Using Rule Induction to Elucidate Co-Occurrence Patterns in Microbial Data

by

K. Kumar Thurimella

A thesis submitted to the

University of Colorado in partial fulfillment

of the requirements for the degree of

Bachelors of Science

Department of Applied Mathematics

2013

This thesis entitled:
Using Rule Induction to Elucidate Co-Occurrence Patterns in Microbial Data
written by K. Kumar Thurimella
has been approved for the Department of Applied Mathematics

_____

Rob Knight

_____

Professor Michael Mozer

_____

Senior Instructor Anne Dougherty

Date _____

Thurimella, K. Kumar (B.S., Applied Mathematics)

Using Rule Induction to Elucidate Co-Occurrence Patterns in Microbial Data

Thesis directed by Associate Professor Rob Knight

Several studies have addressed whether the presence or absence of certain bacteria are linked with a particular phenotype. However, it is plausible that the causitive agent (or the consequence) of a given phenotype is not a single microbe, but groups of them. Rule Induction is a commonly used machine learning tool to infer structure within observational data and build rules to represent respective structures. In this thesis I introduce the application of a method, Rule Induction, to infer co-occurrence patterns in microbial data.

First, I benchmark the methods within Rule Induction and understand how rules are generated with regards to several parameters such as table density, support and confidence. I then subsample data over multiple iterations to understand the robustness of the rules being presented and preserved over each sampling.

Next, I provide insight into different biological variables and examine their effect on rules produced. I compare 16s rRNA region, specficially V1-3 and V3-5 regions. I compare different sequencing technology, specifically 454 and Illumina. I finally compare time, specifically looking over 400 days. Within all these comparisons I aim to understand the differences, but more importantly what is conserved within these variables by the means of rules generated.

Finally, I explore Rule Induction on two microbial datasets and see how strong the rules are in comparison to already known associations. The first dataset I interpret regards a correlation between HIV and the Gut Microbiome. The second dataset distinguishes the Gut Microbiome over varying georgraphical locations. I link each of these rules produced from each dataset with taxonomic information and consildate those rules to give rise the underlying structure within the biological data.

Dedication

To my family and friends.

Acknowledgements

First and foremost I would like to thank Rob Knight for letting me be a part of his lab. He has been a fantastic mentor and someone who I look up to very much. His passion for research is contageous and I have learned so much in my time here, and am very fascinated and intruiged by the research being done in the microbiome field in addition to this lab. I would like to thank Mike Mozer for being a great mentor as well, by giving me great advice specific to this project as well as general life advice. My final committee member, Anne Dougherty, has been nothing short of a phenomenal advisor. It is after talking to her my sophomore year that I switched my major to Applied Mathematics. She has consistently provided great support and I can go to her about anything like another friend.

I would like to thank Jose Clemente for his ideas, support and fabulous mentoring. I have learned a lot from Jose with regards to his perspective on being a researcher. Will van Trueren has been nothing but a great friend and peer mentor in the lab all the while providing great insights to this research. I want to finally thank other members of this lab including Yoshiki Baeza for his help understanding cluster computing, Cathy Lozupone for her HIV data and insight within co-occurrence and many others who were always there for support. I want to acknowledge the Gautam Dantas Lab at Washington University in St. Louis, specifically Kevin Forsberg and Mitch Pesesky. This past summer opened my eyes to exciting avenues of research and that experience provided the direction of research that I continue to this day.

Without my close friends I wouldn't be where I am at today. Thanks to my roommates

David Gillis and Thomas Lynn for helping me out this year. Oriel Eisner for always being interested in (or putting with) our late night talks, mostly regarding science. Many thanks Sathish Subramanian for being a fantastic role model and providing help and support whenever I needed it, as well as late night life talks. I hope to be half the MD/PhD student he is, one day. Will Timbers has always been incredibly fun to talk with and has pushed me to pursue my passions. Andrew Fleming has shared a same passion for science with me in high school and I appreciate all of his support over the years. Without Myke Samuels I don't know where I would be, and it was because of his help that I found my passion and interests and for that I am forever grateful.

Finally thanks to my brother and parents. Coming from a family of computer scientists certainly rubs off and because of all their love and support I have been able to develop my own passions.

Contents

Chapter

Appendix

Tables

Table

Figures

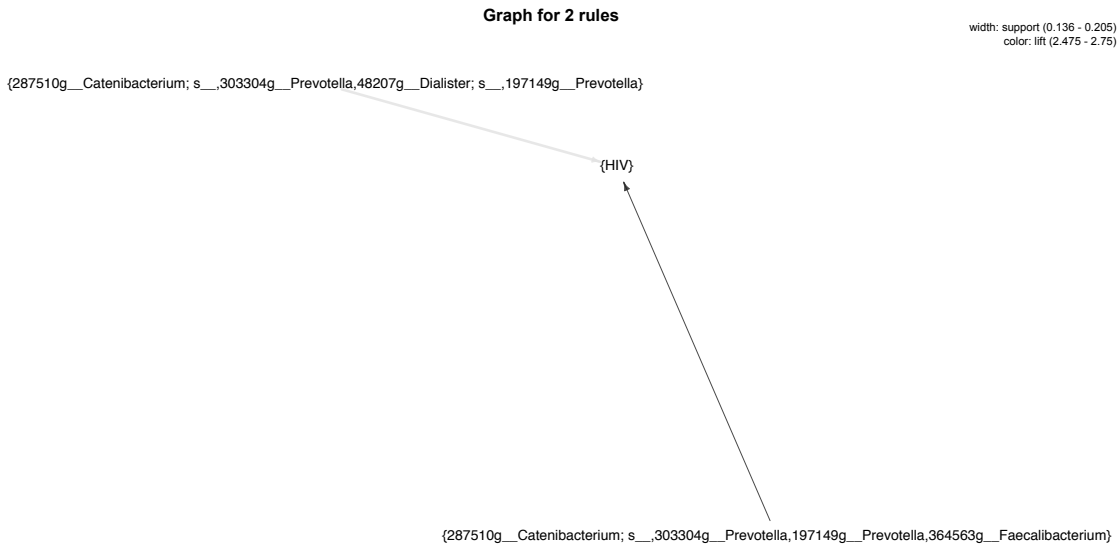
Figure

parskip

Chapter 1

Introduction

From protists to humans, all plants and animals live in close association with microbial organisms. The microbiome is the full collection of these microbes, their genome, and their environmental interactions. Scientists are only now starting to appreciate the complex interactions between microbial communities and organisms. It is important to study these complex interactions because they have been linked with initiating ailments, as well as fostering health. In particular, recent studies have reveled that microbial communities have been associated with diseases such as diabetes, obesity, and rheumatoid arthritis. Furthermore, microbiomes are also being studied in the context of the environment as well. With the amount of recent interest in studying microbiomes, some consider it to be a newly discovered organ.



Although modern technology such as DNA testing has allowed scientists to start exploring microbiomes, the methods in processing and understanding microbial data are still evolving. In other words, although scientists are often able to collect a lot of data, they often have a difficult time interpreting the data. In this thesis we discuss a machine learning principle called rule induction which yields insight into microbial data and help understand the complex interactions that occur between organisms.

This sample document illustrates how to use the `thesis` class, originally written by John P. Weiss. Some requirements of the Graduate School are written into that file; page size, line spacing, appropriate placement of captions for tables and figures, etc. Other tasks of conforming to the requirements are left to other existing LaTeX packages. For example, a common problem is to insert graphics — figures and tables — into the body of the thesis. For this one should use the `graphicx` package, which is part of the standard TeX distribution. Likewise, the Grad School specs say that a large table may be displayed in landscape mode at reduced size, but its caption must also be in rotated position, in the same font and size

as the normal text in the body of the thesis. To accomplish this, the user must invoke the `rotating` package, available online.

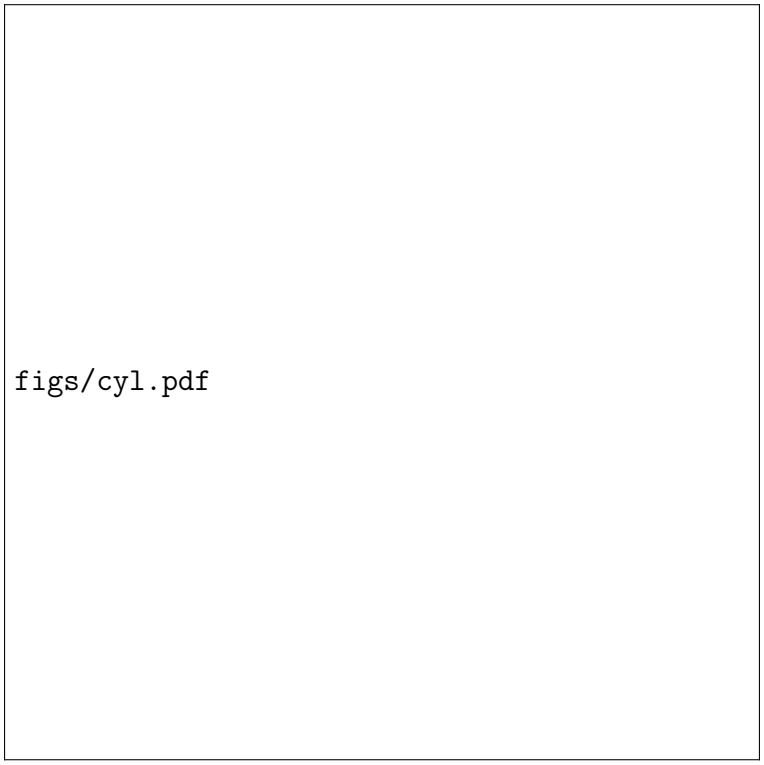
Figure 1 shows something or other; the image is from a PDF file imported into this document using the `graphicx` package. The command `\usepackage{graphicx}`, which appears near the very top of the main LaTeX file, reads in this package which defines the `\includegraphics{}` macro.

1.1 Lists in `thesis` class

In `thesis` class (for Colorado University), lists are defined so that nested lists will be numbered or marked appropriately. First, an itemized (non-enumerated) list prefaces each item with a bullet. Nested itemized list use asterisks, then dashes, then dots. These lists are typed between the `\begin{itemize}` and `\end{itemize}` commands.

- This is "itemized" item A.
- This is "itemized" item B.
- This is "itemized" item C.
    - This is "itemized" subitem A.
        - This is "itemized" subsubitem A.

Figure 1.1: This diagram of a cylinder and various measurements and quantities was actually made using **xfig**, a freeware drawing program for Unix systems. Diagrams can be exported directly to PDF files, the preferred format for vector graphics. Vector graphics can be magnified indefinitely without degradation, whereas bitmap images (JPG and PNG) must be pretty high-resolution if you don't want them looking all pixellated when magnified.



- This is "itemized" subsubsubitem A.

– This is "itemized" subsubitem B.

∗ This is "itemized" subitem B.

• This is "itemized" item D.

Enumerated lists use the commands `\begin{enumerate}` and `\end{enumerate}`, and nested enumerations appear like this.

(1) This is "enumerated" item A.

(2) This is "enumerated" item B.

(3) This is "enumerated" item C.

(a) This is "enumerated" subitem A.

(i) This is "enumerated" subsubitem A.

(i.a) This is "enumerated" subsubsubitem A.

(ii) This is "enumerated" subsubitem B.

(b) This is "enumerated" subitem B.

(4) This is "enumerated" item D.

The work presented here[1] is an extension of Taum[**?**] and Lao et al.[**?**], fictional references that are in the bibliographic source file .bib.

[1] Footnotes are handled neatly by LaTeX.

Table 1.1: Here is an example of a table with its own footnotes. Don't use the `\footnote` macro if you don't want the footnotes at the bottom of the page. Also, note that in a thesis the caption goes **above** a table, unlike figures.

wave form	S (kVA)	P (kW)	Q^* (kVAr)	$D^\dagger$ (kVAd)
Fig. 1a	25.48	25.00	-2.82	4.03
Fig. 1b	25.11	18.02	-9.75	14.52
Table 2.1	24.98	22.26	9.19	6.64
Table 1.1	23.48	15.00	6.59	16.82
Fig. 2.1	24.64	22.81	-0.44	9.3

*kVAr means reactive power.
†kVAd means distortion power.

Chapter 2

Understanding Rule Induction

The objective of this fake thesis document is to demonstrate a multitude of LaTeX features as well as features specific to the thesis class. We start by giving one short formula, and one big hairy multi-line formula (one of the non-dimensional Navier-Stokes equations):

$$A = \pi r^2 \tag{2.1}$$

$$\begin{aligned}\rho\left[\frac{DV_r}{Dt} - M\epsilon^2\frac{V_\theta^2}{r}\right] &= -\frac{\delta^2}{\gamma\ M}\frac{\partial P}{\partial r} + \frac{M\ \delta^2}{Re}\left\{2\frac{\partial}{\partial r}\left[\mu\left(\frac{\partial V_r}{\partial r} - \frac{1}{3}\nabla\cdot\overline{\mathbf{V}}\right)\right]\right. \\ &\left. +\frac{1}{r}\frac{\partial}{\partial\theta}\left[\mu\left(\frac{1}{r}\frac{\partial V_r}{\partial\theta} + \epsilon\frac{\partial V_\theta}{\partial r} - \epsilon\frac{V_\theta}{r}\right)\right]\right.\end{aligned}$$

$$+\frac{\partial}{\partial z}\left[\mu\left(\frac{1}{\delta^2}\frac{\partial V_r}{\partial z}+\frac{\partial V_z}{\partial r}\right)\right]$$
$$+2\,\frac{\mu}{r}\left[\frac{\partial V_r}{\partial r}-\frac{\epsilon}{r}\frac{\partial V_\theta}{\partial \theta}-\frac{V_r}{r}\right]\Bigg\}, \tag{2.2}$$

2.1 Explanation of equations

The latter equation is non-dimensionalized using the following definitions:

$$r=\frac{r'}{R'},\quad z=\frac{z'}{L'},\quad t=\frac{t'}{t'_a},\quad \kappa=\frac{\kappa'}{\kappa'_0},\quad \mu=\frac{\mu'}{\mu'_0},\quad C_V=\frac{C'_V}{C'_{V0}},$$

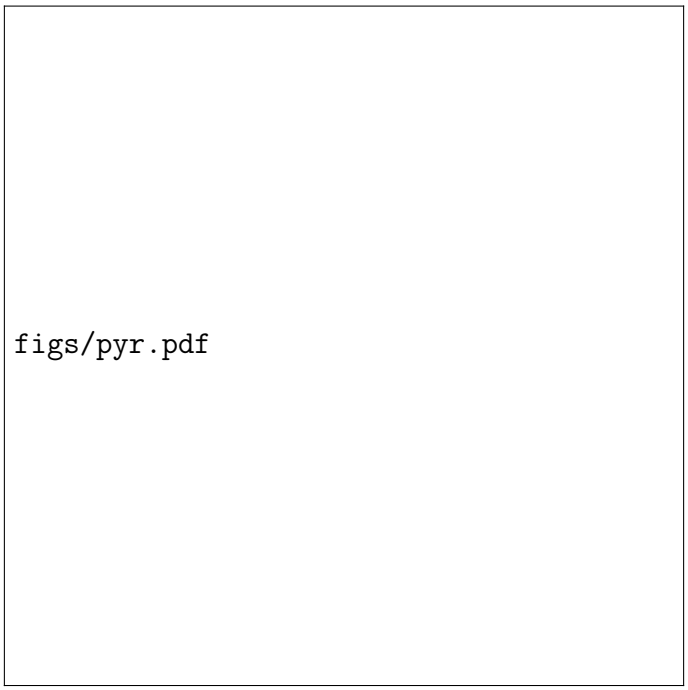
where P'_0 is the initial static pressure in the cylinder, and ρ'_0 and T'_0 are the density and temperature of the fluid being injected from the sidewall.

Here is an example of using the macros `\singlespacing` and `\doublespacing`:

This paragraph was preceded by the command `\singlespacing`. See the Specifications of the Grad School for instructions about when single spacing is appropriate in a thesis.

And now, here is an example of using the macros `\begin{singlespace}` and `\end{singlespace}`; another way to get single-spacing.

Figure 2.1: A triangular pyramid may be cut up as shown, to yield one top pyramid (with one-eighth the volume of the full pyramid), three bottom corner pyramids (which, when joined, are congruent to the top pyramid), three prisms along the bottom edges (the area of whose bottom faces total $B/2$) and the large central prism (volume $= (B/4)(h/2) = Bh/8$). The image, from PDF file "pyr.pdf", was read in using the `\includegraphics` command, from the `graphicx` package.



Two cases are studied in the present work which differ only in the boundary conditions. Each different boundary condition model a different source of instability. The boundary of the first case consists of a steady, axisymmetric sidewall radial velocity boundary and a time-dependent, non-axisymmetric endwall axial velocity boundary. The second case is studied with a fixed impermeable axial velocity along the endwall and a combination axisymmetric steady and non-axisymmetric unsteady radial velocity along the sidewall.

Usually you want to use a table produced by some other software, such as Excel, rather than try to do it using LaTeXmacros. If the table is saved/printed to a PDF file, then it can be displayed using the `\includegraphics` macro inside a `table` environment:

Some of the boundary conditions are:

$$z = 0; \qquad V_z = \begin{cases} 0, & t \le 0 \\ \widetilde{F}_{zw}(r,\theta,t), & t > 0 \end{cases} \tag{2.3}$$

$$z = 0; \qquad V_\theta = V_r = 0 \tag{2.4}$$

$$r = 0; \qquad P, \rho, T, V_r, V_\theta, V_z \text{ finite}, \tag{2.5}$$

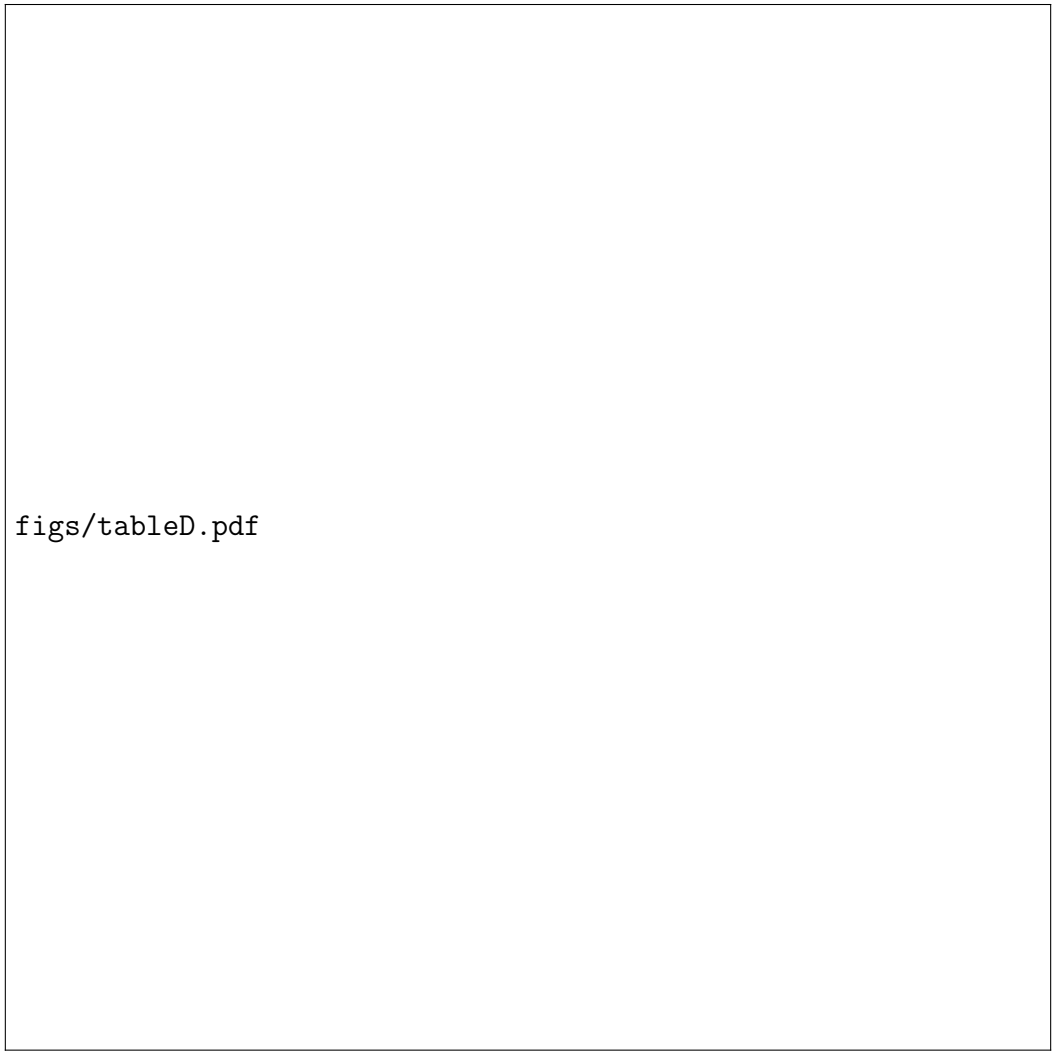
$$r = 1; \qquad V_r = F_{rws}(z), \tag{2.6}$$

$$r = 1; \qquad V_z = V_\theta = 0, \tag{2.7}$$

and solutions must be periodic in θ.

If you don't believe this stuff, check out Mulick[**?**] and Baylor[**?**].

Table 2.1: This table wasn't constructed with LaTeX commands, but resides in PDF file (`tableD.pdf`) created by some other software.



2.2 Yet another section

2.2.1 Just meaningless text to test lines per page

According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According

to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform? According to the Grad School specs. there should be 24–27 lines of print per page of a thesis. This should be true whether the font size is 10, 11, or 12. Count them up; does this document conform?

What is it? This is a labelled paragraph. The heading of the paragraph is emphasized. This is a labelled paragraph. The heading of the paragraph is emphasized.

2.2.2 This is a subsection

This is a subsection. Filler filler filler filler filler filler filler filler. Filler filler filler filler filler filler filler filler.

2.2.3 This is another subsection

This is another subsection. Filler filler filler filler filler filler filler filler. Filler filler filler filler filler filler filler filler.

This is paragraph number 2. It used a `\paragraph{}` header, which are always inlined (with extra space) and boldfaced.

This is the third paragraph of the subsection. Filler filler filler filler filler filler filler filler. Filler filler filler filler filler filler filler filler.

2.2.3.1 This is a subsubsection (1)

This is the first paragraph of the subsubsection. Whether it is numbered or inlined depends on the option selected at the beginning of the thesis.

By default, a `\subsubsection` heading is numbered and set off on a separate line, left-justified.

However. Using the `inlineh4` option, subsubsection headers are inlined. And using the `nonumh4` option suppresses numbering of the subsubsections. Together they make subsubsection headings just the same as paragraph headings.

2.2.3.2 This is another subsubsection (2)

Once again, whether its heading is numbered and/or inlined depends on the class options chosen at the start.

There is no "subsubsubsection" entity, and "subparagraph" gets no special treatment in **thesis** class.

2.3 The End

Finally, this is the end. The bibliography starts on the next page.

Appendix A

Weird Exam Answers

About appendices: Each appendix follow the same page-numbering rules as a regular chapter; the first page of a (multi-page) appendix is not numbered. By the way, the following are supposedly authentic answers to English GCSE exams!

(1) The Greeks were a highly sculptured people, and without them we wouldnt have history. The Greeks also had myths. A myth is a female moth.

(2) Actually, Homer was not written by Homer but by another man of that name.

(3) Socrates was a famous Greek teacher who went around giving people advice. They killed him. Socrates died from an overdose of wedlock. After his death, his career suffered a dramatic decline.

(4) Julius Caesar extinguished himself on the battlefields of Gaul. The Ides of March murdered him because they thought he was going to be made king. Dying, he gasped out: Tee hee, Brutus.

(5) Nero was a cruel tyranny who would torture his subjects by playing the fiddle to them.

(6) In midevil times most people were alliterate. The greatest writer of the futile ages was Chaucer, who wrote many poems and verses and also wrote literature.

(7) Another story was William Tell, who shot an arrow through an apple while standing on his sons head.

(8) Writing at the same time as Shakespeare was Miguel Cervantes. He wrote Donkey Hote. The next great author was John Milton. Milton wrote Paradise Lost. Then his wife died and he wrote Paradise Regained.

(9) During the Renaissance America began. Christopher Columbus was a great navigator who discovered America while cursing about the Atlantic. His ships were called the Nina, the Pinta, and the Santa Fe.

(10) Gravity was invented by Issac Walton. It is chiefly noticeable in the autumn when the apples are falling off the trees.

(11) Johann Bach wrote a great many musical compositions and had a large number of children. In between he practiced on an old spinster which he kept up in his attic. Bach died from 1750 to the present. Bach was the most famous composer in the world and so was Handel. Handel was half German half Italian and half English. He was very large.

(12) Soon the Constitution of the United States was adopted to secure domestic hostility. Under the constitution the people enjoyed the right to keep bare arms.

(13) The sun never set on the British Empire because the British Empire is In the East and the sun sets in the West.

(14) Louis Pasteur discovered a cure for rabbis. Charles Darwin was a naturalist who wrote the Organ of the Species. Madman Curie discovered radio. And Karl Marx became one of the Marx brothers.

Appendix B

Ode to Spot

(Data, Stardate 1403827) (A one-page chapter — page must be numbered!) Throughout the ages, from Keats to Giorchamo, poets have composed "odes" to individuals who have had a profound effect upon their lives. In keeping with that tradition I have written my next poem . . . in honor of my cat. I call it. . . Ode. . . to Spot. (Shot of Geordi and Worf in audience, looking mystified at each other.)

Felus cattus, is your taxonomic nomenclature
an endothermic quadruped, carnivorous by nature?
Your visual, olfactory, and auditory senses
contribute to your hunting skills, and natural defenses.
I find myself intrigued by your sub-vocal oscillations,
a singular development of cat communications
that obviates your basic hedonistic predilection
for a rhythmic stroking of your fur to demonstrate affection.
A tail is quite essential for your acrobatic talents;
you would not be so agile if you lacked its counterbalance.

And when not being utilized to aid in locomotion,
It often serves to illustrate the state of your emotion.

(Commander Riker begins to applaud, until a glance from Counselor Troi brings him to a halt.) Commander Riker, you have anticipated my denouement. However, the sentiment is appreciated. I will continue.

O Spot, the complex levels of behavior you display
connote a fairly well-developed cognitive array.
And though you are not sentient, Spot, and do not comprehend
I nonetheless consider you a true and valued friend.