PHENOSCAPE

http://phenoscape.org

Informatics obstacles for integrating evolutionary phenotype diversity with model organism data



<u>Hilmar Lapp</u>¹, Jim Balhoff¹, Cartik Kothari¹, Todd Vision^{1,2}, Wasila Dahdul³, Paula Mabee³, John Lundberg⁴, Peter Midford⁵, Monte Westerfield⁶

(1) US National Evolutionary Synthesis Center, Durham, North Carolina, USA, (2) University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, (3) University of South Dakota, Vermillion, South Dakota, USA, (4) Academy of Natural Sciences, Philadelphia, Pennsylvania, USA, (5) University of Kansas, Lawrence, Kansas, USA, (6) Zebrafish Information Network and University of Oregon, Eugene, Oregon, USA

Motivation

Model organism databases such as the Zebrafish Information Network (ZFIN) use terms from anatomy and quality ontologies to describe the thousands of phenotypes known from genetic studies. By contrast, the vast stores of evolutionary phenotype diversity data are typically rendered in free-form text in journal articles. This makes it very difficult for researchers to ask questions that cut across these two domains of knowledge, such as "what sister clades in nature differ phenotypically in the same way a given mutant genotype differs from the wildtype?"

Semantic integration through computable assertions

"Meckel's cartilage greatly reduced" decreased size Meckel's cartilage Entity (E) Quality (Q)

Model organism databases have achieved interoperable phenotype annotations by expressing these formally in Entity-Quality (EQ) syntax, and using ontology terms for entity as well as quality terms. An EQ statement asserts that the quality Q inheres in the bearer entity E. A logic reasoner can infer additional statements implied by a body of such assertions and

The Phenoscape project aims to develop the tools necessary to facilitate such questions to be asked on a large scale basis, starting with teleost fish. Here we illustrate the obstacles in effectively integrating the knowledge from the two domains, and the solutions we have devised.



Our initial application domain are the morphological, in particular skeletal, characters that are variable among the Ostariophysi, a clade of teleost freshwater fishes containing zebrafish and approximately 9,000 other species (~28% of all known fish species).

Such phenotypes are traditionally reported in free text form as states of comparative characters in data matrices published in

Outgroup and non-teleostean taxa

Amia calva .- KU 2116, one alc. spec.; KU 3383, seven cl. & st. spec.; KU 6956, one alc. spec.; KU 10187, one alc. spec.; KU 16916, two alc. spec.; KU 21261, one dry sk.; KU 21290, four cl. & st. spec.; KU 21290, four cl. & st. spec.; KU 21337, one dry sk.; KU 21338, one dry sk.; KU 21607, one cl. & st. spec. tractosteus macrobeccus .- KU 21317, one dry sk.; KU 21372, one dry sk. Atractosteus spatula .- KU 18537, KU 18540, KU 18545, and KU dry crania.

Caturus elongatus.- MB. f.3851. †Caturus dus.- MB. f.2931. †Caturus brevicostatus.- MB. f.3849 and MB 1.3850. †Caturus sp.- JM SOS3344. †Caturion - MB. f.3848.

episosteus oculatus.- KU 11163, one alc. spec.; KU 21230, one and KU 21370, one dry sk. Lepisosteus os us.- KU 5, one alc. spec.; KU 1695, one dry sk.; KU 1724, one dry KU 2544, one alc. spec.; KU l one cl. & st. spec.; KU 12645, one cl. & st. spec.; KU 16246, twelve cl. & st. spec.; Ku 12645, one alc. KU 20557, two alc. spec.; KU 22216, two cl. & st. spec. Lepisosteus platostomus -- KU

Table

Data matrix of Taxa Set representing 196 characters belonging to fossil and extant taxa. 0, plesiomorphic state; 1-4, apomorphic states; ?, unclear, owing the preservation of the specimens. Leptolepis coryp., L. coryphaenoides; Leptolepides haert., L. haertesi; Leptolepides sprat., L. sprattiformis; Orthogon., Orthogonikleithrus. †Gyri lus hexa f. 773. †P.

_		1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45
1.	†Allothrissops	11121	2?100	000??	2?010	00011	01201	00001	1001?	00010
2.	Amia calva	00120	00000	00000	00000	00000	00100	00001	20000	00000
3.	[†] Amia pattersoni	00000	00000	000??	0?00?	0000?	00100	00001	2000?	00000
4.	†Anaethalion	00001	2010?	000??	2?0?0	000?1	01110	00001	10000	000?1
5.	†Ascalabos	00001	2?1?0	0?0??	2????	0001?	01?0?	0?001	?0???	00000
6.	<i>†Aspidorhynchus</i>	?0020	0?010	0?0??	20020	0000?	00000	00000	100??	00000
7.	†Belonostomus	00020	0?01?	000??	20020	00000	00000	00000	1000?	00000
8.	Chanos	00011	21100	01011	21010	00011	01201	01001	10011	10000
9.	†Davedium	??000	02011	1100?	????0	00000	00100	00000	100??	000?0

List of characters

The st of characters, the analysis of certain morphological characters, and the phylogenetic relationships of c tain teleosts are based on the features listed below. [0] represents the plesiomorphic character state and 1], [2], [3], and [4] the apomorphic character states. The outgroup used to polarize characters includes

[†]We sonulus eugnathoides, Amia calva, Lepisosteus spp., and others in different analyses. Jith the exceptions indicated, characters 1 to 167 are from ARRATIA (1991, 1996b, 1997) or are new

chal cters. Because of the use of different outgroups, characters 26, 27, 28, 36, 76, 77, 78, 92, 122, 124, 125, 126 28, 129, 130, 137, 140, and 157 changed their polarization with respect to ARRATIA (1996b, 1997), and in q her cases, the presentation of some characters was slightly modified (indicated below). Characters 168 to 1 5 are from GRANDE & BEMIS (1998); characters 176 to 191 are taken from PINNA (1996); and charac rs 192 to 196 are from BRITO (1997).

Ethmopalatine ossification in the floor of nasal capsule articulating with autopalatile. [0] absent; [1] present. (PATTERSON & ROSEN 1977.)

Two paired endoskeletal ethmoidal ossifications: [0] absent; [1] present.

Arratia_1999.xml

Term Info: Amia calv

Amia calva

🗆 Basic Info

□ Synonyms (12)

Links (1)

has_rank

Term: Amia calva

RELATED Amia canina

Amia ornata

Amia viridis

Parents

Children

is_a <u>Amia</u>

Other Properties

a, 1999. Zoological

Catalog ID

2116

3383

6956

10187

16916

21261

21290

computed with an

extensible reasoner

in the database.

+ +

Collection

Parietal bones fused in a median expent: [0] absent; [1] present.



the respective ontologies.

The same formalism can be applied to the descriptions of natural (evolutionary) phenotype diversity in the systematics and taxonomic literature, thereby

making it possible to link divergent phenotypes to their genetic underpinnings.

(~ _ O ×

the systematics literature.

	ha ha a a a a	Data	esec, entre Log, Matrix				_
Set Error Log Matr							×
Taxon	te Intestine coiling	. Parietal length	Posterior marginof maxilla	Symplectic	5	Solid erichordally ossifie	~
5 Chanos chanos	absent	relatively short parietal	convexely rounded or straight	does not art	iculate with lower jaw a	absen	
6 Clupea harengus	?	?	?	?	1	?	h
7 Coilia nasus	?	?	?	?	1	?	
8 Denticeps clupe	absent	as width as long	convexely rounded or straight	does not art	iculate with lower jaw a	absent	
9 Dorosoma cepe	?	?	?	?			2
10 Dorosoma pete	?	?	?	? 7			
11 Elops affinis	absent	as width as long	convexely rounded or straight	does not a			্
12 Elops hawaiensis	absent	as width as long	convexely rounded or straight	does not a			
13 Elops saurus	absent	as width as long	convexely rounded or straight	does not a			
14 Engraulis encra	absent	?	convexely rounded or straight	does not a	la na m		_
15 Engraulis ringens	absent	?	convexely rounded or straight	does not a	Valid Taxon A	Publication Taxo	n
16 Engraulis sp.	absent	?	convexely rounded or straight	does not a	Albula vulpes	Albua vuipes	ric
17 Esox americanus	absent	?	convexely rounded or straight	does not a	Ania calva	Anosa critysocrito	115
18 Esox lucius	absent	?	convexely rounded or straight	does not a	Chanos chanos	Chanos chanos	
19 Esox niger	absent	?	convexely rounded or straight	does not a	Clupea harengus	Clupea harengus	s
20 Ethmidium mac	?	?	?	?	Coilia nasus	Coilia nasus	
21 Gonorvnchus ab	?	?	?	?	Denticeps clupeoides	Denticeps cluper	oid
22 Hiodon alosoides	present	relatively long parietals.	convexely rounded or straight	does not a	Dorosoma cepedianu	m Dorosoma cepec	dia
23 Hiodon tergisus	present	relatively long parietals.	convexely rounded or straight	does not a	Dorosoma petenense	Dorosoma peter	nen
24 Lile stolifera	2	?	?	2	Elops affinis	Elops affinis	
25 Lycengraulis oli	2	?	?	2	Elops hawaiensis	Elops hawaiensis	5
26 Megalops cyprin	absent	as width as long	convexely rounded or straight	does not a	Elops saurus	Elops saurus	
27 Oncorhynchus c	absent	2	convexely rounded or straight	does not a	Engraulis encrasicolus	s Engraulis encras	ICO
28 Oncorhynchus	absent	2	convexely rounded or straight	does not a	Engraulis ringens	Engraulis ringens	S
29 Oneariichthye hi	absent	relatively short parietal	concave or at least notched	does not a	Esox americanus	Esox americanus	5
30 Opsariichthys pl	absent	relatively short parietal	concave or at least notched	does not a	Esox niger	Esox nider	
so opsaniciturys pl	absent	relatively short parietal	concave of acleast flottined	does not a	Ethmidium maculatun	n Ethmidium macu	ulat
	10	Di	splay State Symbol		C		
isplay Publication Nam	e 📘 Display Chara	cter Description	splay State Description			-	_

We developed a tool, **Phenex**, to accelerate transformation of character and state descriptions into ontology-based phenotype assertions.

The human expertise needed for this process (data curation) is the greatest bottleneck in our approach. Phenex records the publication and specimen codes, aids in choosing the

link_node_id_fkey

0 0

Chordata

respective ontology terms for taxon, character entity, and character state, and supports annotation workflows that optimize the time required from experts.

Knowledge Base

Catalog of Fishes, and the multiprovided object quality of a single physical Species Teleost Anatomy Ontology physical object quality 🕀 👩 cellular quality (TAO) as a clone of the Zebrafish 🗄 👩 complexity Composition 🗄 👩 functionality Anatomy Ontology (ZFA). Qualities 🕀 👩 maturity 🖨 👩 morphology come from the Phenotypic Quality 🗄 👩 closure 🗄 👩 deformed 🗈 👩 shape (PATO) ontology. present 🗐 👩 size 🗄 👩 1-D extent 🗄 👩 2-D extent 🗄 👩 3-D extent 😑 👩 decreased size atrophied 🕘 dwarf as normal numbers of present in normal n dystrophic present in fewer present in g umbers in organism parts of type umbers in organism numbers in o 🗄 👩 hypoplastic hypotrophic 🗄 🚯 increased size 🗉 🕤 stubby 🗄 👩 structure 🗉 👩 texture necessity (continuant) 🗄 👩 organismal quality n physical quality Phenoscape Knowledgebase 😴 🕂 Shttp://kb.phenoscape.org/ • Q- Google Welcome to the Phenoscape phenotype database for ostariophysan fishes. Search the Phenoscape Knowledgebase Begin typing to choose a search term from the popup menu The website allows searching by meckel Meckel's cartilage entity term, taxon, and by gene. coronomeckelian Cvprinus meckel synonym ache ahctf1 Amblyceps mangois Ameiurus nebulosus Amphilius atesuensis anterior limb of parapophysis 4 Ariopsis felis Aspredinichthys tibicen Aspredo aspredo atp1a1 Austroglanis barnard autopalatine autopalatine-lateral ethmoid joint Bagre marinus Bagroides melapterus Bagrus bajad Batasio batasio cdh2 cdx4 Cetopsis coecutiens Chaca chaca chd Chrysichthys auratus Chrysichthys longipinnis Chrysichthys nigrodigitatus Clarotes laticeps coronomeckelian dentary tooth Diplomystes chilensis dorsal fin spine 1 epibranchial 1 element epibranchial 2 element

epibranchial 5 cartilage ethmoid cartilage extl3 eva1 fgf8a frontal bone gill raker gpc4 hdac1 Hemibagrus nemurus nyomandibula Ictalurus punctatus infraorbital 1 kita lateral ethmolo Loricaria cataphracta Irrc6

Meckel's cartilage mesethmoid bone metapterygoid mib mycbp2 Mystus nigriceps ndr2 Nematogenys inermis Noturus flavus ntla opercle os suspensorium Pangasianodon hypophthalmus Parakysis verrucosus parapophysis pax2a posterior process of basipterygium pou5f1 premaxilla premaxillary tooth Rhabdalestes septentrionalis Rhamdia laticauda Rhamdia quelen rib Rita rita Scale **Drocess** Schilbe mystus sensory canal shha smo Phenotypes Anatomical Term Zebrafish **Evolutionary** Quality Anatomy Data Data Synonyms: ventral mandibular Meckel's shape 651 taxa <u>1 genes</u> cartilage Values include: shape cartilage Definition: Meckel's cartilage is the bilaterally paired, rod-like, cartilaginous Meckel's structure ventral component of the lower jar, or None <u>35 taxa</u> cartilage Values include: structure ventral mandibular arch. It is typically resorbed in adults. Meckel's count None 1 genes cartilage Values include: absent Meckel's quality <u>2 genes</u> None is a type of: pharyngeal arch cartilage cartilage Values include: deformed, malformed is part of: ventral mandibular arch develops from: retroarticular Meckel's Values include: decreased length, size, <u>3 genes</u> 638 taxa cartilage eased size, decreased size Result pages integrate spatial quality de: angular placement, relational <u>380 taxa</u> 1 genes between genetic and evolutionary data. Acknowledgments. We thank NSF DBI 0641025, NIH HG0002659, and the National Evolutionary Synthesis Center (NESCent), NSF bent #EF-0423641, for funding. We are also inheres_in ndebted to our curators, and the semantic supraorbital bone data integration experts and OBD work in the Berkeley Bioinformatics & Ontologies Project.

Decomposing "post-composition" relations

Rule: $\forall Q, E: inheres_in(Q, E) \implies inheres_in(inheres_in(Q, E), E)$

Rule: \forall Q, E: inheres_in(Q, E) \Rightarrow is_a(inheres_in(Q, E), Q)

Phenotype annotations are typically "post-composed", where an entity and quality are combined into a Compositional Description. For example, an annotation about the quality decreased size (PATO:0000587) of the entity Dorsal Fin (TAO:0001173) may be post-composed into a Compositional Description that looks like PATO:0000587^OBO_REL:inheres_in(TAO:0001173). Instances of is_a and inheres_in relations are extracted from post compositions like this. In the above example, the reasoner extracts

1. PATO:0000587^OBO_REL:inheres_in(TAO:0001173) OBO_REL:inheres_in TAO:0001173, and 2. PATO:0000587^OBO_REL:inheres_in(TAO:0001173) OBO_REL:is_a PATO:000058

Phenoscape-specific rules

This section describes the Phenoscape-specific rules added to the OBD reasoner

PATO Character State relations

The Phenotypes and Traits Ontology (PATO) contains definitions of qualities, many of which are used in phenotype descriptions. These qualities are partitioned into various subsets (or slims) such as attribute slims, absent slims, and value slims. Attribute and value slims are mutually exclusive subsets. Attribute slims include qualities that correspond to Characters of anatomical entities, Color or Shape for example. Value slims include gualities, which correspond to States that a Character may take, for example Red and Blue for the Color character and Curved and Round for the Shape character. These relationships are not explicitly defined in the PATO ontology but can be inferred using the relations shown below

1. PATO:0000587 oboInOwl:inSubset value_slim

2. PATO:0000587 OBO_REL:is_a PATO:0000117

3. PATO:0000117 obolnOwl:inSubset attribute_slim From these definitions, the relationship

1. PATO:0000587 PHENOSCAPE:value for PATO:0000117

can be inferred by the reasoner. Ideally, the inference rule for this can be represented as

Rule: $\forall V, A: in_Subset(V, value_slim) \land is_a(V, A) \land in_subset(A, attribute_slim) \Rightarrow value_for(V, A)$

All ontologies (in OBO format), phenotype assertions (in NeXML files), and the genetic data and phenotypes from ZFIN (tabular) are integrated in a semantic web-like database (OBD) that aims to be a scalable triple-store implementation in SQL.

