Applying Reproducible Research in Scientific Discovery

Robert Gentleman

Hutchinson Center

rgentlem@fhcrc.org

BioSilico

Oct 2005

FRED HUTCHINSON CANCER RESEARCH CENTER

Outline

- What is reproducible research
- How do we implement it
- Publishing
- Problems
- Future

Reproducibility

- means different things in different disciplines
  - Physics tends to require a much higher level of reproducibility than biology (lower between experiment variance)
  - computation should be essentially reproducible, but we seem to be holding it up to a lesser standard
- we are not talking about data collection, but rather about data analysis

Scientific Publication: Author

- an author selects a set of data
- they transform that data, to produce pictures, tables and statistics
- from these they draw conclusions

Scientific Publication: Reader

- get a static document
- you can read it, try to understand what the author did
- electronic publishing has made it easier to
  - get the document
  - carry it around
  - search it
- but not to understand or comprehend it

Dissection

- Data collection
  - **not** reproducible or **limited** reproducibility
  - **not** part of our discussion
  - but, in all cases, the author must have selected some specific version to analyze
  - our interest begins there
  - from that point on, things should be reproducible

Dissection

- Data analysis: creation of figures, statistics, tables
  - should be reproducible
  - the data have been fixed
  - well defined and understood statistical methods have been applied
  - many readers should be able to understand the complete details

Dissection

- Conclusions
  - depend on the data and the analysis
  - given them, we can agree or disagree with the author
  - if we don't understand the analysis how do we agree with the conclusions?
  - download their data - reanalyze…*frustration*…unable to reproduce…**my fault**?…**their fault**?

How to do better

- these documents are integrated collections of text code and data, we call them *compendiums*
- we need a better set of authoring tools
- they should allow us to write structured documents, with alternating code and text segments
- there should be a place for data
- a transformer takes the raw document and transforms it
  - text chunks are formatted
  - code chunks are evaluated and replaced by their output
  - document is created (pdf, html, etc)

Early Work

- Claerbout's lab at Stanford
  - use of Makefiles
- Buckheit and Donoho (1995)
  - plots should be reproducible
- Don Knuth
  - Literate Programming
- Duncan Temple Lang
  - extensible dynamic docs
- Tony Rossini
  - Literate Data Analysis
- Fritz Leisch
  - Sweave
- Vincent Carey
  - Literate programming

Sweave

```
\section{Data}

We see an interesting pattern in
    Figure~\ref{F1}
<<F1, fig=TRUE>>=
plot(data.x,data.y)
@
And so we like it.
```

- on the left we see a section of an Sweave document
- first, standard LaTeX and then a small code chunk that is R code
- after weaving the code chunk will be replaced by the code to include the plot (which is in eps or pdf)

Publishing

- we mean merely *making available*
- such documents do not need to be published in the sense that others can take them and *read* them
- could be kept internal to a company, or research group
- our major use of them is to train new post-docs, researchers, etc
- from a compendium a new person can take over a project and quickly be productive
- next they learn to produce their own

Problems

- reliance on external libraries/software
- reliance on operating system
- use of large external data sources
- security
  - can we protect private data but still allow some modeling
  - encryption systems

Immediate Future

- an XML implementation
- XSLT for the transformer
- code chunks in more languages
  - Perl
  - Python
  - SAS
  - Excel

More distant future

- Authors need: .
  - better authoring tools
  - documents become trees, where some branches are explored by those who want to know the details, others will skip over these
  - dissemination
- Readers need:
  - ways to get and read compendiums
  - tools to interact; sliders, menu selections
  - navigational tools (compendiums could have a tree like structure)

Some Examples

- you can look at some R-centric examples at
- www.bioconductor.org
- Follow the links for Publications

Thanks to

- Vince Carey, Channing Lab
- Duncan Temple Lang, UC Davis