The PracTEX Journal, 2007, No. 4
Article revision 2007/12/11

LaTeX Tools for Life Scientists (BioTEXniques?)

Senthil Kumar M

Email murugapiran.senthil@gmail.com
Address Systems Biology Lab, Ajou University, Suwon, Republic of Korea

Abstract LaTeX has been a long time favorite of mathematicians and physicists alike. However now, many packages are available, that have tremendously extended the capabilities of LaTeX beyond routine typesetting and provide biologists new avenues to not only typeset documents, but also help in the visualization of membrane proteins and in the analysis of DNA or amino acid sequences by multiple sequence alignment. I will discuss with examples some of the LaTeX packages and tools that are presently available for the biologists. Scientific journals (for biological research) now accept TeX/LaTeX formatted manuscripts, although they are still a rarity. This article will provide the references of those sources that might be helpful to prospective authors from life sciences that want to submit manuscripts in TeX/LaTeX format. This article is written in the perspective of a biologist who might be interested in creating better documents using LaTeX & friends.

"Uncle Cosmo ... why do they call this a word processor?"
"It's simple, Skyler ... you've seen what food processors do to food, right?"
– Jeff MacNelly, "Shoe"

The advantages of LaTeX over WYSIWYG applications are well known[1]. It has not only been the traditional application of choice to typeset mathematical formulae, but also had been employed to typeset music scores, games like crossword, chess and bridge. In this article, I will explain some of the tools that can be effectively used by life scientists for preparing documents in LaTeX, briefly explain about two LaTeX packages related to biology namely, TeXshade and TeXtopo, developed by Eric Beitz, discuss about XFig and Inkscape, two well known drawing programs and also give some examples illustrating the use of XyMTeX, a package for drawing chemical structures. I will end my article with some references to online sources that might be helpful to life science researchers in search of a style file or a bibliography file suitable for a particular journal and other LaTeX sources.

Regular readers of this journal would have read an earlier article by Peter Flom[1]. It is one of the good places to start for academicians with little or no LaTeX experience, since it provides a good introduction and shatters several myths about LaTeX. The present article is written to help mainly biochemists and molecular biologists. A general background on using TeX or LaTeX would be useful, but not essential to try the things described in this article.

1. For example: http://nitens.org/taraborelli/latex and links therein.

1 Importance of multiple sequence alignment

Proteins are polymers of amino acids and they form the building blocks of life. Enzymes that catalyze the chemical reactions in our cells, hemoglobin that carriers oxygen to our cells through bloodstream, actin that provides stability to cells are some examples of proteins. Proteins are encoded by specific genes. Information in DNA (in the form of 4 letters: A or T or G or C) are translated into amino acids (20 different letters) during protein synthesis. Therefore, analyzing DNA or protein sequences is pivotal to determine the nature of how genes and thus proteins have changed during evolution. Aligning multiple sequences of either DNA or protein sequences from many different organisms is called as multiple sequence alignment. It provides an overall view of comparative changes in the sequences under study, that may have occured due to mutations.

1.1 Aligning sequences using TEXshade

Availability: CTAN

Author: Eric Beitz

Current Version : 1.17

TEXshade, a LATEX package provides an ideal solution of displaying the key changes in DNA (4 letters: A, T, G and C) or protein (20 different amino acids, indicated by unique single letters) sequences with great control. Other programs do exist that provide a way to display sequence similarities in multiple sequence alignment `prettyplot`[2], included in EMBOSS[3], a open source software for molecular biology), is one good example. However, only TEXshade provides:

- LATEX quality output.
- Flexibility: The alignments can be typeset as per the needs for example, while writing a paper using LATEX. TEXshade provides different modes of shading, therefore, it takes less time to modify an alignment.

DNA or Protein sequences are aligned *a priori* using alignment programs like `clustalW` to generate an output file (usually a plain text ASCII file) containing the alignment of sequences from different species. TEXshade can then be used to highlight the level of similarity/identity/dissimilarity among the sequences.

TEXshade provides four predefined shading modes: identity, similarity, diversity and functionality. They come useful depending on what one requires. For example, if one would like to determine the sequence similarity (as shown in Figure 1), `similar` option highlights the residues that are similar in all the species. If one is interested in showing

2. http://bioweb.pasteur.fr/docs/EMBOSS/prettyplot.html
3. http://emboss.sourceforge.net/

the identity (*i.e.* residues that occur above a given threshold level). It is possible to show only those residues that differ among a set of sequences using diversity mode.

Example 1: A simple example using TEXshade. The output is shown in Figure 1.

```
\usepackage{texshade}
\begin{texshade}{protein-texshade.aln}
%Acceptable file formats are: ALN, MSF & FASTA
\residuesperline*{50}
\shadingmode[allmatchspecial]{similar}
\setends{1}{5..300}
\hideconsensus
\feature{top}{1}{25..25}{fill:$\downarrow$}
{First line of each block shows the human IF2 protein}
\feature{bottom}{1}{75..76}{brace}
{Residues that are absent in the human protein}
\end{texshade}
```

Under functionality, six different options are further available that truely shows the strong capabilities of TEXshade. So one can choose to highlight amino acid residues based on: charge, hydropathy, structure, chemical nature, standard area (surface area of amino acid sidechain), accessible area (by solvent molecules).

Apart from all these features, TEXshade can be further configured to read standard protein secondary structure files from the following format: DSSP, STRIDE, PHD or HMMTOP. These structural information provided from these files can then be used along with the sequence alignment to provide much more information. For an exhaustive list of options and examples please refer TEXshade documentation provided with the package[4] and the original research paper[2]. TEXshade can also be used through a web interface. Biology Workbench, Ver 3.2, available at the San Diego Supercomputer Center website, is a collection of bioinformatics tools and anyone can register and are allowed to use these online tools. TEXshade output options are clearly indicated by simple radio buttons or pull down menus (Figure 2).

2 TEXtopo and transmembrane proteins

Availability: CTAN

Author: Eric Beitz

Current Version : 1.4

4. Available here:
http://www.ctan.org/tex-archive/macros/latex/contrib/texshade/texshade.pdf

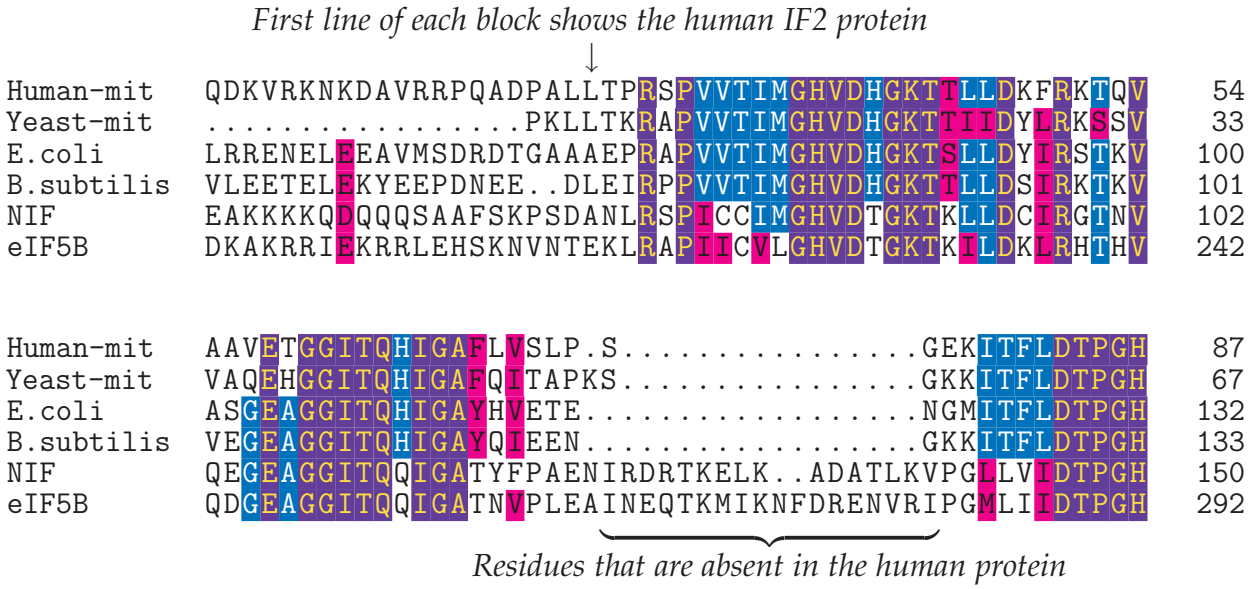


Figure 1: Protein sequence alignment using TEXshade. Alignments can be annotated to describe the features of a particular stretch of residues.

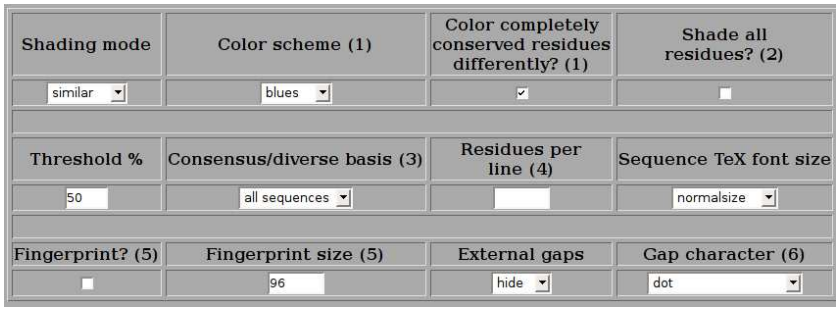


Figure 2: Biology Workbench (Ver 3.2) at the San Diego Supercomputer Center's site provides TEXshade through a simple, easy to use web interface.

Transmembrane proteins are an important class of proteins that play significant role in signaling, ATP production etc. Several such proteins are known and there are comprehensive databases that cater to biologists studying them[5].

2.1 A quick example of using TEXtopo

As an example, I will illustrate the use of TEXtopo using a `Swissprot` file. A sample file is shown here:

A SwissProt file.

```
ID   C56D1_HUMAN             Reviewed;         229 AA.
...
PE   2: Evidence at transcript level;
KW   Transmembrane; Transport.
FT                                /FTId=PRO_0000151034.
FT   TOPO_DOM      1     24       Cytoplasmic (Potential).
FT   TRANSMEM     25     45       Potential.
FT   TRANSMEM    129    149       Potential.
FT   TRANSMEM    170    190       Potential.
FT   TRANSMEM    194    214       Potential.
FT   TOPO_DOM    215    229       Cytoplasmic (Potential).
FT   DOMAIN       22    224       Cytochrome b561.
SQ   SEQUENCE   229 AA;  25424 MW;  43978DAF7D8EC218 CRC64;
     MQPLEVGLVP APAGEPRLTR WLRRGSGILA HLVALGFTIF LTALSRPGTS LFSWHPVFMA
     LAFCLCMAEA ILLFSPEHSL FFFCSRKARI RLHWAGQTLA ILCAALGLGF IISSRTRSEL
     PHLVSWHSWV GALTLLATAV QALCGLCLLC PRAARVSRVA RLKLYHLTCG LVVYLMATVT
     VLLGMYSVWF QAQIKGAAWY LCLALPVYPA LVIMHQISRS YLPRKKMEM
//
```

TEXtopo takes the features from this file (indicated by FT at the start of each line in the `SwissProt` file) and the sequence is depicted across a membrane with the number of transmembrane domains that the protein has. In this case, the number of transmembranes are Six. An example is given below.

Example 2: A simple example using TEXtopo. The output is shown in Figure 3.

```
\usepackage{textopo}
\begin{textopo}
\getsequence[make new]{SwissProt}{Q8N8Q1.SP}
\end{textopo}
```

Apart from Swissprot files, in TEXtopo (just like TEXshade), other file formats like PHD, HMMTOP can be used. Alignment files can also be used or even a raw sequence

5. For example see: http://www.expasy.org

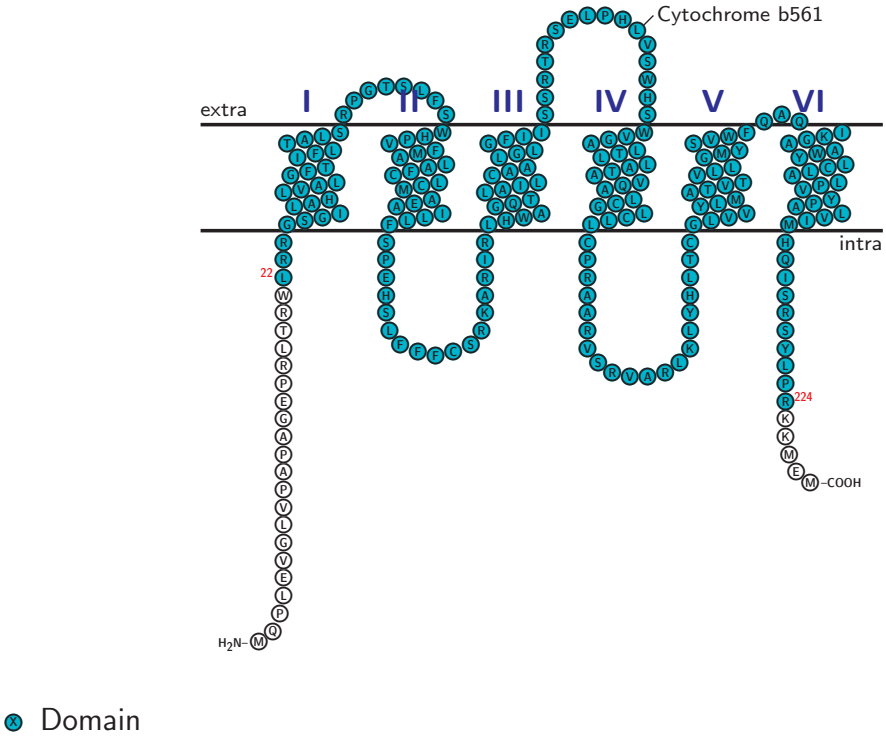


Figure 3: Human Cytochrome b561 domain-containing protein 1 rendered with TEXtopo. Go on, zoom in and see it better!

can be entered using `\sequence` command. More examples are available from the package documentation[6] or in the original paper[3].

3 Preparation of publication quality figures

Traditionally, figures for papers in biological research have been prepared using presentation software[7]. Image files from Southern or Northern or Western blots, SDS-PAGE gels, Footprinting experiments would be imported in presentation software, then further text and other information regarding the experiments are added and the file is then saved as TIFF or JPEG image formats. These image files are then cropped using a image editing software to generate correctly cropped image files for publication.

6. Available here:
http://tug.ctan.org/cgi-bin/getFile.py?fn=/macros/latex/contrib/textopo/textopo.pdf
7. By presentation software, I mean Openoffice Impress/Draw or the more popular and commonly used Microsoft Powerpoint™.

3.1 XFig & Inkscape

Preparing figures from molecular biology experiments as we have have seen is much different from mathematics or physics where generally not much post processing is involved once a graph is ready. Basically what we generally do to make a figure, is:

1. To import images of gels, blots from properietary format to a general format like `TIFF` or `eps`.
2. Add text labels and other experimental details.
3. Export the finished product as a commonly used image format file.
4. Include this in your document.

XFig and Inkscape could be employed for doing these with high efficiency and it produces figures that are neater. Figures that take hours in presentation software could be easily prepared using XFig.

I have included one example that takes advantage of scalable vector graphics (Figure 3.1) which shows a map (drawn to scale) of different plasmid constructs. The documentation for XFig is widely available and the software itself is either installed default in many GNU/Linux distributions or can be downloaded from the web.[8] Other software available for scalable vector graphics are: Inkscape, Mayura (Windows) etc. So, XFig or Inkscape along with LaTeX is a powerful tool that is available to authors not only from mathematical background, but also from life sciences.

Some people might be baffled by the lack of an easy to use interface provided by XFig. For those, I heartily recommend Inkscape a versatile Scalable Vector Graphics creator and editor with a "modern" user interface. Creating figures using Inkscape is very intuitive (as it is in XFig, but with Inkscape it is more fun). Figure 5 shows such an example of a simple figure.

4 Creating chemical formulae with XyMTeX

XyMTeX[9] is a macro package for TeX, written by Fujita Shinsaku for rendering high quality chemical structures. A simple example to draw a phenol molecule (as shown in Table 1) is given:

```
\usepackage{xymtex}
\begin{document}
.
.
\bzdrh{4==OH}
\end{document}
```

8. Please refer: http://xfig.org
9. http://homepage3.nifty.com/xymtex/fujitas3/xymtex/indexe.html

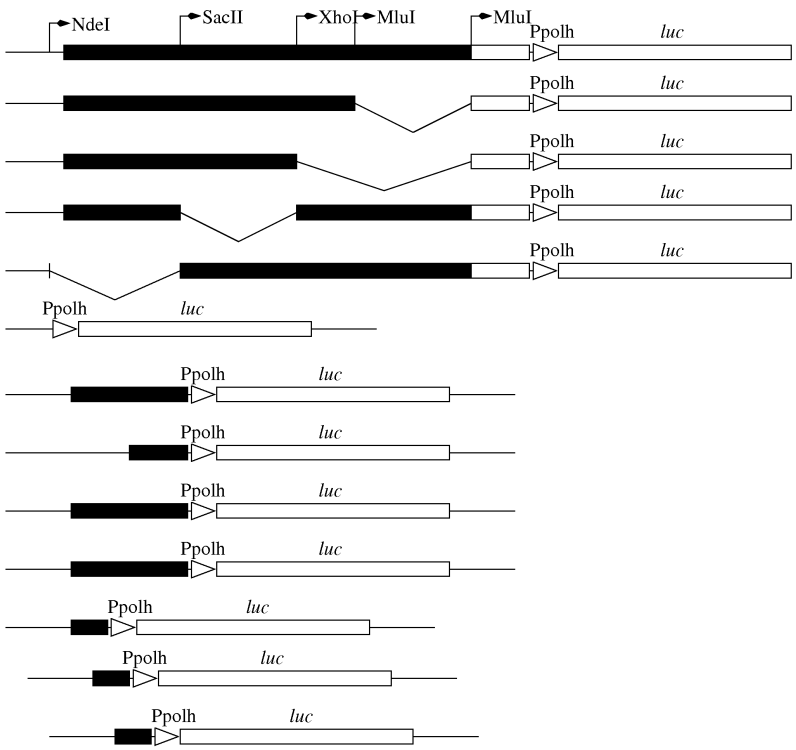


Figure 4: A simple example using XFig. Since the placement of objects in XFig is accurate, it is easy to create line drawings according to scale. Complex chemical structures could also be drawn without much effort. Moreover, the text can be flagged as "special" and LaTeX commands can then be included (Figure reprinted from the author's PhD thesis).

From simple chemical formulae to chemical structures that are complex, including steroid rings, plant products such as flavonoids can also be easily typeset using XyMTeX. Table 1 shows such a list of plant products. A good place to start will be the article written by Fujita Shinsaku[4].

5 Ready to publish your next article in LaTeX?

LaTeX has been the choice of publishers of mathematics and physics journals. The number of journals related to other subjects including biochemistry, molecular biology that accept TeX/LaTeX formatted manuscripts is slowly increasing.

Many of the journals such as *Cell*, *Science*, *Proceedings of the National Academy of Sciences, USA* that publish articles on biology (apart from physical sciences) now accept TeX/LaTeX formatted manuscripts. Well defined style files for preparing the main text of the manuscript and bibliography are available from the journal's website for prospective authors or from people who have created the style and bibliography files for their own use and decided to share it with others. However, LaTeX still has a long way to go before it supercedes `.rtf` & `.doc` as the preferred document format for manuscripts in

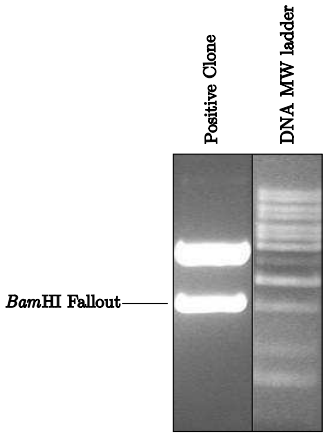


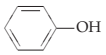
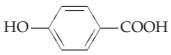
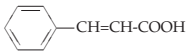
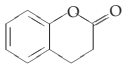
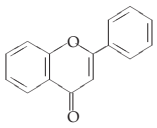
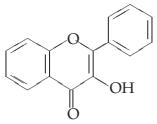
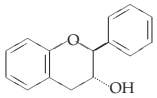
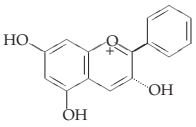
Figure 5: A figure to illustrate the ease of use of Inkscape. Notice the word *Bam*HI. In XFig, it takes a couple of tricks only a few people will attempt to get the italics and the "normal" font in the same word. In Inkscape, LaTeX typesetting commands can be entered even by novices and this feature provides the complete set of LaTeX tools available for authors that are interested in creating correct technical names accompanied by the figures, that they are trying to describe.

biology. Here are some of the sources that the readers will find useful for preparing their manuscripts in LaTeX for submission to journals in the area of Life Sciences.

- TeX FAQ: TeX Frequently asked questions. **The** place usually suggested by TeXperts for people learning TeX & LaTeX.[10]
- TeX showcase: Contains several excellent samples made with LaTeX. If you are not lured by this site into learning LaTeX, I doubt if anything else can make you do it.
- Tom Schneider's Page[11]: Well up to date. Provides style files and `.bst` files for various journals. Has several links to similar sites and lots of useful information.
- LaTeX Bibliography Styles Database: Searchable database of style files.
- Elsevier's site: Provides guidelines for preparing manuscripts in LaTeX for submission to Elsevier journals.[12]
- BioMed Central: Provides instruction for preparing manuscript in LaTeX format for publication in BMC journals.

10. Mailing lists provide a platform for discussing doubts, questions *not* in the FAQ. For example, texhax, TUGIndia mailing lists.
11. Inspired this author to take up using LaTeX seriously.
12. Please note that FEBS Letters, though being a Elsevier's journal, accepts only `.rtf` files.

Table 1: A list of chemical structures typeset using XyMTeX

Class	Basic Skeleton	Basic Structure	Examples
Simple phenols	C_6		Cresol, Thymol
Benzoic acids	C_6-C_1		Gallic acid, Vanillic acid
Cinnamic acids	C_6-C_3		*p*-Coumaric acid, Ferulic acid
Coumarins	C_6-C_3		Umbelliferone, Aesculetin
Flavone	C_6-C_3-C_6		Apigenin, Luteolin, Chrysin
Flavonol			Quercetin, Kaempferol, Myricetin
Flavan-3ols			Catechin, Epicatechin, Epigallocatechin
Anthocyanin			Cyanidin, Malvidin, Delphinidin

6 Conclusion

We saw in this article, the capabilities of the various TEX packages that are available for Life Scientists. Although this list is not comprehensive, the features of these packages discussed here only a glimpse, I hope that the readers will give a serious thought about using these tools that are wide open for them to explore.

About the author

Senthil finished his PhD work on *Gene regulation in very late promoters of Baculovirus*, from the Centre for DNA Fingerprinting & Diagnostics (CDFD), India. He claims to have written his entire PhD thesis, armed with nothing but Emacs + AUCTEX, LATEX and a toothbrush. He is currently working as a *Postdoctoral Research Associate* at the

Systems Biology Lab headed by Dr. Sangdun Choi, at Ajou University, Suwon, Republic of Korea.

References

[1] Peter Flom. LaTeX for academics and researchers who (think they) don't need it. *The PracTeX Journal*, (4), 2005.

[2] Eric Beitz. TeXshade: shading and labeling of multiple sequence alignments using LaTeX2ε. *Bioinformatics*, 16(2):135–139, 2000.

[3] Eric Beitz. TeXtopo: shaded membrane protein topology plots in LaTeX2ε. *Bioinformatics*, 16(11):1050–1051, 2000.

[4] Shinsaku Fujita. XyMTeX for drawing chemical structural formulas. *TUGboat*, 16(1):80–88, 1995.