PHENOTYPED

Paula Mabbe

June 4, 2018
Second Annual Digital Data in Biodiversity Research Conference: Emerging Innovations for Biodiversity Data

University of California, Berkeley
The Promise

Mabee and her colleagues at NESCenct initiate a bootcamp with the MOD World. They fall for semantic phenotypes but are unaware of the perils ahead.

Missing Link

Moving beyond fundamentals, the scientists begin to use the System's logic on the Phenotyped. Dark secrets about the genetic bases of phenotypes come to light and others have ideas to reverse the logic.

Beyond the Observed

When the scientists attempt to make a synthetic supermatrix of phenotypes, they realize that they can dramatically expand knowledge.

The Internet

Worried about their ability to predict phenotype, they use the internet.
Mabee and her colleagues at NESCent initiate a bootcamp with the MOD World. They fall for semantic phenotypes but are unaware of the perils ahead.
Rich legacy, digital, but not computable
Difficult for machines....

<table>
<thead>
<tr>
<th>OMIM Query</th>
<th># of records</th>
</tr>
</thead>
<tbody>
<tr>
<td>“large bone”</td>
<td>785</td>
</tr>
<tr>
<td>&quot;enlarged bone&quot;</td>
<td>156</td>
</tr>
<tr>
<td>&quot;big bones&quot;</td>
<td>16</td>
</tr>
<tr>
<td>&quot;huge bones&quot;</td>
<td>4</td>
</tr>
<tr>
<td>&quot;massive bones&quot;</td>
<td>28</td>
</tr>
<tr>
<td>&quot;hyperplastic bones&quot;</td>
<td>12</td>
</tr>
<tr>
<td>&quot;hyperplastic bone&quot;</td>
<td>40</td>
</tr>
<tr>
<td>&quot;bone hyperplasia&quot;</td>
<td>134</td>
</tr>
<tr>
<td>&quot;increased bone growth&quot;</td>
<td>612</td>
</tr>
</tbody>
</table>
Discovery impeded

“Find all images of keywords = gill arch or gill arch skeleton”

Results = 0

But there are many images of parts of gill arches

From Morphbank [28 May 2018]
MOD world: using ontologies, model organism databases link phenotypes—genes
Semantics provide meaning

- gill arch skeleton (UB: 0011153)
- replacement bone (UB: 0012075)
- basihyal bone (UB: 0011618)
- basihyal cartilage (UB: 0011615)

relationships:
- part_of: http://purl.obolibrary.org/obo/BFO_0000050
- develops_from: http://purl.obolibrary.org/obo/RO_0002202

Dahdul et al., 2010
Systematic Biology
A (semantic) word is worth a thousand images...

<table>
<thead>
<tr>
<th>Gill arch part</th>
<th>Number images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basihyal bone</td>
<td>147</td>
</tr>
<tr>
<td>Basibranchial</td>
<td>236</td>
</tr>
<tr>
<td>Ceratobranchial</td>
<td>224</td>
</tr>
<tr>
<td>Hypobranchial</td>
<td>110</td>
</tr>
<tr>
<td>Epibranchial</td>
<td>220</td>
</tr>
<tr>
<td>Pharyngobranchial</td>
<td>92</td>
</tr>
<tr>
<td>Copula</td>
<td>70</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>1099</strong></td>
</tr>
</tbody>
</table>

From Morphbank [28 May 2018]
Phenoscape (2007)

**Promise:** Find, compare, integrate computable phenotypes
Perils

Lots of resource & software development: phenotype ontologies, annotation software, Knowledgebase (KB), tools
Phenotyped: 4,260 species
but >1 billion biological specimens in U.S. natural history collections!

Phenoscape Knowledgebase

Anatomy studies
651,660 phenotypes
>5,000 taxa
161 studies
>20,000 character states

Model organism datasets
309,383 phenotypes
15,447 genes
+ 732,658 anatomical expression
34,731 genes
Phenotyped specimen data

- from legacy literature
- from images
- ‘born semantic’ phenotypes
- predicted phenotypes
Moving beyond fundamentals, the scientists begin to use the System’s logic on the Phenotyped. Dark secrets about the genetic bases of phenotypes come to light and others have ideas to reverse the logic.
Semantic phenotypes link species & genes

Model organism

Evolutionary species

Gene A → Phenotype B → Taxon C
Brachyplatystoma capapretum is a variant of Pimelodidae. tfap2a is a sequence-specific DNA binding transcription factor activity that influences some ethmoid cartilage exhibiting some tfap2a*ts213/ts213. This split that inheres in some ethmoid cartilage influences some round that inheres in some ethmoid cartilage.

Gene-Phenotype connection:
- tfap2a
- ethmoid cartilage
- olfactory region
- round
- split
- shape
- chondrocranium cartilage
- has_function
- part_of
- variant_of
- exhibits some
- influences some
- is_a
- is_a
KB is a Hypothesis generator

Interoperability yields 100,000’s of Gene-Phenotype hypotheses

Mabee et al., 2012
E.g., How did the catfish lose its tongue?

Flathead Catfish; Photo by USFWS, used under Creative Commons License
Find the genes with a similar phenotype

Candidate genes

Candidate genes:
- brpf1
- disc1
- disp1
- fac
- foxd3
- hand2
- myst3
- sox9a
- unm_th9
- unm_tn20c
- unm_ty5

Edmunds et al., 2016, Mol. Biol Evol.

Photo: Richard Edmunds

Ictalurus punctatus

Siluriformes taxa:

changes in brpf1?
Phenoblast: find the genes with a similar set of phenotypes to a taxon

Set of phenotypes for a species
Reverse: Find the species with a phenotype similar to a gene…

eda, edar
‘Candidate species’ (195)

- including all eels, some catfish, etc.
- voucher specimens are in natural history collections
Phenotyped natural history specimens automatically linked with genetics

https://celebrating200years.noaa.gov/datasets/fishcollection/
Beyond the Observed 10 min

When the scientists attempt to make a synthetic supermatrix of phenotypes, they realize that they can dramatically expand knowledge.
Genetic supermatrix
Phenotype supermatrix (+/-)

**OntoTrace**: Software tool that returns +/- synthetic matrices for desired taxa and characters
Infer presence, from presence of part

• “Bottom up”

• If part of an entity is present, then the entity is present

• If a species has toe disks, it has digits

digit

manual digit

toe disk

digit is present

manual digit is present

toe disk is present (Heyer, 1988)

toe disks
Infer presence, from quality of entity

- If an entity is asserted to have a quality other than absence, it is present.
- E.g., If radials are jointed, they are present.

Modified from Shubin et al. 2006

Radials present

Radials jointed

Radials are present

Radials are jointed (Zhu & Ahlberg, 1998)
Infer absence, from absence of parent

- “Top down”
- If parent is absent, all children and parts are absent
- If snakes lack limbs, they lack digits
Infer absence, from absence of developmental precursor

- If developmental precursor is absent, entity is absent
- If species lacks limb buds, they lack limbs
Query: “fin, limb, girdle & parts” “Sarcopterygii”

‘Synthetic morphological supermatrix’

Inference provides most of the data for this phenotype

Enabled: Isolation, quantification of data conflicts

774 conflicted cells (0.5%) of 146,451 populated cells
Enabled: Phenotype enrichment

- Heat map:
  - Cell color reflects number of character states:
    - for each anatomical entity (column)
    - for each taxon (row)
  - Dark blue = empty cells

Enabled: Large-scale ancestral state reconstruction

Enhancement to sparse data

Jackson, Balhoff, Hanscom, Franando, & Mabee (2018)
Systematic Biology doi.org/10.1093/sysbio/syx098
Potential: scale up phenotypic data to drive genomic discoveries
Worried humans reached out to the Machine for help in predicting climate change. Cleverly connecting phenotype with collection data, It told the humans which species would be likely to win the competition.
Help me…

- Related taxa with similar phenotypes from dissimilar environments
- Unrelated taxa with similar phenotypes from similar environments
- Understand the direction of phenotypic change in a particular species over time
- Predict the likely changes in phenotype and species’ interactions based on collections data
Reasoning across phenotype & environment

Figure 2 Manual workflow conceptual diagram. This diagram shows the manual workflow to link phenotype and environment data sets using current tools and services.
Phenotypes a portal to understanding the integration between organism and environment
The Phenoscape project is currently funded by NSF ABI Innovation collaborative grants (1661529, 1661356, 1661456, 1661516) and an ABI Development grant (1062542). Phenoscape was previously funded by NSF 1062404 and 0641025, and supported by NESCent, NSF 0905606.

- NSF Phenotype Research Coordination Network (phenotypercn.org), DBI-0956049

- Phenoscape contributors, Advisory Board, Data sources (see: http://phenoscape.org/wiki/Acknowledgments)

- Phenoscape collaborators