iMADS
A sustainable software collaboration for predicting transcription factor binding specificity

Dan Leehr
Duke Center for Genomic and Computational Biology
github.com/Duke-GCB
genome.duke.edu
Why We Develop Scientific Software

Promote dissemination, exploration, and impact of a method or dataset
Unsustainable
Informatics @ GCB

Professional data management, computational reproducibility, software engineering know-how, made accessible to GCB research labs
Figure 7. The iMADS Framework for Integrative Modeling and Analysis of Differential DNA-Binding Specificity between Paralogous TFs

https://doi.org/10.1016/j.cels.2018.02.009
iMADS: Informatics Collaboration

**Researcher**
Science: methods and datasets.

Discovery driven

**Informatics**
Engineering: Practices, tools

Ordered, deliberate
How to build a Minimal Viable Product

"Learn from users using the product as soon as possible"
Milestone

Generate whole-genome predictions and publish as browser tracks
Engineering Practices

"Not everyone needs to know everything"
Tracy Teal on research driven software, 2 days ago

- Automated Tests
  - Always good for software
  - Helps bridge collaboration gap
- Building a Portfolio
  - Libraries for foundation, code re-use
  - Data formats, tools, resources
  - Architecture and Web Frameworks


Audience: Domain scientists with CLI experience
Predicting in Parallel

Assembly

| hg19 | hg38 |

Chromosome

| chr1  | chr2  | ... | chrX  | chrY  |

Model: Protein+Core

| c-Myc CACATG   | c-Myc CACGAG   | ...   | Runx2 GTGGG | Runx2 GTGTT |

2 × 24 × 41 = 1,968
Automating Track Hubs

- Open source pipeline built in Common Workflow Language (CWL)
- Completely automated and reproducible Track Hub
- Publishes genome browser tracks (BigBed) automatically
- Expressive and easy to maintain

https://github.com/Duke-GCB/TrackHubGenerator
Prediction Track Hub

Audience: Domain scientists that use existing resources
Milestone

Online database with scores, visualizations, and on-demand predictions
Searchable Database

• Common use cases: search by genes or genomic ranges

• Visualize high scores (SVG) and download data (CSV)

• Iterative and collaborative

• More engineering, risks of scope

Audience: Domain scientists looking for integrated queries
### Predictions: Scores and Coordinates

<table>
<thead>
<tr>
<th>TF</th>
<th>Prediction</th>
<th>Chrom</th>
<th>Start</th>
<th>Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELK1</td>
<td>0.533</td>
<td>chr1</td>
<td>3020</td>
<td>5020</td>
</tr>
<tr>
<td>ELK1</td>
<td>0.847</td>
<td>chr2</td>
<td>33932</td>
<td>34501</td>
</tr>
<tr>
<td>ELK1</td>
<td>0.122</td>
<td>chr8</td>
<td>53932</td>
<td>63932</td>
</tr>
</tbody>
</table>

### Genes: Names and Coordinates

<table>
<thead>
<tr>
<th>Gene List</th>
<th>Symbol</th>
<th>Chrom</th>
<th>Start</th>
<th>Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>refGene</td>
<td>WASH7P</td>
<td>chr1</td>
<td>5021</td>
<td>5041</td>
</tr>
<tr>
<td>refGene</td>
<td>7SK</td>
<td>chr2</td>
<td>34499</td>
<td>34519</td>
</tr>
<tr>
<td>refGene</td>
<td>8SK</td>
<td>chr8</td>
<td>73932</td>
<td>73952</td>
</tr>
</tbody>
</table>
"Show me the genes where c-Myc binding is highest"
1. **Web** upload sequences

2. **Worker** fetches sequences, predicts, uploads results

3. **Web** view results (BED/CSV/SVG)
On-Demand Predictions

Audience: Domain scientists that want to use our models
Goal

Sustainable, maintainable development and reproducible deployment
Open Source, Portable, Reproducible, Maintainable

- Prediction Engines & Model Training
- Annotation Track Synthesis Pipelines
- Web Application and Database
- Container descriptions and Orchestration

Audience: Ourselves, collaborators, partner IT
Wrap-Up

• Building sustainable research software and requires institutional support, collaboration, and targeted skills

• Constantly target specific audiences with useful deliverable milestones to guide for success

• Consider reproducibility in scientific methods, development and deployment

• Building relationships
Acknowledgements

- Gordan Lab - gordanlab.cs.duke.edu
- Raluca Gordan, Ning Shen, Josh Schipper
- Duke GCB Informatics - genome.duke.edu
- John Bradley, Hilmar Lapp

Published in Cell Systems, Volume 6, Issue 4, 470-483.e8, 25 April 2018.
https://doi.org/10.1016/j.cels.2018.02.009
https://doi.org/10.5281/zenodo.1205525
https://imads.genome.duke.edu
Thank You

github.com/Duke-GCB/iMADS
imads.genome.duke.edu
dan.leebr@duke.edu
@Leehro