pIR: An *R* package for isoelectric point prediction based on amino acid sequences

Yasset Perez-Riverol

June 17, 2015

Abstract

Accurate estimation of the isoelectric point value (pI) based on the amino acid sequence becomes critical to perform proteomics experiments. Also, it is one of the most useful electrostatic properties to study peptides and proteins. Different methods has been proposed to compute the theoretical isoelectric point of peptides and proteins using several pK sets [1, 2, 3]. This vignette provides a brief overview of the available interface and functionality as well as a short use case.

Keywords: proteomics, peptides, proteins, electrophoresis, mass spectrometry, isoelectric point, tutorial.

Contents

1 Authoring Sweave / LaTeX package vignettes

To use with Sweave, add the following to your package 'DESCRIPTION' file:

```
Suggests: BiocStyle
```

and add this code chunk to the preamble (between the \documentclass{article} and \begin{document} latex commands) of your .Rnw file:

```
<<style-Sweave, eval=TRUE, echo=FALSE, results=tex>>=
BiocStyle::latex()
@
```

To use with *knitr*, add the following to the 'DESCRIPTION' file:

```
VignetteBuilder: knitr
Suggests: BiocStyle, knitr
```

this to the top of the `.Rnw` file:

```
%\VignetteEngine{knitr::knitr}
```

and this to the preamble:

```
<<style-knitr, eval=TRUE, echo=FALSE, results="asis">>=
BiocStyle::latex()
@
```

See `?latex` for additional options. *BiocStyle* automatically attaches the following LaTeX packages: `color`, `enumitem`, `fancyhdr`, `geometry`, `hyperref`, `parskip`, `sectsty`.

Provided the package has been installed, a convenient way to view the vignette as it is being written is with the command

```
R CMD Sweave --pdf vignette.Rnw
```

A short-cut useful for checking the LaTeX portion of vignettes is to toggle evaluation of code chunks to `FALSE`

```
SWEAVE_OPTIONS="eval=FALSE" R CMD Sweave --pdf vignette.Rnw
```

When using *knitr*, the command to process the vignette is

```
R CMD Sweave --engine=knitr::knitr --pdf vignette.Rnw
```

By default, *knitr* automatically caches results of vignette chunks, greatly accelerating the turnaround time required for edits. Both the default and *knitr* incantations create PDF files using *texi2dvi –pdf*; many versions of this software incorrectly display non-breaking spaces as a tilde, ~. This can be remedied (as on the *Bioconductor* build system) with a final run of

```
R CMD texi2dvi --pdf vignette.tex
R CMD pdflatex vignette.tex
```

2 Style macros

BiocStyle introduces the following additional markup styling commands useful in typical *Bioconductor* vignettes.

Software:

- `\R{}` and `\Bioconductor{}` to reference *R* software and the *Bioconductor* project.
- `\software{GATK}` to reference third-party software, e.g., *GATK*.

Packages:

- `\Biocpkg{IRanges}` for *Bioconductor* software packages, including a link to the release version landing page, *IRanges*.
- `\Biocannopkg{org.Hs.eg.db}` for *Bioconductor* annotation packages, including a link to the release version landing page, *org.Hs.eg.db*.
- `\Biocexptpkg{parathyroidSE}` for *Bioconductor* experiment data packages, including a link to the release version landing page, *parathyroidSE*.
- `\CRANpkg{data.table}` for *R* packages available on CRAN, including a link to the FHCRC CRAN mirror landing page, *data.table*.
- `\Githubpkg{rstudio/rmarkdown}` for *R* packages available on GitHub, including a link to the package repository, rstudio/rmarkdown.
- `\Rpackage{MyPkg}` for *R* packages that are *not* available on *Bioconductor* or CRAN, *MyPkg*.

Code:

- `\Rfunction{findOverlaps}` for functions `findOverlaps`.
- `\Robject{olaps}` for variables `olaps`.
- `\Rclass{GRanges}` when referring to a formal class *GRanges*.
- `\Rcode{log(x)}` for *R* code, `log(x)`.

Communication:

- `\bioccomment{additional information for the user}` communicates *comment: additional information for the user*.
- `\warning{common pitfalls}` signals *warning: common pitfalls*.
- `\fixme{incomplete functionality}` provides an indication of *fixme: incomplete functionality*.

General:

- `\email{user@domain.com}` to provide a linked email address, `user@domain.com`.
- `\file{script.R}` for file names and file paths 'script.R'.

3 Title, running headers, and table of contents

Create a title and running headers by defining the `\bioctitle` and `\author` commands in the preamble

```
\bioctitle[Short title for headers]{Full title for title page}
%% also: \bioctitle{Title used for both header and title page}
%% or... \title{Title used for both header and title page}
\author{Iman Author\footnote{iman@author.org}}
```

Use `\maketitle` at the start of the document to create the title in the document.

Use `\tableofcontents` for a hyperlinked table of contents, `\section`, `\subsection`, `\subsubsection` for structuring your vignette.

Formatting of subsections and subsubsections are as follows.

3.1 This is a subsection

3.1.1 This is a subsubsection

4 Figures

Besides the usual LaTeX capabilities (`figure` environment and `\includegraphics` command), 'Bioconductor.sty' defines a macro `\incfig[placement]{filename}{width}{shorttitle}{extendedcaption}`, which expects four arguments:

filename The name of the figure file, also used as the label by which the float can be referred to by `\ref{}`. Some *Sweave* and *knitr* options place figures in a subdirectory; unless `short.fignames=TRUE` is set the full file name, including the subdirectory and any prefixes, should be provided. By default, these are '<sweavename>-' for *Sweave* and 'figure/' for *knitr*. Please note the different naming scheme used by *knitr*: figure files are named '<chunkname>-i' where *i* is the number of the plot generated in the chunk.

width Figure width.

shorttitle A short description, used in the list of figures and printed in bold as the first part of the caption.

extendedcaption Continuation of the figure caption.

The optional **placement** specifier controls where the figure is placed on page and takes the usual values allowed by LaTeX floats, i.e., a list containing `t`, `b`, `p`, or `h`, where letters enumerate permitted placements. If no placement specifier is given, the default `tbp` is assumed.

Figure 1: **A curve.** The code that creates this figure is shown in the code chunk.

For incfig with Sweave, use

```
<<figureexample, fig=TRUE, include=FALSE, width=4.2, height=4.6>>=
v = seq(0, 60i, length=1000)
plot(abs(v)*exp(v), type="l", col="Royalblue")
@
\incfig{LatexStyle-figureexample}{0.25\textwidth}{A curve.}
  {The code that creates this figure is shown in the code chunk.}
as shown in Figure˜\ref{LatexStyle-figureexample}.
```

This results in

```
> v = seq(0, 60i, length=1000)
> plot(abs(v)*exp(v), type="l", col="Royalblue")
```

as shown in Figure 1. When the option `short.fignames` is set to `TRUE`, figure names used by `\incfig` and `\ref` do not contain any prefix and are identical to the corresponding code chunk labels (plus figure number in case of *knitr*). For example, in Sweave the respective code for the above example would be `\incfig{figureexample}{...}{...}{...}` and `\ref{figureexample}`, while in *knitr* these are expected to be `\incfig{figureexample-1}{...}{...}{...}` and `\ref{figureexample-1}`.

For `\incfig` with *knitr*, use the option `fig.show='hide'` rather than `include=FALSE`. The *knitr*-equivalent code for Figure 1 is:

```
<<figureexample, fig.show='hide', fig.width=4.2, fig.height=4.6>>=
v = seq(0, 60i, length=1000)
plot(abs(v)*exp(v), type="l", col="Royalblue")
@
```

Note the difference in option names setting the figure width and height compared to *Sweave*. Unless `short.fignames=TRUE` is set, use the default 'figure/' prefix when inserting and referring to figures, e.g.:

```
\incfig{figure/figureexample-1}{0.25\textwidth}{A curve.}
  {The code that creates this figure is shown in the code chunk.}
```

A custom prefix for figure file names can be passed to `latex` using the `fig.path` option. When `short.fignames=TRUE`, figures can be referred to directly by code chunk labels, as described earlier in this section.

5 Session info

Here is the output of `sessionInfo` on the system on which this document was compiled:

```
> toLatex(sessionInfo())
```

- R version 3.1.3 (2015-03-09), `x86_64-apple-darwin10.8.0`
- Locale: `en_GB.UTF-8/en_GB.UTF-8/en_GB.UTF-8/C/en_GB.UTF-8/en_GB.UTF-8`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Loaded via a namespace (and not attached): BiocStyle 1.2.0, tools 3.1.3

References

[1] Yasset Perez-Riverol, Enrique Audain, Aleli Millan, Yassel Ramos, Aniel Sanchez, Juan Antonio Vizcaíno, Rui Wang, Markus Müller, Yoan J Machado, Lazaro H Betancourt, et al. Isoelectric point optimization using peptide descriptors and support vector machines. *Journal of proteomics*, 75(7):2269–2274, 2012.

[2] Benjamin J Cargile, Joel R Sevinsky, Amal S Essader, Jerry P Eu, and James L Stephenson. Calculation of the isoelectric point of tryptic peptides in the ph 3.5–4.5 range based on adjacent amino acid effects. *Electrophoresis*, 29(13):2768–2778, 2008.

[3] Bengt Bjellqvist, Graham J Hughes, Christian Pasquali, Nicole Paquet, Florence Ravier, Jean-Charles Sanchez, Séverine Frutiger, and Denis Hochstrasser. The focusing positions of polypeptides in immobilized ph gradients can be predicted from their amino acid sequences. *Electrophoresis*, 14(1):1023–1031, 1993.