**Phenoscape Project Meeting**

**NESCent**

**Wed April 25 and Thurs, April 26, 2012**

**Wednesday morning**

Slides:

\* Homology WG: http://dl.dropbox.com/u/1288730/Homology%20WG%20update.pdf

\* Semantic similarity (aka Phenoblast) and capstone slides: http://Phenoscape.org/wiki/wiki/File:Semsimilarity\_apr2012.pdf

\* Knowledgebase development: http://dl.dropbox.com/u/6704325/Phenoscape%20all-hands%20April%202012-KB.pdf

**#1 Monte: Model Organisms ontology dev and data curation**

a. expand and maintain phenotype (limb and fin) ontologies

b, curate the literature + capture data from large scale phenotype projects

## milestone 1. curate phenotypes and gene expressions

ZFIN 61000 phenos for 4k genes

800 fin/limb phenotypes

(future) include gene expression data

MGI 6100 limb phenotypes for 2500 genotypes, curated as MP terms, MP mapped to EQ and this is complete and there is a mechanism to keep up to date

(future) continue curation, maintain mappings, get emap with data released, GET DATA INTO Phenoscape

Xenbase - XAO complete and current, VAO terms added to XAO

gene expression curated from all 123 legacy papers, 80 genes

(future) continue curation, import data into Phenoscape

## milestone 2. incorporate into KB

Challenges

* MGI - anatomy ontology is organized by developmental stages; close to emap with developmental staging information captured with stage ranges. Immediate future goal is to release emapA version to Phenoscape within a few months.
* Xenbase - backend support for phenotypes unfunded. 2. states of anatomical elements

Discussion

1. Paula, how do we focus work on intersection of gene expression at similar stages that have correlated with phenotypes that relate to fin/limb transition.

2. Gather all genes from mouse, fish, Xenopus that have been examined.... what is the intersection of genes?

3. Hoxd13 discussion

**#2 Paula: ontology development and data curation (evolutionary)**

a. dev and align vert anat / focus on limb/fin and transition

- decided to start with initial individual focus on teleosts, amphibians, amniotes (mostly skeletal)

- VAO (not including FMA)

-- Agreed on uniform terminolgoy for limb terms...added to VAO

-- need to add in Human (obtain extra funding )

-- consensus big / 'settled' between evol and mod groups

- XAO/AAO synchronization, April 2012

- ref VAO for MODs

--done for ZFA

--XAO and MGI in progress

- incorporate new terms as needed into CL

b. develop support VAO

c. immediate plans

-- move to Protege4 with OWL

--ontological modularization strategy

--single VAO

---scope unclear

---requires lots of knowledge about term taxonomic distrubution

---frequent need to obsolete and replace duplicate terms in subontologies

---rationale includes speeding up workflow, expert subsections can be combined and support

--- need to capture information in comments/notes including authority data

d. UBERON can we use it for annotation

c. Phenex updated to allow for collaborative editing using file synchroniation software

d. 66 high-priority papers for curation, ~20 for fishes, ~40 for amphibians (many for early tetrapods), ~4 for amniotes/archosaurs

**#3. AAO report from David**

- approx. 2100 terms, about 10% xref'd

- AAO still has 200 unclassified terms, but progress made on some in Feb 2012

- expert curated by a community of people, so lots of quality terms but lacking many definitiions and many relationships are inaccurate

- now integrating with XAO, xrefs, adding terms Eric Segerdell contributor

- XAO - 1500 stage relationships, complete for definitions and is\_a complete

**#4 Peter Midford VTO report**

a. use NCBI taxa, Paleo DB, IOU checklist...

b. use Catalog of Fishes and Amphibia Web to supplement TTO and ATO that feed into VTO. (Taxa 101k, species 89k, synonyms 102k)

c.(future) Tighter integration with NCBI taxonomy; using NCBI ID s as primary ID s

- where name match use NCBI ids as alt-ids

d. (future) define workflow to update VTO more frequently

c. (future) moving to OWL and taxa as individuals

- (future) integrate with AAO/VAO

**#5 Scalable Workflow - Hong**

- Convert character information to EQ phenotype expressions

- Objectives

- automatic extraction of terms from character descriptions

- semantic matching extracted terms to ontology terms faster and more accurately

- status: investigating alternative algorithms

- automatically generate EQ statement from character description

- status: prototype system demonstrated at biocreative

- NLP integration with Phenex

- Semantic Term matching: Similarity-Vector and Similarity-Lesk

- requires premade glossary; fails to distinguish synonyms from antonyms

-Internal Evaluation Result

- Term (original text) based EQ 90%, label (ontology) based EQ 51%

-Biocreative Evaluation Result

-

CharaParser Effect

Inter-Curator Agreement

Factors

- Development corpus size (3 papers)

- System not well configured for Sereno style curation

- standard parsing did not work with Sereno style curation

- special Sereno parsing worked with Sereno style pdf, but Gold standard testing did not

allow triggering Sereno parser

- Gold standard used in BioCreative

- Diagram for future integration with Phenex and Ontobroker

- Q: how does Phenoscape direct requests to Bioportal, PATO system, etc

- A: bulk vs. single term additions

- Discussion of curator consistency and granularity issues

**#6 KnowledgeBase Development - Jim**

Motivation

- Integrate data and enable discovery through multiple interfaces

Overall objectives

- Supporting semantic similarity search and homology reasoning

- Integrating gene expression

- Support for correlated phenotypes

- Enable open-ended data mining and repurposing of KB

Participants - Everyone is involved in feedback

Progress to data

- Phenoscape I (kb.Phenoscape.org)

- Integrated KB

- Web services API

- Web UI

- Official release candidate now up

- Other Progress

- translating incoming data to OWL

- reasoning pipeline for precomputing inferences with slices of data and ontologies

- loading OWL statements into Virtuoso RDF triplestore

- testing performance using SPARQL

Trello board - immediate future plans

Challenges

- Implementing new backend technology while maintaining current KB interface

- Backend needs to support broader project goals

- Attracting users to the web UI

A Really Useful web user interface

- Objective: enabling discovery

- Refocus the UI on character descriptions, but navigate and retrieve using ontologies

- Does any user want to compose an EQ expression and receive lists of ontology

terms in return

- Lessons from the current KB

David B: correlated genotype question is not the ultimate selling point for Phenoscape

Need great use cases

Aaron: Need to start publishing papers and demonstrate utility of system

Paula: review of Phenoscape I user needs workshop

Judy: Connecting to human disease phenotypes

Making the model organisms continue to be relevant to NIH's focus on human disease

Monte: but need to expand evo use cases, too

**#7 Homology reasoning - Hilmar**

Summary of April 2012 Chicago meeting

Motivation

- Terms are defined in terms of structural criteria, no evolutionary history

- have default homology already - phenotypes inhering in the same entity are considered related or the same if quality is the same (same name -> same structure)

- but what about homonyms?, synonyms?

Objectives

- create a formal framework for stating which entities are "the same"

- collect a comprehensive collection of cross-vertebrate fin/limb homologies

- OWL2 framework

Participants and Roles

Hilmar: coordination and reasoning

Chris: reasoning

Paula, David, Paul, Nizar: collecting use cases

Progress

- logical model for phylogenetic homology

- test dataset (in progress)

- test for formal framework validation

Chicago meeting

- Flesh out use cases

- converge group-specific AOs to a single vertebrate AO

- Homonymous terms to have have only a single class, with evolutionary meaning expressed through homology assertions

- Defining "sameness" and necessary user-control for discovering phenotypes is challenging

Trello board

- reasoning use cases

- finishing rules for handling default homology

- iterative homology

- testing and validation

Chris: we can use vocabulary from BGEE as well

Challenges

- how to handle and deduce default homology in a scalable manner

- what explanations and control do users need for homology-based inferences

- how to treat homology sameness for semantic similarity

- How to assess correctness and comprehensiveness of our model

**LUNCH WED**

**8. Semantic similarity (Todd)**

profile = set of phenotypes associated with a genotype or taxon(lineage)

motivation: find similar profiles among lineages, genotypes, or between evolutionary lineages and genotypes

phenoblast now not conceived of as a separate tool, but encompassed by kb

similarity issues:

- terms need not match lexically

- there need not be a one-to-one matches between phenotypes for biological and methodological reasons

- similarity between common phenotypes is less informative

- match in quality alone not meaningful (entity most important in searching across g-evo profiles

**9. Outreach (Nizar, Paula)**

-centered at U Chicago, Project Exploration module 'Junior Biocurator' (based off of Junior Paleontologist, a proven successful program)

- 6 weeks, financed by Phenoscape

- showed possible new interface/splash page for KB

- Junior biocurators will take images of a small set of taxa (5-10 standard species, e.g., chick,), learn photoshop; could add images that would be displayed in such an interface; will give final presentation;

- Jim: interfaces look great for what they show; could they expand to include views of the data and cut across characters and genes?

- Monte: have representative figures for anatomical terms in ZFA; special workshop this summer;

- David: discussed Digimorph scans

- Judy: each summer has 2-3 interns who do exploratory data analyses to answer a particular problem. Not like this;

**Wednesday afternoon BREAKOUTS (ranked by \* below)**

* KB use cases: who needs this? +human phenotypes for >MOD use; need to reassess use cases! \*\*\*\*\*\*\*
* new interface ideas from usability testing and from Chicago & new Phenoscape II needs, including images from junior biocurator program; Incorporation of images into ontologies, UI\*
* branch profiles & semantic similarity search (Jim)\*\*\*\*\*\*\*\*\*
* integration of CharaParser w/ Phenex (Hong, Jim, Alex, Nizar)\*\*\*
* \*biocreative results and next steps for scalable curation research and CharaParser integration\*

**DISCUSSION**:

**KB use cases:**

No users; where to go? Wouldn't it be cool if....? hold a workshop?

Aaron: need to publish for exposure; need to identify individuals to use this and initiate collaboration with them;

- difficult to use the KB

- people need to see the end product; not many people will spend the afternoon working on this;

- collaboration - ask what they want to know, help them

- onus on us to identify people who might be interested

- Gloria etc. not using it

- MOD folks not interested because no human data

- not enough devo-evo folks in biomed community

Aaron: could imagine finding candidate genes across all taxa in kb, but can already find by going to mod databases

- David: users would come at this from the standpoint of biodiversity and species descriptions

- Suzi: phenotype variation tree should be up front; make it more appealing;

- Aaron: workshop should end with publication

- has to show that it is more useful than a google search --

- evo-devo folks are very interested when they are exposed

- Suzi: need multiple carts for saving search results; download report based on these results

- get gene network people involved, because networks are changing;

- contact 12 people who are interested -- can we answer a question for you? At this point, too new for them to use. Do the work and write the paper;

- Monte showed cytoscape demo of William's syndrome phenotypes with likely mouse and zebrafish genes; MD’s are very interested and going to google to understand;

- gets at complex phenotypes, but evo-devo too small; maybe genetics folks; qtl and pop bio people;

Judy: demo'd gene weaver -- can add own data set and do private analysis; based on gene sets; value of statistic tools in the system, applied to multiple data sets

- Paula: workshop with invited possible collaborators

Suzi: need two workspaces so you could do set operations

--bioconductor and intermine -- repurposing of cytoscape, reuse grid-display from geneweaver (do phenome map; set of genes and pubs);

Judy: outcome of this conversation: need to intersect Phenoscape with other group tools; can create private domains; make it a recognized module/value-added in context of other system; http://geneweaver.org/

Facebase interaction?

Need workshop that targets several communities:

Purpose of workshop:

-use cases to drive development; Brian storm broadly

-invite software developers to build/connect their tools to Phenoscape

Need immediately:

1. ability to download results

2. be able to save more than one search (and feed into other programs such as cytoscape or geneweaver) or do set operations e.g., to see what is in common.

Jim: Output format? what are they? see cytoscape and geneweaver --

Monte: wants to know WHY people want the tools

Todd: we could each seek a collaborator or broadcast for collaborators

**Thurs morning I**

* Anatomy ontology: relationship of Uberon and vertebrate anatomy ontologies
* VAO scope:
* currently skeletal system, muscular system; will bring together VAO, TAO and AAO terms (~2,500 in TAO, ~2100 in AAO); Uberon is currently tracking VAO (align with); they will obsolete
* TAO musculoskeletal becomes VAO; AAO musculoskeletal becomes VAO;
* bring in other (non-skeletal) systems from Uberon into VAO
* stewardship: within Uberon
* Editing control: Phenoscape curators, uberon curators
* Melissa summary of what was on the board:
  + Uberon has an overlapping representation of musculoskeletal terms with VAO that have been partially reconciled. We will figure out a better strategy for this integration. for now, calling this part of combined set uberon/vao.
  + TAO/skeletal terms have not really been put into VAO, but are applicable to vertebrates. Same with XAO.
  + TAO doesn't really have much in terms of nonskeletal elements that aren't already in uberon or XAO. It does have a lot of skeletal elements that need migration into uberon/VAO.
  + AAO has maybe @200 terms that need migrating to uberon, and a similar number (?) as TAO that need migrating/reconciling with uberon/VAO
  + Uberon has xrefs to ZFA and XAO, and some vice versa
    - Uberon will integrate with musculoskeletal representation (technical details to be finalized)
    - AAO and TAO will no longer be supported, and the nonskeletal terms will go into uberon/MODs, skeletal terms into uberon/vao
* Monte/Jim/Judy/Peter/Aaron: sketch out pipeline for ZFIN, Xenbase, mgi data into Phenoscape; including MGI staging issue\*\*

**Thursday**

**1. Model organism expression data and gene homology (BREAKOUT)**

\* Monte/Jim/Judy/Peter V./Aaron

\* Sketch out pipeline for ZFIN, Xenbase, mgi data into Phenoscape; including MGI staging issue\*\*

\* Jim will be reworking the pipeline for import of ZFIN data to make it OWL compliant. When this is done, he will use the same mechanism for MGI and Xenbase.

\* MGI probably already provides phenotype and expression data in an equivalent format from their download pages

\* Xenbase can provide equivalent data for gene expression at this point, phenotypes in the future

\* Jim will spend some time right away examing the MGI and Xenbase files to ensure he understands the format and that all the data he needs are accessible;

**ACTION ITEM**: Jim get back to Judy and Aaron if they need to generate a new report; contacts are both Judy and Terry for MGI, Peter V. for Xenbase

\* For gene expression, we will limit support at least initially to in situ data from the MODs, i.e. rather than GEO data, because they are annotated with the anatomy ontologies

\* We also discussed the issue of orthology:

\* Does Phenoscape need to pay attention to orthology and, perhaps, provide information (assertions) of orthology?

\* HomoloGene might be sufficient for a first pass

**ACTION ITEM:** Judy and Monte might do some comparisons of HomoloGene predictions of mouse/zebrafish orthologs versus the ZFIN hand curated orthology assertions

* Xenopus is not yet represented in HomoloGene.

**ACTION ITEM**: Aaron will contact Richard Harland (who has genome annotation) HomoloGene to get Xenopus included

\* Phenoscape will take responsibility for getting homology assignments into the KB

\* We think we could write a paper about this process of combining gene expression and phenotypes from three organisms, mapping orthology, agreeing on common formats, "Maps in the Forest"

**2. Anatomy ontology notes (BREAKOUT)**

Upcoming presentations:

iEVOBio (July 10-11), bazaar presentations only

Joint Meeting of Ichthyologists & Herpetologists in August 2012: Paula, in Vancouver BC

Society of Vertebrate Paleontology in Raleigh, NC in October; both Nizar and Alex presenting

Phenotype RCN on behavior in February 2013

**PUBs**

1. Suggestion to do a Database paper - deadline due soon

2. Paper from Phenoscape I

3. Post launch announcement for KB: jonathan eisen; zimmer; page;

4. OBD

5. VAO bone paper(s)

**ACTION ITEM**: suggestion to monitor tweets for 'Phenoscape' (clipping service)

**ACTION ITEM**: Need to write up guidelines for authorship

Publication model: opt-in model: present plan (e.g., wiki page and author listing), transparency; use 'Phenoscape consortium' if limited in number of authors;

Posters

**ACTION ITEM**: Idea (suzi) to have a hackathon to develop tools to visualize phenotypes; gmod, rcn, nescent;

* **Other to dos (not discussed enough)**
* homology! flesh out use cases and requirements
* term broker\*\*\*\*
* discussing ideas for capstone to understand implications for present activities
* taxonomy ontologies: NCBI, PDBD\*\*\*\*\*
* VTO workflow in relation to PBDB and tracker (flesh out wiki page in process)
* consider how to get additional community input into...ontology development and phenotypes (breakout with...) Todd could give nanopresentation on nanoattribution; discuss direct submissions of evolutionary phenotypes; matrix-based vs comparative vs one-off species descriptions; community input\*\* (who would be the submitters?)
* anatomy ontology breakout
* Auto-definition tool - from differentiae and annotations; taxon specific definitions/comments. We need to modify ontology format!
* Support for preferred labels (low priority, Thurs)
* Melissa would like to meet with Jim to discuss issues involved in move to OWL
* All curators: svn organization -- Hilmar, Jim, Melissa (hard to find where source files are kept; need to be able to open AmAO in owl and have it automatically grab vao release version; very unclear)
* Xenbase Phenote config (Jim & VG)
* ALL: discuss Fishbase Consortium; links out to Fishbase glossary and species pages
* ALL: Discuss/plan and list somewhere on wiki the upcoming planned or potential Phenoscape presentations over next year.