Phenotype Ontology Coordination Workshop Report

April 27-28, 2009

Organizers: Paula Mabee, Monte Westerfield, Todd Vision

Phenoscape and NESCent sponsored a 'Phenotype Ontology Coordination Workshop' at the National Evolutionary Synthesis Center (NESCent) April 27-28, 2009. The goal of the meeting was to bring together scientists with interests in computing on phenotype to prioritize strategic community goals for the purpose of developing a RCN (Research Coordination Network) and/or INTEROP (Community-based Data Interoperability Networks) proposal for NSF. We invited representatives (see list in Table 1) from the model organism, evolutionary biology, paleontology, comparative morphology, bioinformatics and computer science communities who were leading efforts that involved comparing phenotypes and with interest or experience in using ontologies to represent phenotype. A phenotype community that includes these participants is new, thus fitting the type of group well supported by RCN funding. The goals of this nascent community, however, are quite unified and specific, thus also well served by an INTEROP grant.

Our objective was to discover what the participants viewed as the specific research questions to be answered and thus goals to be achieved through a community-based grant. As participants introduced themselves, they described their personal research questions/objectives that would be furthered through such funding. This was followed by breakout groups in which the following four questions concerning community needs were discussed.

1. What are the key gaps in tools to develop your work?
2. What are key obstacles to data exchange and data integration?
3. What do you perceive as the key ontologies that are lacking?
4. What are the gaps in expertise in understanding, developing and applying ontologies in the community?

The groups itemized and prioritized the many tools and ontologies that are lacking for the phenotype community, as well as the specific obstacles to data integration and exchange and community buy-in (Table 2).

The tool gaps that were viewed as most critical were specific missing features for annotation and ontology development (Table 2A). Annotation tools currently lack rich pick-list interfaces so that users/curators can easily find appropriate terms from one or more ontologies. Additionally, these tools do not yet facilitate collaborative annotation. One high priority missing ontology development tool feature is an 'ontology request broker' (ORB) that would allow a community member to acquire a temporary ontology term. Annotation could thus proceed without a time lag (waiting for community to vet the term and gatekeeper to add), and without the use of multiple applications. Related to this is a desired feature that would allow multi-ontology aware editing and alignment, i.e. it would allow the ontology editor who is the process of adding a term or synonym to see similar terms from other ontologies and to add new terms or cross reference and align with existing ones accordingly. Such an 'ontology request broker' fits between annotation vs. ontology development and could be plugged into multiple databases. Other high priority ontology development features include ontology alignment tools [term matching] and tools to extract semantic data (terms and synonyms) from the literature. An important tool that came up repeatedly in this workshop was a mechanism to 'phenoblast', i.e. perform cross-database searches for similar phenotypes (Table 2A).

In relation to ontologies, participants identified the alignment among currently siloed anatomy ontologies as particularly important (Table 2B). Formalizing standards and practices for phenotypes and phylogenetic data was also considered high priority. These include standards for representing homology relationships, phenotype syntax, file formats and best practices for ontology development. Several types of key ontologies in the community are lacking. One of these is a common reference ontology for anatomy for a particular taxonomic subset (e.g. protostome animals) with an associated homology model. This type of ontology and associated

homology mapping are required to address a basic search, for example, for corresponding parts among *Drosophila*, trilobites, and beetles. The lack of a homology model was also identified as a critical obstacle to data exchange and data integration (Table 2B). Anatomy ontologies are also required for many groups that currently lack them. A second type of missing ontology is a broad taxonomy ontology. The NCBI taxonomy is comprehensive but missing most extinct taxa, and it is not integrated with others. This is a central point for data integration. For ontologies that describe time (development, ecology, evolution), standard ways of representing its passage need to be worked out. Finally, functional ontologies, including whole organism processes (behavior, biomechanics, physiology) are critical to develop and augment (e.g. the GO).

The key obstacles to data exchange and integration were identified as the lack of skilled personnel and the sociological issues that apply to communities adopting new technologies and ideas. These are tied in part to lack of training, and they include cultural inertia, communication difficulties, difficulties in compromising, adopting standards, etc. (Table 2C).

Finally, the group recognized significant gaps in community understanding of ontologies and their applications. This was viewed as due to lack of skilled manpower and sociological factors (Table 2C) but also to lack of knowledge, because information about ontologies is not a typical part of a biologists' training. The community viewed their peers as not understanding what an ontology is, not knowing what ontology-related resources are available, and most importantly, not knowing what an ontology enables. The development of a 'killer application' that is non-biomedical was recognized as critical in providing the impetus for the community to contribute to ontology development and further applications. Community training is critical to increasing the talent pool, and it must involve a comprehensive, structured documentation of tools, data, and annotation best practices (Table 2D).

It was recognized that remediation of the above missing tools, ontologies, and standards, and community issues must be driven by the desire to address an over-

arching community research question (sometimes referred to as the unifying grand challenge [GC] question). Prototyping a 'killer application' to demonstrate that this question can be most effectively addressed with ontologies, annotations, etc. would be an effective way to evoke interest and enthusiasm in this approach.

The dialogue during the remainder of the workshop was focused on defining specific use cases and generalizing to the driving research questions. Specific grant aims and requirements based on these questions were identified.

Use cases/Driving questions:

The most commonly articulated use case was the analog of a gene blast search, i.e. the desire to find 'matching' phenotypes in other taxa, to explore the nature of the matches (homology, common function), and to build a list of candidate genes from all (model) taxa that possessed that structure to better understand the genetic and developmental bases of its evolution. This use case is elaborated below under 'I. Deep Time Phenotype Matching'. The overarching question for this use case might also be articulated with a different emphasis (e.g. below, 'II. Match or mismatch between phenotypes and genes?'). The specific aims and outcomes are similar.

I. Deep Time Phenotype Matching:

The first question from this use case is: '*What are the structures corresponding to [structure x] in [taxon y] in other taxa (extinct or extant)?* For example, I may study the anatomy of vestigial hindlimb skeletons in boas (snakes) and want to know what the homologous parts are in fishes. I may also want to ask specifically what the vestigial remnants of legs known as 'anal spurs' in boas are homologous to in lizards. I will want to know what the evidence is that was used to assert this homology relationship. I may also want to find all the structures in vertebrates that share the function of anal spurs, and I may want to simply search on 'spur' and see what the semantic matches are.

The second question from this use case is '*What are the candidate genes for that structure?*' The set of candidate genes for evolutionary species can be built from the gene-phenotype mappings from the model organism(s) that possess matching

structures. A related question 'What is the underlying biochemical pathway for a particular phenotype' could be solved similarly, by searching for a similar phenotype from any species where the pathway is known.

Several related use cases were suggested from a molecular approach to the data. One was 'What changes in phenotype are associated with particular changes in a protein/gene structure?' Data-mining for correlations in gene-phenotype data across multiple taxa and databases could potentially in the discovery of some signature for particular types of phenotypic change. Another use case was based in knowing which genes influence the phenotypes in model organisms, and searching for genes in organisms that have no homology statements relating their structures. One could then examine what phenotypic effect those genes have, using those data to assess phenotype homology in that organism.

Requirements: The major goal described here, of querying integrated phenotype ontologies, requires several specific aims. The first is the identification of relationships, such as homology and functional similarity, among anatomical structures from organisms across various levels of phylogeny. Correspondence among anatomical entities may be based on one or more types of similarity: homology (similarity due to common ancestry), functional similarity, and text string similarity (semantic matching). These relationships must be asserted by experts and be linked to evidence codes. Only when these relationships are in place, most likely in relation to a reference ontology at a particular node, can they be used to make assertions about phenotype across the phylogeny of life. Interoperability will require solving the technical problem of mapping homologies between those organisms where there is potential for it (i.e. where there is phenotypic overlap). Developing the tools to search for these similar phenotypes, i.e. 'phenoblasting' multiple databases, and an interface to view results is a requirement for this work.

Expected outcome: Phenoblast may be used to explore relationships among data across the anatomical ontologies and to find relationships among data that are linked to the anatomical structures via the ontologies (e.g., genes, developmental processes, etc.). Phenotypes that are similar due to common ancestry (homology)

can be filtered from similar phenotypes that have evolved independently (i.e. may have only functional similarity). Similarities and differences in their underlying candidate gene set, for example, can then be explored.

Moreover, given that we know how a structure has been shaped over evolutionary time (through phenoblasting over the tree), we can explore how its evolution may have been influenced by climate change by connecting to paleoclimate data. Did past climate changes result in the evolution of functionally similar structures? I.e. at particularly critical climate times, do we see the emergence of similarly functioning structures? Do they have a common or dissimilar genetic basis?

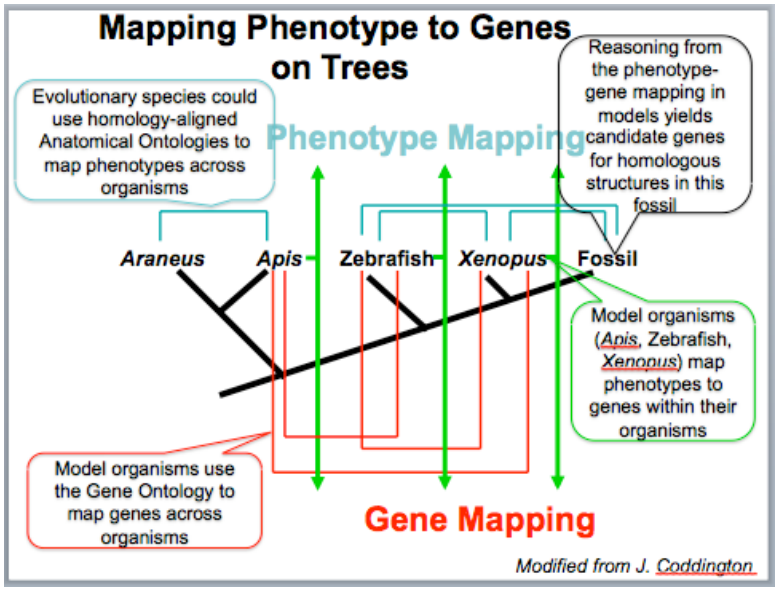
With the addition of an ontology that addresses biomechanical function, one could search for links, e.g., between body form and acceleration in fishes. This would certainly return known form-function links, but may also highlight less obvious associations that would lead to further biomechanical research. If one had ontologies that included wing shape, taxonomy, phylogeny and flight characteristics, one could identify links between wing form and function and track these characteristics over the evolution of a lineage. This could highlight patterns of stasis and innovation in wing design. By adding courtship complexity to morphological diversity, one could query examples in which courtship complexity is inversely correlated with display ornaments. This could challenge longstanding models of the evolution of courtship and ornamentation. Finally, knowledge of homology and knowledge of similarity that is not homologous allows one to identify examples of convergence, which is a powerful tool for exploring the operation of functional constraints.

II. Match or mismatch between phenotypes and genes?

That there is a high level of conservation in genes across disparate organisms is one of the key biological discoveries of our time. The apparent discordance, however, between the high level of genetic conservation underlying varying species phenotypes is not well understood and has not been empirically characterized. It is not clear, in fact, that there is necessarily a discrepancy in 'the phylogenetic reach' of

gene homology vs. phenotype homology. Testing the generally accepted idea that there is a significant difference in level of gene vs. phenotype conservation will provide a major insight into the connection between genes and phenotypes during evolution. Levels of homologous gene conservation can be and have been measured using existing tools (I think), but levels of phenotype conservation have not been to date.

How closely does the phylogenetic extent of phenotype homology match that of gene homology? I.e. is there a difference in taxonomic distance between phenotypic homologues vs. genetic homologues? What aspects of phenotype match closely, which don't? Do they scale with 'level' of homology?



From an evolutionary standpoint, unveiling the *large-scale patterns of phenotypic evolution against the patterns of genetic evolution* is paramount. What are the rules or tendencies of phenotypic change in evolution?

Importance: Comparative biology is based on relationships among entities (e.g., homology, functional similarity, etc.) that are key to evolutionary theory, which in turn is the key to integrating genomic and phenotypic information. Being able to make these assertions and linking structural information will build the foundation

that allows us to integrate genotype and phenotype, integrate extinct and extant phenotypes, function, and further, to integrate climatic changes in relation to all three. Developing the anatomy reference ontologies, homology mapping and phenoblast tool as suggested in this workshop, will not only address a broad question about the evolution of life on our planet, but it will provide immediate practical benefit to many ontology efforts, as they will have a reference to plug into/align with, and they will be able to immediately reference their work to the larger body of phenotypes (extinct and extant), genes, function, and evolution.

Databases: The group considered the set of databases that would enable interoperability. These include model organism databases (such as ZFIN), databases with phylogenetic information (Treebase, Paleobiology Database), databases with character data from diverse organisms such as Spider ATOL, Hymenoptera ATOL, LepTree (these data are not yet tagged with ontologies), Phenoscape, MorphBank, GenBank, and Encyclopedia of Life.

Table 1

Participant	Affiliation
Judy Blake	The Jackson Laboratory
Jonathan Coddington	Department of Entomology, Smithsonian Institution
Lindsay Cowell	Duke University Medical Center
Andy Deans	Hymenoptera Tree of Life and Department of Entomology, North Carolina State University
Betsy Dumont	Department of Biology, University of Massachusetts, Amherst
Eva Huala	The Arabidopsis Information Resource
Hilmar Lapp	Phenoscape and NESCent
Suzi Lewis	Berkeley Bioinformatics and Ontology Project
Paula Mabee	Phenoscape and Department of Biology, University of South Dakota
Anne Maglia	Missouri University of Science and Technology
Austin Mast	Florida State University
Peter Midford	Phenoscape and Dept. of Ecology and Evolutionary Biology, University of Kansas
Cyndy Parr	Smithsonian Institution and Encyclopedia of Life
Greg Riccardi	College of Information, Florida State University
Paul Sereno	University of Chicago
Arlin Stoltzfus	Center for Advanced Research in Biotechnology, University of Maryland
Todd Vision	Phenoscape, NESCent, and Department of Biology, University of North Carolina at Chapel Hill
Peter Vize	Department of Biology, University of Calgary
Monte Westerfield	Phenoscape, Zebrafish Information Network (ZFIN) and Institute of Neuroscience, University of Oregon

Table 2. Complete list of missing features of tools and ontologies and obstacles to data exchange and integration and community involvement. 1

[1]A. Key gaps in tools, the lack of:

- 9-Annotation tools (missing features)
  - that include rich 'pick-list' interfaces for ease of finding terms (=ontology-provisioning interface?)
  - collaboration, command syntax & protocol for data exchange
- 8-Ontology development tools (missing features)
  - ontology request broker (ORB)
  - ontology alignment tools
  - multi-ontology aware editing and alignment
  - ontology term extraction from text including synonyms (=tools for converting non-semantic data)
  - ways to test large ontologies for gaps and/or errors in logic
  - ways to do collaborative ontology building
- 6-Phenoblast -interfaces and mechanisms for cross-database phenotype searches
- 5-Tools to transform legacy phenotype data
- 5-Common repository for annotations
- Automated capturing of annotations and ontology terms and a way to proof them
- Tools to do automated reasoning over primary data
- Visualization tools
- A common ontology editor (obo vs. owl)

B. Key ontologies, the lack of:

- Alignment among anatomy ontologies (currently they are siloed)
- 6-Formalized standards (e.g. file format; phenotype syntax) and practices
- Ontologies:
  - Common reference ontology for anatomy at specific phylogenetic nodes, with associated homology model
  - Taxonomic ontologies and integration with those from NCBI, Paleodb, Species 2000
  - Functional ontologies that include whole organism processes, e.g. behavioral, biomechanical, and physiological axes
  - Ecological/habitat ontologies
  - Multispecies anatomy ontologies for groups of organisms with distinct body plans
  - Ontologies that describe time (ecological, evolutionary) and space
  - Ontologies capturing variation and variability over space, time, individuals
  - Ontologies that capture methodologies
  - Developmental ontology
  - Ontology of evolutionary processes
- Higher-order ontological relations (e.g. homologous_to)

[1] The numbers in front of top items are the number of votes/ranking from our workshop.

Table 2. Complete list of missing features of tools and ontologies and obstacles to data exchange and integration and community involvement.

C. Key obstacles to data exchange and integration, the lack of:

- 8-Skilled manpower, resources, funding
- 5-Sociology of people
    - different world views that are equally legitimate
    - general difficulty in communicating among stakeholders
    - lack of commitment to sustainable, accessible solutions (tools, formats)
    - cultural reluctance to share or change
    - cultural inertia in tools use practices
    - lack of compromise
    - lack of agreement on where to start
- 5-APIs for data silos
- Homology model in relation to ontologies
- Centralized caching for data/genes across repositories
- Repositories for EQ statements
- Repository data exchange standards
- Annotation standards
- GUID for data objects
- Scalability and aggregation of reasoning frameworks
- Lack of intelligent access to services
- Philosophically compatible ontologies

D. Key gaps in expertise in understanding, developing and applying ontologies in the community, the lack of:

- Killer applications that are non-biomedical (no demonstration case that is compelling)
- 4 -Comprehensive structured documentation of tools, data, annotation best practices
- 4-Awareness of what is available
- Understanding what an ontology is and fear that they are not understandable
- Understanding how to integrate or evaluate ontologies
- Understanding ontological logic (biologists)
- Reward/incentives for using ontology (in publications), developing terms, contributing to community digital resources
- Recognition for data curation/annotation
- Understanding the role of ontology gatekeepers