Open Source Software Packages: Using and Making your contributions

June 5, 2017

Bioinformatics MS Interest Group
Your hosts

Meena Choi
Post doc.
Northeastern University

Statistical methods for quantitative proteomics

Samuel Payne
Scientist
Pacific Northwest National Lab

Integrative Omics
Outline

• General Intro – Meena Choi
• mzRefinery/proteowizard - Sam Payne
• openMS - Oliver Kohlbacher
• Skyline - Brendan MacLean
• General discussion on open source
  • Ask questions for the General Discussion http://bit.ly/2qNZVBU
  • Shout-out for Open Source tool http://bit.ly/2qVHV07
Oliver Kohlbacher

- The chair of Applied Bioinformatics at University of Tübingen & fellow at the Max Plank Institute
- OpenMS (openms.de)
Brendan MacLean

• Principal developer for Skyline (skyline.ms)
• University of Washington
Ask questions or comments:
http://bit.ly/2qNZVBU

• Why have open source?
• What are the advantages and disadvantages between open source and private closed-source software?
• How should a developer consider the question of making a project open source or not?
• What is appropriate level of guide/documentation to help new developers?
• How to incentivize people to contribute to open source software?
• Provide the framework for developing package: basic structure, requirements...
  • Requirements:
    1. pass check or BiocCheck on all supported platforms (their own checking system)
    2. Documents
      • DESCRIPTION, NAMESPACE, vignette, help file, NEWS
    3. Review process (2-5 weeks)
      • submit a GitHub repository
      • a reviewer will be assigned and a detailed package review is returned.
      • the process is repeated until the package is accepted to Bioconductor.
• Maintaining the packages across release cycles (twice a year) + deprecate packages
• Import or depend on other packages in Bioconductor or CRAN
R package as software

- Easy to make open source software for new method development.
- Reproducible: R script, R markdown
- Can improve GUI with Shiny
- Not easy to work with other language
Skyline

Targeted Mass Spec Environment

Reflections on open source projects

Brendan MacLean
MacCoss Lab
University of Washington
Personal Open Source History

- Microsoft 1991 – Microsoft Foundation Classes
- BEA Systems 2001 – Apache XML Beans
  - Last release 2012 – Retired 2014
- LabKey Server 2003 (originally CPAS)
  - X! Tandem contributions – pluggable scoring & k-score
  - TPP X! Tandem pipeline
- Skyline as a ProteoWizard subproject
  - Drove vendor acceptance of open source licensing
- Panorama as a module in LabKey Server
Skyline Project Overview

- 9th year in development
- 6 developers
- over 500,000 lines of code – open source
- 60,000 lines of code for testing
- over 8000 registered users
- over 7500 instances started each week
- funded by NIH and 6 instrument vendors

Overstated?
Chromatography-based Quantification

- SRM – Selected ion chromatograms
- PRM – Extracted ion chromatograms
- DIA – Extracted ion chromatograms
- DDA – Extracted ion chromatograms from MS1-only

<table>
<thead>
<tr>
<th>Acquisition</th>
<th>Targeted</th>
<th>Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>More Selective</td>
<td>PRM</td>
<td>DIA</td>
</tr>
<tr>
<td>Less Selective</td>
<td>SRM</td>
<td>DDA</td>
</tr>
</tbody>
</table>
Aggregating and Publishing

- Publish fully annotated Skyline documents
- Build chromatogram libraries
- Aggregate lab QC data

- Free hosted version ([http://panoramaweb.org](http://panoramaweb.org))
  - >220 separate projects so far (CPTAC, LINCS & ABRF sPRG)
  - >2500 data sets uploaded (+7000 QC documents)
  - User controlled security

- Locally installable server application
  - Roche, Genentech, [unnamed], Merck, CPTAC, Amgen

- Free and open source (Apache 2.0)

Sharma, et. al, *J. Proteome Res. 2014*
Skyline/Panorama Workflow

1. **Refine**
2. **Acquire Data**
3. **Build Method**
4. **Evaluate Results**

- **Download** Skyline documents and chromatogram libraries
- **Publish to Panorama**
- **Panorama**
  - Store in a relational database
  - Explore results in a web-browser
  - Search
  - Analyze
  - Export

- **Skyline**

- **Person**
On Open Source Licensing

- Public Domain
- Apache 2.0
  - Explicit rights to patents
- Berkeley Software Distribution (BSD)
  - Use as you please
- Artistic, Mozilla …
  - No branching allowed, adaptations must be public
- LGPL
  - Backwards architecting of using software
- GPL
  - “Viral” – users must open their own source
More Permissive Has Benefits

- May inspire broader adoption
  - Adoption is critical to all software
- May inspire trust from funders
  - Public grants
  - For profit companies
    - Instrument vendors
    - Pharma companies
- May inspire more outside contribution
Why people worry about going open

- **Loss of “control”**
  - Others will jump in and push the code places I don’t want

- **Loss of revenue opportunity**
  - Having a free offering limits revenue potential

- **Loss of advantage**
  - Others can see and steal my best ideas

- **Exposure**
  - My code is not ready to share with others
The biggest open source fallacy

- **Overestimating interest in your project**
  - If I open source the project, it will go faster or last longer…
  - More likely benefits:
    - Extra scrutiny
    - Occasionally, inspired contributions
    - Broader adoption and interest

- If I open source, others will read the code and find it lacking…
- If it is important enough to read, they may offer improvements.

- Let’s open source that project that grad student left behind
What motivates involvement?
Skyline Team

- Nick Shulman
- Don Marsh
- Brian Pratt
- Max Horowitz-Gelb
- Vagisha Sharma
- Nat Brace
- Kaipo Tamura
- Yuval Boss
OpenMS – Fundamentals and Getting Involved

Oliver Kohlbacher
University of Tübingen and
MPI for Developmental Biology
KohlbacherLab.org | @okohlbacher
OpenMS

- **OpenMS** – an open-source C++ framework for computational mass spectrometry
- Jointly developed at ETH Zürich, FU Berlin, University of Tübingen
- **Open source**: BSD 3-clause license
- **Portable**: available on Windows, OSX, Linux
- **Vendor-independent**: supports all standard formats and vendor-formats through proteowizard

**TOPP – The OpenMS Proteomics Pipeline**
- Building blocks: One application for each analysis step
- All applications share **identical user interfaces**
- Uses PSI **standard formats** and integrates seamlessly with other applications supporting these formats

**Tools** can be integrated in various **workflow systems**
- TOPPAS – TOPP Pipeline Assistant
- Galaxy
- WS-PGRADE/gUSE
- Proteome Discoverer/Compound Discoverer
- KNIME

Kohlbacher et al., Bioinformatics (2007), 23:e191
OpenMS 2.x - Features

- **Currently 185 distinct tools**
- **Utilities** – extract information from files, file conversion, visualization
- **PTX identification** – interface to DB search engines, de novo search, RNA-protein XL MS, protein inference, RT prediction, proteotypicity prediction
- **PTX quantification** – label-free, TMT, iTRAQ, SILAC, MRM, OpenSWATH (DIA), ProteinSIP (metaproteomics), RT alignment
- **MTX quantification** – nontargeted metabolomics, MRM
- **MTX identification** – accurate mass DB search, spectral matching, composition
- **Miscellaneous** – MRM scheduling, LC-MS simulator, ...

Röst et al., Nat. Methods, 2016, 13:741
Tool Documentation

- Documentation for each tool is available as part of the OpenMS documentation (www.OpenMS.org)

**FeatureFinder**

The feature detection application for quantitation.

![Diagram of FeatureFinder and related tools]

This module identifies "features" in a LC/MS map. By feature, we understand a peptide in a MS sample that reveals a characteristic isotope distribution. The algorithm computes positions in rt and m/z dimension and a charge estimate of each peptide.

The algorithm identifies pronounced regions of the data around so-called seeds. In the next step, we iteratively fit a model of the isotope profile and the retention time to these data points. Data points with a low probability under this model are removed from the feature region. The intensity of the feature is then given by the sum of the data points included in its regions.

How to find suitable parameters and details of the different algorithms implemented are described in the TOPP tutorial.

**Note:**

- that the wavelet transform is very slow on high-resolution spectra (i.e. FT, Orbitrap). We recommend to use a noise or intensity filter to remove spurious points first and to speed-up the feature detection process.

Specialized tools are available for some experimental techniques: SILACAnalyzer, ITRAQAnalyzer.

**The command line parameters of this tool are:**

```
FeatureFinder -- Detects two-dimensional features in LC-MS data.
Version: 1.7.0 Sep 3 2010, 15:13:04, Revision: 7349

Usage:
FeatureFinder <options>
```
Tool Implementation

- Very easy to implement thanks to the OpenMS framework
- Usually short (200 lines of code on average, mostly concerned with parameter handling)
- Use of the OpenMS core library

**IDMapper.C:**

```cpp
[...]
vector<ProteinIdentification> protein_ids;
vector<PeptideIdentification> peptide_ids;
String document_id;
IdXMLFile().load(getStringOption_  
    ("id"), protein_ids,peptide_ids, document_id);
IDMapper mapper;
[...]
ConsensusXMLFile file;
ConsensusMap map;
file.load(in, map);
mapper.annotate(map, peptide_ids, protein_ids, false);
file.store(out, map);
```
Workflows as an Abstraction
Workflow Repositories

- OpenMS website contains a workflow repository with selected example workflows ([www.OpenMS.org](http://www.OpenMS.org))
- General-purpose workflow repository: [www.myexperiment.org](http://www.myexperiment.org)
  - Collects workflows from arbitrary workflow engines
  - Numerous applications, can be used to document data analysis
Open (Data | Source | Science)

Röst et al., Nat. Methods, 2016, 13:741
Who uses OpenMS? And what parts of it?
Getting Involved

• **Contributing algorithms**
  • Got interesting algorithms you developed?
  • Anything missing in the library/tools?

• **Contributing interfaces**
  • Got a tool we have no other solution for?
  • Let’s discuss interfaces and levels of integration

• **Contributing scripts**
  • Solved an interesting problem once?
  • Need help hacking a quick prototype for something?

• **Contributing workflows**
  • Put something together that solves a particular problem?

Get in touch with us – ideally via the mailing list or the talk to me during the conference!
Materials

• **OpenMS Website:** [http://www.OpenMS.org](http://www.OpenMS.org)
  - Documentation
  - Tutorials
  - Online lecture ‘Computational Proteomics and Metabolomics’ (Kohlbacher, Reinert, Nahnsen) [http://bit.ly/2d2kBSq](http://bit.ly/2d2kBSq)

• Downloads
  - Binaries
  - Source code
  - Plugins for Proteome Discoverer

• Access to mailing lists – this is where you can get help 24/7