions to a ontology-driven database platform.

The aims of the current award include:

1. Curation of data for a much broader taxonomic clade - living and extinct vertebrates and inclusion of genetic data from multiple vertebrate model organisms (mouse and *Xenopus*, in addition to zebrafish). We anticipate that, at completion of this project, the knowledgestore will contain 2.5 billion phenotype assertions, and in the process we will have developed several new and expanded anatomical and taxonomic ontologies as community resources.
2. A scalable workflow for data annotation and ontology development, incorporating Natural Language Processing (NLP) to automatically generate candidate ontology terms and phenotype assertions from text. In the process, we will produce enhanced curation software.
3. A framework to capture and reason over homology. We aim to create a logically rigorous and generic methodology to reason over assertions of homology and annotate homology assertions from the literature for the vertebrate skeleton.
4. Fast semantic similarity search. We aim to develop a statistical methodology coupled to software that leverages ontology structure and patterns of data annotation to discover phenotypically similar profiles between organisms and/or genotypes.
5. An expanded Phenoscape Knowledgebase that integrates the ontologies, homology assertions, phenotypic assertions, genetic data, semantic search and interfaces to support an expanded set of use cases that leverage this integrated resource.
6. As a capstone, we will validate the capabilities of this suite of tools by testing how well it identifies known developmental pathways for well-studied skeletal transitions in vertebrate evolution and how well it scales to a datastore containing billions of phenotypes.
7. Outreach activities include:
    - Summer internships in bio-ontologies for undergraduate/graduate students, in partnership with the DeepFin RCN.
    - A junior curator program for advanced high school students through Project Exploration (http://projectexploration.org)
    - Undergraduate internship and community outreach to the Lakota population through the Getting American Indians to Information Technology (GAIN-IT) at U. South Dakota.

In addition to these specific aims, we would like at the conclusion of the project to have a community of passionate knowledgebase users and thus be in a position where we can make the case to stakeholders for sustainability of the knowledgebase; support efficient curation of additional taxonomic groups and model organisms (e.g. plants, insects); incentivize contributions of data from individual researchers; have community participation in maintenance of the ontologies and software. We welcome guidance from the Advisory Committee on how we can tune our strategy to achieve these aims.

Information and Links:

- Website: http://phenoscape.org (most content world readable, meeting minutes and select pages only accessible to project team)
- Phenoscape Knowledgebase: http://kb.phenoscape.org
- Blog: http://blog.phenoscape.org
- Mailing lists: see http://phenoscape.org/wiki/Contact

AGENDA

1. Introduction (Paula Mabee)
2. Project updates
   - Curation workflow (Hong Cui, Jim Balhoff)
   - Ontology development (Melissa Haendel, Peter Midford)
   - Phenotype curation (Alex Dececchi, Monte Westerfield)
   - Homology reasoning (Hilmar Lapp)
   - Semantic similarity (Todd Vision)
   - Outreach (Nizar Ibrahim and PM)
   - Knowledgebase development (JB)
3. Upcoming plans (Todd Vision)
4. Q&A and discussion (Board)

MINUTES

Teleconference participants

- Advisory Board: John Day-Richter, Brian K. Hall, Cyndy Parr, Paul Schofield, Peter Vize, Alan Ruttenberg
- Phenoscape project members present: Paula, Peter, Terry, Jim, Aaron, Hilmar, Todd, Alex, Melissa, Monte, Judy, Hong, Nizar
- Phenoscape project members not present: David B., Paul S., Chris M., Wasila D.

Slides:

PDF: http://phenoscape.org/wiki/File:ABmeetingSlidesDec2012.pdf

1. **Introduction (Paula Mabee)**
   - Self introductions of Advisory Board (AB) members (new advisory board member, Cyndy Parr) and new project team members (postdoc Prashanti Manda, to start Feb 2013)
   - Recap of Phenoscape history, goals; Spring 2012 AB recommendations to be addressed here by project team members

2. **Project updates**
   - **Curation workflow - NLP** (Hong Cui)
     - (Schofield) How are you dealing with atomic terms (such as unossified) to turn this into an EQ expression - to pull this apart

on an automated basis? Some terms may only be able to be done through human curation.
    - Hong - Indeed this is rather difficult. Taking cues from how humans curate these terms.
  - (Alan) A question is where such pulled apart information is kept? Typically you would have a lexicon to capture information about both automatically and humanly determined composite terms, and then the parsing code would use the lexicon as a resource.
    - Hong - That is right. Keeping this knowledge in one place is important. We will develop a rule-bank like module to hold those heuristic rules, so if some rules change, the algorthm/CharaParser will act accordingly.
- **Curation workflow - Term broker** (Jim Balhoff)
  - (Day-Richter) Where to find documentation (more detailed technical specifications) about this? For example, use of backed and authentication technology.
    - Jim - Doesn't exist - not much that's worth pointing to - will take this as an action item.
  - (John) Interested in backend; will contact Jim via email
- **Ontology development - Anatomy** (Melissa Haendel)
  - (Parr) How large is the community around the ontology curation tool development.
    - Melissa - Really just looking at 4 people for curating the phenoscape-ext.owl file, but UBERON community overall is much larger (~50 people submitting terms/editing the ontology)
  - (Alan) Congrats on adopting OWL2/P4 successfully.
- **Ontology development - Vertebrate Taxonomy Ontology** (Peter Midford)
  - (Ruttenberg) Should make equivalence axioms rather than cross-references for NCBI terms that are trimmed out due to redundancy.
  - (Parr) Let's talk later about ways to leverage EOL and Global Names, etc. for taxonomy ontology (include Jim)
- **Phenotype curation - Evolutionary Phenotypes** (Alex Dececchi)
  - (Ruttenberg) Are the PATO term additions available in one place? PATO may need some rethinking. Would be willing to look into for where our terms should best go.
    - Some on PATO tracker; some on the ORB (adding to PATO pending)
- **Phenotype curation - Model organisms** (Monte Westerfield)
  - (Ruttenberg) Some of the EQs looked like they were only partial for mappings of MP to EQ.

    - There are indeed some MP terms that are not completely described by their EQ mappings, though they should suffice for Phenoscape's purposes.
  - (Schofield) How about issues with sets of digits? How far have you got with mapping HPO for limb phenotypes - this has proven to be very difficult. Is UBERON being used for all of these mappings?
    - There haven't been issues with sets of digits for MP. For HPO mapping to logical definitions, this is on the agenda for the LAMHDI project in the near future, but hasn't been tackled within Phenoscape yet.
    - Melissa-automate pipeline, uberon in combination with logical combination to decompose HPO. Have on LAMHDI agenda to help with HPO with logical definitions
    - Monte-we have embraced the human model organism, but besides what Melissa has discussed we have not addressed it very much
- **Homology reasoning** (Hilmar Lapp)
  - (Hall) Homologizing elements within a limb; are we homologizing between fore and hindlimb? How are we doing this, serial homology?
    - yes, we are distinguishing between phylogenetic and serial homology. Elements within a limb or between limbs of the same organism would, where appropriate, have serial homology asserted.
  - (Ruttenberg) What does it mean that there needs to be an OWL model as well as one in RDF? I would think it one or the other?
    - For now the OWL model would be directly loaded into the (RDF-based) KB. It's still possible that the OWL will need to be transformed for better KB performance, but not expected as of now.
- **Semantic similarity** (Todd Vision)
  - (Schofield) Would appreciate more detail (perhaps before the April meeting) on the medianIC and the semantic similarity reasoning and scoring.
    - Todd will share this.
- **Outreach** (Nizar Ibrahim and PM)
  - (Hall) Are there plans for the coming year to attract more students through the outreach program?
    - Won't have more students at U Chicago due to equipment and resource constraints. Same at USD.
- **Knowledgebase development** (JB)
  - (Schofield) What is pre-reasoning?
    - We create fairly complicated OWL expressions, which can be difficult to query with a simple RDF reasoner.

Therefore, create some hierarchies of subsuming class expressions in advance.

3. **Upcoming plans (Todd Vision)**
    - No comments/questions specific to this.

4. **Questions from the Advisory Board/Discussion** (Board)
    - (Hall) Have we considered capturing data on gene regulation for model organisms rather than just gene expression? (noted in literature that gene regulation in medaka differs from zebrafish)
        - Monte - This is question for MODs. Gene regulation and regulatory element data are beyond the resources of the MODS to capture or curate at the moment. Note that Gill Bejerano, one of the first people to develop automatic methods for finding regulatory elements in genomes, is attending the Feb Phenoscape workshop.
    - (Schofield) Impressed with how the project has been addressing the feedback from the AB. Concern was primarily on "pushing the meat", and that's what seems to have a lot of forward movement. Utility is 90% of finding ways for sustainability.
    - (Hall) Integrating the Model organisms and the annotations required a huge amount of work.
    - (Day-Richter) How are we thinking about creating communities of users and incentivizing contribution. What measures have you taken so far. Have we received annotations from outside? Will send more questions by email.
        - Todd - Having all curation centralized benefits quality as MODs show, but obviously isn't very scalable. We haven't tried to crowdsource annotations yet. The attribution of curators by their ORCIDs is the beginning of our thinking about nano publications, but going beyond that is still pie in the sky.
        - Paula - We have people approach us wanting to curate data. However, curation process and adding terms to the ontologies is still rough for people who aren't well trained first. Not suitable yet for naive curators to jump in and curate.
    - (Parr) Would encourage to think about where we are with respect to MODs and what that can tell us about scaling. Are our current methods really going to be working a few years down the road.

Action items

- Jim add system documentation for ontology broker (Alan)
- Alex, Paula: Send Alan R, the PATO term requests added + ORB request link
- Todd: share medianIC methodology/results with Paul
- Jim add documentation for OWL reasoning system prior to next advisory board meeting

- Terry: collect the cases that are a little more difficult -so we don't lose that experience (Alan)
- All: Try google hangouts! (John)

Source:
http://phenoscape.org/wg/phenoscape/index.php?title=WG:Advisory_Board_Meeting_2012-12-04&oldid=10428