RNA Folding: Local Versus Global Optimization

Max Ward (20748588)

This report is submitted as partial fulfilment
of the requirements for the Honours Programme of the
School of Computer Science and Software Engineering,
The University of Western Australia,
2014

Abstract

I am definitely going to need to write this at some point. This is a short report on how to use the `cshonours.cls` class to prepare dissertations using the latest LaTeX version, LaTeX2e. This class is based on the standard class `report.cls`.

Keywords: Honours, report, dissertation, UWA, RNA, bioinformatics
CR Categories: Not, really, sure

Acknowledgements

Going to need something here too. This class is designed to produce reports that look the same as those produced by the older `cshonours.sty` style for LaTeX2.09, which was modified by Nick Spadaccini from a style provided by Ken Wessen.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 DNA and RNA

Deoxyribonucleic Acid (DNA) is the basic genetic building block upon which the classification of genetic material into genes and chromosomes is based. The role of DNA as the hereditary unit of genetics was determined in the 1940s [1]. Soon thereafter, Watson & Crick [11] published a highly acclaimed paper describing the fundamental chemical structure of DNA. In it, they outlined a double helix formation which has since become as iconic as it is canonical (see Figure 1.1). Each strand of the helix Watson & Crick discovered is essentially a chain of 'nucleotides' which are made of a sugar-phosphate backbone, attached to a single 'base'. The bases of each strand form hydrogen bonds which hold the double helix together. The most astonishing and important of their findings was that these bases bond in a reciprocal fashion. They described four bases: Adenine (A), which always bonds to Thymine (T), and Guanine (G), which always bonds to Cytosine (C).

The reciprocal bonding relationships between bases is what allows replication to occur; a copy of the DNA can be made by simply allowing the correct bases to bond to one of the strands making up its helix. This gives a model for inheritance and cellular replication. However, there remains the question of how DNA can actually code for protein. Proteins are made up amino acids bonded in a specific sequence [1]. The DNA must therefore code for amino acids. This code, which can be thought of as the 'digital' representation for the 'analogue' protein used by our cells, needs to be carried to ribosomes which translate it into protein [1]. This is a task carried out by Ribonucleic Acid (RNA). RNA is very much like DNA in that it can bond reciprocally to another strand with matching bases. The main difference is that it is single stranded in structure, and has Uracil (U) in place of Thymine [1]. It is important to note that in RNA molecules G and U pairings are also possible. RNA bonds to DNA and, in a sense, reads it. This results in the production of a copy of the DNAs genetic payload. This

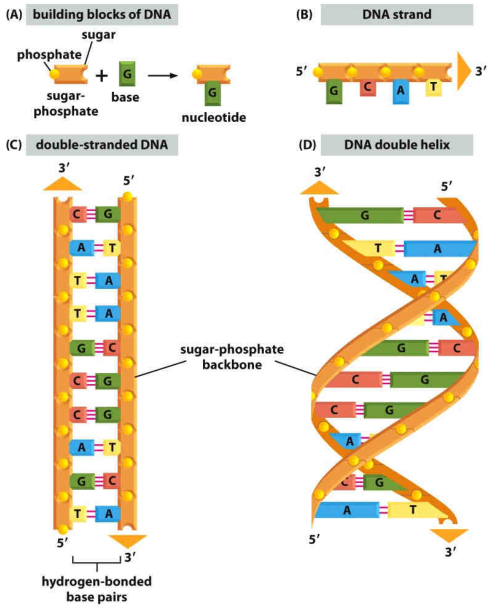

Figure 1.1: The structure and composition of DNA. Diagram taken from ”Essential Cell Biology” [1].

‘downloaded’ information is then carried away to be translated into protein [1]. An example of this is depicted in Figure 1.2, in which we see a Messenger RNA molecule bonding to and thus making a copy of a section of DNA. As depicted in Figure 1.2, the 3’ end of a DNA or RNA molecule is the end onto which new nucleotides are added. The 5’ end is chemically stable, and nucleotides are not usually appended to it [1].

For many years the conventional wisdom was that DNA contained genes which coded for functional proteins used by the cell [1]. Though this is undoubtedly true, there was a problem: much of the human genome, and the genomes of other species, contains DNA which does not appear to code for anything [2]. Many theories have been put forward to explain this. It was argued that this ‘junk’ DNA is the perennial build-up of mutation, and that natural selection simply cannot act with strong enough selective force to cull this free-loading DNA [2]. Surprisingly, much of this non-coding DNA is actually transcribed into RNA, despite having no apparent function [8]. As it turns out, RNA is more than a simple messenger for encoded proteins. Recent research has found myriad important functions for RNA. For example, RNA can act as a catalyst for RNA splicing and peptide bond formation, and can also alter the regulation of genes [12]. It seems that much of our genome contains templates for non-coding RNAs

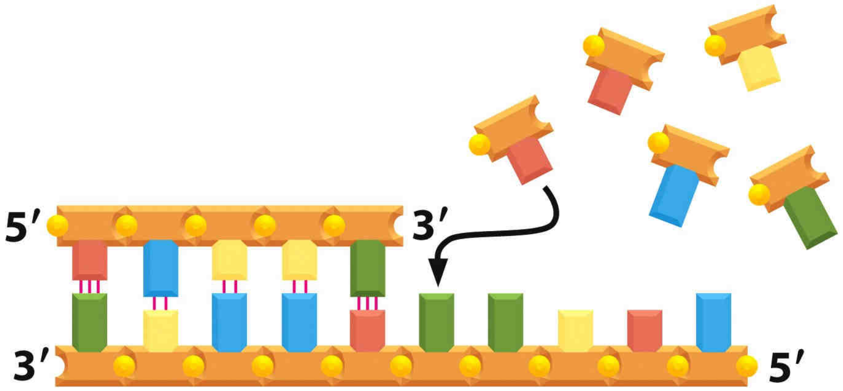

Figure 1.2: RNA transcription. Diagram taken from ”Essential Cell Biology” [1].

(ncRNAs). These RNAs perform essential cellular functions without actually being translated into protein at any point in their life-cycle [8]. Because of its inherently single stranded nature, RNA forms bonds with itself, folding into secondary and tertiary structures [4].

It is axiomatic that chemical structure is tantamount to biological function; RNA is no exception. For this reason there has and continues to be an intense interest in predicting the secondary structure and tertiary structure of RNA molecules. This is in part because it will elucidate the underlying principles of RNA structure formation and function [4], but also because it will allow the detection and classification of unknown RNAs, enable prediction of novel RNA function, and assist the design of new RNA based drugs [3]. In fact, RNA is an extremely versatile molecule, and as such is attractive from both an engineering and computational point of view. Small combinatorial computation problems have been solved by representing the solution set using RNAs. Furthermore, a theory of computation has been put forward using self assembling RNA molecules [3]. As if to comment on the upheaval of a protein-centric view of biology in recent years, researchers have found that RNA is capable of supporting all the processes required for life without the need of protein [3]. The secondary structure of RNA is also highly conserved during evolution, indicating its importance [6]. Secondary and tertiary structures can be treated hierarchically, as a result it is possible to predict the secondary structure of an RNA without understanding the tertiary structure. The tertiary structure in turn builds upon the secondary structure [10]. This paper will focus on secondary structure prediction.

It holds to reason that an algorithm for RNA secondary structure prediction can never be realised if we do not understand how these structures form, or their general morphology. For this reason it is important to understand how true RNA secondary structures can be determined, and the limitations of these techniques. DNA and RNA molecules can be analysed using X-ray crystallographic methods. These types of approaches work because the wavelengths of some X-rays are the same as the dimensions of DNA and RNA inter-atomic bonds. The diffraction of X-ray light by these molecules can thus be observed and their structures can subsequently be inferred by analysis of the resulting data. Nuclear Magnetic Resonance (NMR) is another technique which can be applied to the analysis of DNA/RNA. It relies on the spin of atoms when in a magnetic field. These spin signals can be used to determine the atomic composition and topology of a molecule. This has the advantage of not requiring the molecule under analysis to be crystallized before analysis. Arguably this gives a better in vivo view of RNAs/DNAs, which are fundamentally flexible structures. NMR also has some disadvantages; for instance, it is less accurate than X-ray crystallography, and cannot be used on extremely large molecules. The reason these techniques cannot be used for all RNA structural assays is that they are extremely expensive and time consuming [9].

RNA secondary structure prediction techniques can be broadly broken into two categories: those that use auxiliary information to assist in prediction, and those that predict structure ex nihilo—that is, with nothing but the ‘proband’ sequence we require a structure for. The former approach typically does consensus matching between some sequences for which a user already knows the secondary structures, and a sequence for which the structure is unknown [6]. In this paper I investigate the latter approach because it requires deeper knowledge about why and how RNAs fold. Also, it is the more general of the two.

1.2 Getting Started

The `cshonours` files are located in:

```
/cslinux/cstex/local
```

In order to use the `cshonours` class you need to tell TEX how to find it. To do this simply add the following to your shell resource file (ie. `.zshrc`, `.bashrc`, etc):

```
export TEXINPUTS=$TEXINPUTS:.:/cslinux/cstex//
```

(The double-slash `//` tells TEX to search the tree from this point.) Then open a new shell window to run LATEX in.

If you are using a machine that doesn't mount `cslinux` or a stand-alone system such as a home machine, you can take a copy of the `cshonours.cls` file and put it on your own machine. Please copy the class file directly from the original in the above directory to make sure you have an unadulterated copy.

Once you have told TEX how to find the class file, the easiest way to get started is to copy this example file, `cshonours.tex`, and the accompanying example bibliography file, `cshonours.bib`, from the above directory, give them a new name, and start modifying the text.

1.3 What does it all mean?

The example file is pretty self explanatory, but here's a little elucidation for those who are interested.

```
\documentclass{cshonours}
```

. . . tells LATEX to use the `cshonours` class. The commands between here and the `\begin{document}` command are known as the *preamble* of the latex document. Font size is automatically set to 12pt in this class.

```
\bibliographystyle{acm}
```

. . . sets the bibliography style. Default is the style used in Transactions of the ACM.

```
\usepackage{graphics}  %optional
```

. . . this is only needed if you want to include postscript images.

```
\title{The Honours Dissertation Class for \LaTeX2e}
\author{Cara MacNish}
```

. . . same as usual.

```
\keywords{Honours, report preparation, \LaTeX}
\categories{A.2, I.7.2}
```

. . . keywords and Computing Reviews classification numbers. These will be put at the bottom of the abstract page.

```
\begin{document}
```

. . . so much for the preamble, now we start the document proper.

```
\maketitle
```

. . . produces the title page using the title and author stored earlier. Unlike the standard `report` class it also starts roman page numbering.

```
\begin{abstract}
This is a short report...
\end{abstract}
```

. . . produces the abstract page, including the keywords and categories stored earlier.

```
\begin{acknowledgements}
This style is designed...
\end{acknowledgements}
```

. . . produces the acknowledgements page.

```
\tableofcontents
\listoftables  %optional
\listoffigures  %optional
```

. . . you guessed it! `\listoftables` and `listoffigures` can be omitted if you have no tables or figures respectively.

```
\chapter{The Honours Dissertation Style Guide}
```

. . . and so the first chapter begins. Unlike the standard `report` class the first `\chapter` command also switches pagenumbering to arabic.

The main body is created using the usual LaTeX commands. At the end we come to:

```
\appendix
```

. . . starts off the appendices.

```
\bibliography{cshonours}
```

. . . puts in the bibliography, generated in this case from the file `cshonours.bib`.

1.4 Carrying on. . .

The rest of the document proceeds in the usual way, with all standard LaTeX commands available. These are described in [7], which is written by the author of LaTeX, Leslie Lamport, and commonly known as the LaTeX "Bible".

For those who are feeling ambitious, a wealth of contributed packages, some of which are included in our distribution, and some of which you would need to download yourself, are described in [5], commonly known as the "Doggie Book".

1.5 Including Postscript Files

Most drawing packages (such as `xfig` and `xpaint`) and image manipulation packages (such as `xv` and `gimp`) allow you to save your work as (encapsulated) postscript, which can be easily included in your LaTeX document. The recommended (and simplest!) way of doing this is by including the command `\usepackage{graphics}` in the preamble (see Section 1.3) and then include the postscript file using the `\includegraphics` command.

For example, Figure 1.3 shows a Gnu, produced by the following code:

```
\begin{figure}
\begin{center}
\includegraphics{gnu}
\end{center}
\caption{This is a Gnu.}
\label{gnu}
\end{figure}
```

You can scale graphics using the `\scalebox` command. For example, Figure 1.4 shows a smaller Gnu, produced as follows:

Figure 1.3: This is a Gnu.

Figure 1.4: This is a smaller Gnu.

```
\begin{figure}
\begin{center}
\scalebox{0.6}{\includegraphics{gnu}}
\end{center}
\caption{This is a smaller Gnu.}
\label{smallergnu}
\end{figure}
```

1.6 Producing Postscript Output

LaTeX produces a `.dvi` file which you can convert to postscript using `dvips`. If you have included encapsulated postscript figures the bounding boxes of those figures sometime confuse the printing routines. To overcome this it is recommended you use the `-K` option to strip bounding box comments out. The full recommended format is:

```
dvips -K -f myfile.dvi > myfile.ps
```

To save typing I just use a simple script for all my LaTeXing. Just create a file, called say `laps`, containing something like:

```
latex $1
dvips -K -o $1.ps $1.dvi
```

Then make it executable, run LaTeX with the command

```
laps myfile
```

and view with ghostview (with “State” set to “Watch file”).

1.7 Producing PDF

If you prefer PDF output you can produce this using `pdfelatex`. This does not use `dvi` as an intermediary so you just say:

```
pdfelatex myfile
```

Note that any graphics you include must also be in suitable PDF. There are conversion programs, such as `ps2pdf`, but I've found they often don't work well. It is better to produce the graphics directly in `pdf`.

I've also found that `pdf` viewers don't tend to refresh well, so that you need to keep opening the file. As a result I tend to postscript where possible.

1.8 Emacs and LaTeX

Gnu Emacs and Xemacs recognise both `.tex` and `.bib` files, and provide a number of tools for preparing them. For example you can select `.bib` entry templates from a drop-down menu. Simple commands like `C-c C-e` (puts in the `\end` command to finish an environment) save lots of typing.

1.9 Appendices

After the main body comes the appendices. See Appendix A and Appendix B.

1.10 Bibliography

Finally, the bibliography can be produced automatically from a `.bib` file using `bibtex` in the usual way. This is described in [7].

The bibliography is the only change from the LaTeX2.09 `cshonours` style file. The bibliography now comes after the appendices, in line with printed books, and uses alphanumeric citation tags to make reading (and marking) easier.

APPENDIX A

Original Honours Proposal

You must include as your first appendix an exact copy (in wording) of your original project proposal. This aids other readers to establish what was the initial focus of the project.

APPENDIX B

Another Appendix

Other appendices might include pseudocode for your implementation, a Users Manual, an important data file, etc.

Bibliography

[1] Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. *Essential Cell Biology*, third ed. Garland Science: New York, 2009.

[2] Beaton, M. J., and Cavalier-Smithf, T. Eukaryotic non-coding dna is functional: evidence from the differential scaling of cryptomonad genomes. *Proceedings of the Royal Society of London. Series B: Biological Sciences 266*, 1433 (1999), 2053–2059.

[3] Condon, A. Problems on rna secondary structure prediction and design. In *Automata, Languages and Programming.* Springer, 2003, pp. 22–32.

[4] Conn, G. L., and Draper, D. E. Rna structure. *Current opinion in structural biology 8*, 3 (1998), 278–285.

[5] Goossens, M., Mittelbach, F., and Samarin, A. *The LaTeX Companion.* Addison-Wesley, 1994.

[6] Hofacker, I. L. Rna consensus structure prediction with rnaalifold. In *Comparative Genomics.* Springer, 2008, pp. 527–543.

[7] Lamport, L. *LaTeX : A Documentation Preparation System User's Guide and Reference Manual*, second ed. Adison-Wesley, 1994.

[8] Leung, Y. Y., Ryvkin, P., Ungar, L. H., Gregory, B. D., and Wang, L.-S. Coral: predicting non-coding rnas from small rna-sequencing data. *Nucleic acids research 41*, 14 (2013), e137–e137.

[9] Neidle, S. *Principles of nucleic acid structure.* Academic Press, 2010.

[10] Tinoco Jr, I., and Bustamante, C. How rna folds. *Journal of molecular biology 293*, 2 (1999), 271–281.

[11] Watson, J. D., Crick, F. H., et al. Molecular structure of nucleic acids. *Nature 171*, 4356 (1953), 737–738.

[12] Xu, Z., Almudevar, A., and Mathews, D. H. Statistical evaluation of improvement in rna secondary structure prediction. *Nucleic acids research 40*, 4 (2012), e26–e26.